APPLICATION OF BAYESIAN HIERARCHICAL MODELS IN GENETIC DATA ANALYSIS

A Thesis

by

LIN ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Bani K. Mallick
	Veera Baladandayuthapani
Committee Members,	Raymond J. Carroll
	Garry Adams
Head of Department,	Simon J. Sheather

December 2012

Major Subject: Statistics

ABSTRACT

Genetic data analysis has been capturing a lot of attentions for understanding the mechanism of the development and progressing of diseases like cancers, and is crucial in discovering genetic markers and treatment targets in medical research. This dissertation focuses on several important issues in genetic data analysis, graphical network modeling, feature selection, and covariance estimation. First, we develop a gene network modeling method for discrete gene expression data, produced by technologies such as serial analysis of gene expression and RNA sequencing experiment, which generate counts of mRNA transcripts in cell samples. We propose a generalized linear model to fit the discrete gene expression data and assume that the log ratios of the mean expression levels follow a Gaussian distribution. We derive the gene network structures by selecting covariance matrices of the Gaussian distribution with a hyper-inverse Wishart prior. We incorporate prior network models based on Gene Ontology information, which avails existing biological information on the genes of interest. Next, we consider a variable selection problem, where the variables have natural grouping structures, with application to analysis of chromosomal copy number data. The chromosomal copy number data are produced by molecular inversion probes experiments which measure probe-specific copy number changes. We propose a novel Bayesian variable selection method, the hierarchical structured variable selection (HSVS) method, which accounts for the natural gene and probe-within-gene architecture to identify important genes and probes associated with clinically relevant outcomes. We propose the HSVS model for grouped variable selection, where simultaneous selection of both groups and within-group variables is of interest. The HSVS model utilizes a discrete mixture prior distribution for group selection and group-specific Bayesian lasso hierarchies for variable selection within groups. We further provide methods for accounting for serial correlations within groups that incorporate Bayesian fused lasso methods for within-group selection. Finally, we propose a Bayesian method of estimating high-dimensional covariance matrices that can be decomposed into a low rank and sparse component. This covariance structure has a wide range of applications including factor analytical model and random effects model. We model the covariance matrices with the decomposition structure by representing the covariance model in the form of a factor analytic model where the number of latent factors is unknown. We introduce binary indicators for estimating the rank of the low rank component combined with a Bayesian graphical lasso method for estimating the sparse component. We further extend our method to a graphical factor analytic model where the graphical model of the residuals is of interest. We achieve sparse estimation of the inverse covariance of the residuals in the graphical factor model by employing a hyper-inverse Wishart prior method for a decomposable graph and a Bayesian graphical lasso method for an unrestricted graph.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Bani Mallick. He has been the ideal supervisor for me for his sage advice, insightful criticisms, and patient encouragement aiding the completion of this dissertation. Most importantly, his excellent guidance on how to conduct research and develop professional career will benefit me remarkably as a researcher in the long run.

I am deeply grateful to my co-advisor, Dr. Veera Baladandayuthapani, who provides me the opportunity to work with him and the excellent team at the University of Texas M.D. Anderson Cancer Center, and guides me in the research on real data analysis. His sincerity and unfailing support has been and will be an inspiration for me to progress in my professional career.

I would like to thank Dr. Garry Adams, Dr. Raymond Carroll, and Dr. Michael Longnecker for their patience and encouragement to my completing Ph.D. study.

I also wish to thank Dr. Kim-Anh Do, Dr. Patricia Thompson, Dr. Melissa Bondy, and the team at the University of Texas M.D. Anderson Cancer Center, who helped me in my research project.

I also own my thanks to my colleagues Soma Dhavala, Abhra Sarkar, Anindya Bhadra, and Xiaolei Xun for helping me with my work and inspiring me in my research.

Finally, this dissertation is dedicated to my parents and husband for their love, endless support and encouragement, giving me the strength to continue and not give up.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

1. INTRODUCTION

1.1 Problem Formulation

Significant advances in DNA sequencing strategies over the past decade have revolutionized the field of genomic research, allowing for development of many genome-wide technologies like microarray, serial analysis of gene expression (SAGE), RNA sequencing, and molecular inverse probes (MIPs) experiments. These high-throughput technologies make deep genome sequencing and transcription quantification, and provide information on up to thousands of genes simultaneously. Availability of vast amounts of high-dimensional data opens up a new opportunity to understand the mechanism of biological processes, and, as well, brings up challenges in methodology development for analyzing data of different types and characteristics. In this dissertation, we concern several important issues in genetic data analysis including graphical network modeling, feature selection, and covariance estimation. We propose novel statistical methods and models to address the nature of different types of genetic data, and attempt to move towards more structured approaches to leverage information in statistical analysis.

1.2 Organization

In Section 2, we propose an algorithm for modeling gene networks based on discrete gene expression data. We specifically focus on the discrete expression data from serial analysis of gene expression experiments (Velculescu et al., 1995). We assume that the observed counts of mRNA transcripts are from independent Poisson processes, with the mean rates to be the true transcriptional levels. The log ratios of the mean counts are considered to follow a multivariate normal distribution, whose inverse covariance matrix gives the conditional independence structure of the gene network model. We utilize a conjugate prior for the covariance matrices, the hyper-inverse Wishart distribution introduced by Dawid and Lauritzen (1993), and an MCMC-based algorithm to identify graphical models. Furthermore, we propose a prior for the graphical models based on GO information, which utilizes prior information on the genes of interest obtained in biological research as well as inducing sparsity in the graphical models as is assumed in gene regulatory networks. We conduct simulation studies to examine the performance of our discrete graphical model and apply the method to real discrete datasets in identifying the gene regulatory networks.

In Section 3, we concern the issue of hierarchical feature selection in analysis of copy number data. Changes in chromosomal copy numbers have been identified as important causes of cancer (Pinkel and Albertson, 2005), and hence analysis of chromosomal copy number alterations has the potential to identify genetic markers and treatment targets for cancers. In this section, we consider a high-dimensional copy number profile obtained from molecular inversion probes experiments which measure probe copy number changes. Our goal is to ascertain probe-specific copy number alterations that are correlated with patient clinical characteristics. Since the probes located in the coding region of one gene can be taken as a natural group, we propose a Bayesian variable selection method, the hierarchical structured variable selection (HSVS) method, which accounts for the natural grouping structures in the data and simultaneously selects both gene groups and within-gene probes. The HSVS model utilizes a discrete mixture prior distribution for group selection and group-specific Bayesian lasso hierarchies for variable selection within groups. We further accounts for serial correlations within a gene by incorporating Bayesian fused lasso methods for within-group selection. Through simulations we establish that our method results in lower model errors than other methods when a natural grouping structure exists. We apply our method to an MIP study of breast cancer and show that it identifies genes and probes that are significantly associated with clinically relevant subtypes of breast cancer.

In Section 4, we consider the problem of estimating high-dimensional covariance matrices of a particular structure, which is a summation of low rank and sparse matrices. This covariance structure can be applied to multiple statistical models such as factor analytical model, random effects model, and conditional covariance model. We propose a novel Bayesian method of estimating the covariance matrices with such decomposition structure by rewriting the covariance model in the form of a factor analytic model where the number of latent factors is unknown. Our object is to estimate the covariance as well as recovering the rank of the low rank component and the support of the sparse component. We estimate the rank of the low rank component through factor selection with latent binary indicators, and use a Bayesian graphical lasso selection prior for the sparse component estimation. Simulation studies show that our method can recover the rank and the sparsity of the two components respectively with high frequencies. We further extend our method to a graphical factor analytic model, by which we recover the number of factors as well as the graphical model of the residuals. To induce sparsity in the inverse covariance of the residuals, we employ a hyper-inverse Wishart prior method for modeling decomposable graphs, and a Bayesian graphical lasso method for unrestricted graphs. We show through simulations that the extended model can recover both the number of latent factors and the graphical model of the residuals successfully when the sample size is sufficient relative to the dimension.

Finally, we summarize our main findings and suggest future research directions in Section 5.

2. GRAPHICAL MODEL INFERENCE FOR DISCRETE GENE EXPRESSION DATA

2.1 Introduction

A gene network is a collection of genes that influence the expression levels of each other indirectly through their RNA or protein products. Gene network inference is a task critical for revealing signaling pathways in cells, understanding the occurrence and development of diseases like cancers, and identifying target genes for disease treatment. With the development of genome-wide technologies like RNA fingerprinting, expressed sequence tag sequencing, serial analysis of gene expression (SAGE), and microarrays, high dimensional gene expression data become available for mapping the interactions between thousands of genes simultaneously. Current statistical modeling of gene networks is primarily based on continuous gene expression profiles obtained from microarray experiments, with the gene expression data assumed to follow a Gaussian distribution, which has many well-established properties.

In this section, we propose an algorithm for modeling gene networks based on discrete gene expression data. We specifically focus on the discrete data from SAGE experiments (Velculescu et al., 1995). In a SAGE experiment, all the mRNA transcripts of a cell sample are collected and a 10-base DNA fragment is released from each mRNA transcript, which is called a SAGE tag. The number of the tags with the same nucleotide sequence is then counted in a cell sample. Since the nucleotide sequence of a tag is specific to the mRNA from which the tag is released, the count of the tags of a particular sequence gives the amount of their corresponding mRNA transcripts in a cell sample. Note that the counts of mRNAs from SAGE experiments are relative quantities with respect to the total number of transcripts collected in a cell sample. In a typical SAGE experiment, a large number of mRNA transcripts (often from 30,000 to 100,000) are collected from each cell sample.

Similar to microarray, SAGE produces a snapshot of gene expression profiles by measuring the levels of mRNA transcription in a cell sample. However, SAGE provides gene expression profiles with a higher level of genome coverage than microarray, because SAGE is not limited to expression analysis of known genes, as is microarray. Furthermore, SAGE experiments give the relative amount of each gene's mRNAs with respect to the total mRNA transcripts in a cell sample. Thus we can compare mRNA levels among libraries generated by different laboratories. Due to these unique features of SAGE experiments, analysis of SAGE data plays an important role in biological and biomedical areas, such as prediction of new gene function and identification of target genes for disease treatment. In this section, we explore discrete SAGE datasets for gene network structures with an undirected graph.

There have been many approaches to network modeling with Gaussian graphical models. In a Bayesian setting, Gaussian graphical models are based on hierarchical specifications for the covariance matrix (or precision matrix) using global conjugate priors on the space of positive-definite matrices, such as the inverse Wishart prior or its equivalents. Dawid and Lauritzen (1993) introduced an equivalent form as the hyper-inverse Wishart (HIW) distribution. This construction enjoys many advantages, such as computational efficiency, due to its conjugate formulation and exact calculation of marginal likelihoods (Scott and Carvalho, 2008). Giudici (1996) used a prior for the covariance matrix that is a mixture of HIW priors with fixed parameters over decomposable graphs and calculated the posterior probability of each graph. Armstrong et al. (2009) extended this method by proposing a prior that assigns equal probabilities over graph sizes and utilized a conditional Markov chain Monte Carlo (MCMC) sampler. These methods have been extended for nondecomposable graphs using reversible-jump algorithms (Giudici and Green, 1999; Brooks, Giudici, and Roberts, 2003). Moreover, the G-Wishart prior distribution has been proposed as a generalization of HIW priors that is suitable for nondecomposable graphs (Roverato, 2002; Atay-Kayis and Massam, 2005). Gaussian graphical models have been widely used to infer the regulatory relationship among genes for continuous gene expression data at the transcriptional level (Wu, Ye, and Subramanian, 2003; Dobra et al., 2004; among others). The conditional independence arising out of a Gaussian graphical model is flagged by the zero off-diagonal elements in the inverse covariance matrix.

In this section, we develop Bayesian graphical models for discrete gene expression data. We assume that the observed counts of mRNA transcripts in a SAGE experiment are from Poisson processes, with the means to be the true transcriptional levels. The log ratios of the mean counts are considered to follow a multivariate normal distribution. That is, the expression levels of genes are regulated by each other through a Gaussian graphical model underlying the log ratios of the means, whose inverse covariance matrix gives the conditional independence structure of the undirected gene network. We utilize the conjugate HIW prior to sample the covariance matrices and an MCMC-based algorithm to identify graphical models. Furthermore, we propose a prior for the graphical models based on GO information, which utilizes prior information on the genes of interest obtained in biological research as well as inducing sparsity in the graphical models as is assumed in gene regulatory networks.

We obtain the GO information from the GO consortium, which provides a controlled vocabulary of terms describing gene product characteristics in the aspects of cellular component, molecular function, and biological process (Ashburner et al., 2000). For each of the three fields, GO terms are organized in a hierarchical directed acyclic graph (DAG) structure, reflecting the associations between ontology terms. For example, the biological process terms "calcium-mediated signaling" and "leukemia signaling" are two daughter terms of the term "intracellular signaling," meaning that they are two kinds of intracellular signaling; and "intracellular signaling" is a daughter term of "signaling transduction." Hence, two genes sharing the same or similar GO terms in biological process may have the same or similar cellular functions. Based on this idea, methods have been developed to measure the semantic similarity between GO terms and gene products (Resnik, 1999; Wang et al., 2007).

These gene similarity measures based on associated GO terms have been used in gene clustering and gene function prediction (Kustra and Zagdanski, 2006).

In our method, we apply GO-derived semantic similarity measurements to gene network modeling. We measure the functional semantic similarity of each pair of the genes of interest based on the relatedness of their associated GO terms. The semantic similarity score is then taken as the prior probability of an edge between the two genes in the gene network. Using this method, we derive a prior for the graphical model by taking the product of the prior probabilities of the edges in the graph. This GO-based prior on the gene network incorporates biological information of the genes into gene network modeling as well as bringing scientifically interpretable sparsity in the inferred graphical models.

We introduce our Bayesian hierarchical model and the GO-based prior derivation in Section 2.2. We describe a model selection method based on false discovery rates (FDRs) for inferring graphical models from posterior samples in Section 2.3. In Section 2.4, we show the results of a simulation study evaluating the performance of the discrete graph modeling method. In Section 2.5, we present the result of a real SAGE data analysis to model the gene networks in breast cancer cells. Finally, a short summary of our method is included in Section 2.6. The schemes of posterior sampling are detailed in Appendix A.

2.2 Probability Model

Let X denote an $n \times q$ matrix of discrete gene expression profiles, with X_{ij} to be the observed mRNA count of gene j ($j = 1, \cdots, q$) obtained in a SAGE experiment for the i^{th} ($i = 1, \cdots, n$) individual. Since a SAGE experiment counts the transcripts of a gene given a large total number of transcripts in a cell sample, we assume that each count, X_{ij}, follows a Poisson distribution with mean λ_{ij}. We consider λ_{ij}, the expected count of the transcripts given a total number of transcripts, as the true transcriptional expression amount of gene j in the i^{th} cell sample. We assume that the log ratios of $\boldsymbol{\lambda_i} = (\lambda_{i1}, ..., \lambda_{iq})'$ for $i = 1, \cdots, n$ follow a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. The likelihood is specified as follows:

$$\begin{aligned} X_{ij} &\sim Pois(\lambda_{ij}), \\ \log(\boldsymbol{\lambda_i}) &\sim \mathcal{N}_q(\boldsymbol{\mu}, \Sigma). \end{aligned}$$

This is a special case of popular generalized linear mixed models (Zeger and Karim, 1991; Breslow and Clayton, 1993). In this framework, we assume a graphical model through Σ to account for the association structure among the underlying log ratios of the expression levels in a cell sample. Our focus is to infer the graphical model by selecting the covariance matrix Σ.

2.2.1 Bayesian Gaussian Graphical Models

In a Bayesian framework, Gaussian graphical modeling is based on hierarchical prior specifications for the covariance matrix Σ at the two levels: a prior distribution for Σ under each graph and a prior distribution over different graphs. Before giving the details about the hierarchical priors, we first describe the notations on Gaussian graphical models.

An undirected graph is a pair of $G = (V, E)$ with a vertex set $V = \{1, ..., q\}$ and an edge set $E \subseteq V \times V$. Nodes i and j are adjacent or connected in G if $(i, j) \in E$, whereas i and j are conditionally independent if $(i, j) \notin E$. A graph G with $E = V \times V$ is called a complete graph. Complete subgraphs $C \subset V$ are called cliques; the joint subset of two cliques is called a separator S. If a graph G could be partitioned into a sequence of subgraphs $(C_1, S_2, C_2, ..., C_K)$ such that $V = \bigcup_k C_k$ and $S_k = C_{k-1} \bigcap C_k$ are complete for all $k = 1, ..., K$, G is called a decomposable graph (Lauritzen, 1996). In this section, we consider the decomposable graphs. For a covariance matrix Σ, let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or the precision matrix. Nodes i and j are conditionally independent, given other nodes, if and only if $\Omega_{ij} = 0$. Thus, the undirected graph G is given by the configuration of nonzero off-diagonal elements of Σ: $E = \{(i, j) : \Omega_{ij} \neq 0\}$.

Let $M(G)$ be the set of all symmetric positive-definite matrices Σ satisfying $E = \{(i, j) : \Omega_{ij} \neq 0\}$. Given a decomposable graph $G = (V, E)$, Dawid and Lauritzen (1993) introduced the HIW distribution for a covariance matrix $\Sigma \in M(G)$, with parameters (δ, Φ), denoted by $\Sigma \sim \text{HIW}(G, \delta, \Phi)$. The probability density function (pdf) is given by

$$p(\Sigma|G, \delta, \Phi) = \frac{\prod_{k=1}^{K} p(\Sigma_{C_k}|\delta, \Phi_{C_k})}{\prod_{k=2}^{K} p(\Sigma_{S_k}|\delta, \Phi_{S_k})},$$

where $\delta \in \mathbf{R}^+$ is a degree-of-freedom parameter, $\Phi \in M(G)$ is a symmetric positive-definite scale matrix, and C_k and S_k are the cliques and separators of the graph G respectively. The terms $p(\Sigma_{C_k}|\delta, \Phi_{C_k})$ denote the inverse Wishart (IW) density of $\Sigma_{C_k} \sim \text{IW}(\delta, \Phi_{C_k})$ with the pdf

$$p(\Sigma_{C_k}|\delta, \Phi_{C_k}) \propto |\Sigma_{C_k}|^{-(\delta/2+|C_k|)} \exp\left\{-\frac{1}{2} tr(\Sigma_{C_k}^{-1}\Phi_{C_k})\right\}.$$

The HIW distribution is a conjugate prior distribution for the covariance matrix $\Sigma \in M(G)$. Specifically, if q-dimensional random variables $\mathbf{X_i}$ follow an independent and identical (iid) multivariate normal distribution $\mathcal{N}_q(\mathbf{0}, \Sigma)$ for $i = 1, \ldots, n$, and Σ follows $\text{HIW}(G, \delta, \Phi)$, the posterior of Σ is $\Sigma|X, G \sim \text{HIW}(G, \delta + n, \Phi + X'X)$. The closed form of the posterior distribution for Σ plays a key part in the posterior inference based on an MCMC algorithm.

2.2.2 Hierarchical Model

To facilitate computation and notation, we reparameterize $\log(\lambda_{ij})$ as θ_{ij} throughout the rest of the section. We assume the complete hierarchical model for a discrete gene expression dataset as follows:

$$\begin{aligned} X_{ij} &\sim Pois(\lambda_{ij}), & (2.1)\\ \boldsymbol{\theta_i} &\sim \mathcal{N}_q(\boldsymbol{\mu}, \Sigma)., & (2.2)\\ p(\boldsymbol{\mu}|\Sigma, G) &\propto \text{constant}, & (2.3)\\ \Sigma|\delta, r, G &\sim \text{HIW}(G, \delta, rI_q), & (2.4)\\ r &\sim \text{Unif}(0, c), & (2.5)\\ G &\sim \pi(G), & (2.6) \end{aligned}$$

where δ, r, and c are fixed, positive hyperparameters and I_q is a $q \times q$ identity matrix. Equations (2.3) and (2.4) specify the prior for the mean and covariance matrix, respectively. We assume an improper constant prior for $\boldsymbol{\mu}$, as our focus is on the structures of Σ^{-1}. The prior for Σ is HIW(G, δ, Φ) as described in Section 2.2.1. We restrict the graph G of Σ to be decomposable so that the prior for Σ is a mixture of HIW distributions over all decomposable graphs. We consider $\delta = 3$ as reflecting the lack of prior information on Σ, and specify the hyperparameter Φ as rI_q, where r is assumed to follow a uniform hyperprior on the interval $(0, c)$ as in equation (2.5) for some large value of c.

Notice that given the priors for $\boldsymbol{\mu}$ and Σ as specified above, we can integrate out $\boldsymbol{\mu}$ and Σ and obtain a marginalized prior on θ given the graph G as

$$p(\theta|G, \delta, r) \propto \frac{h(G, \delta, rI_q)}{h(G, \delta + n - 1, rI_q + S_\theta)},$$

where $S_\theta = \sum_{i=1}^{n}(\boldsymbol{\theta_i} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta_i} - \bar{\boldsymbol{\theta}})'$. The term $h(G, \delta, rI_q)$ is the normalizing constant for the HIW(G, δ, rI_q) distribution given by

$$h(G, \delta, rI_q) = \frac{\prod_{k=1}^{K} |\frac{rI_{C_k}}{2}|^{(\frac{\delta+|C_k|-1}{2})} \Gamma_{|C_k|}\left(\frac{\delta+|C_k|-1}{2}\right)^{-1}}{\prod_{k=2}^{K} |\frac{rI_{S_k}}{2}|^{(\frac{\delta+|S_k|-1}{2})} \Gamma_{|S_k|}\left(\frac{\delta+|S_k|-1}{2}\right)^{-1}},$$

where $\Gamma_q(x) = \pi^{q(q-1)/4} \prod_{j=1}^{q} \Gamma(x + (1-j)/2)$ is the multivariate gamma function. The marginalized prior leads to a collapsed Gibbs algorithm in sampling G, which substantially accelerates the graphical model search task and is valued when the graph G is our focus in the inference.

We induce the prior $\pi(G)$ in equation (2.6) by assigning an independent prior probability of an edge, $p(e_{ij})$, to each pair of nodes (i, j), so that $\pi(G) = \prod_{(i,j)\in E} p(e_{ij} = 1) \cdot \prod_{(i,j)\notin E} p(e_{ij} = 0)$. Without prior information, a choice of $p(e_{ij})$ could be the Bernoulli-Beta hierarchical prior. Scott and Berger (2010) showed that when the hyperparameters of the Beta distribution is $(1, 1)$, the marginalized prior probability of a graphical model containing k edges out of $q(q-1)/2$ potential edges in a graph G is $p(k) \propto \binom{q(q-1)/2}{k}^{-1}$. Hence, such choice of prior encourages sparsity in the inferred graphical models. In the context of gene expression network modeling as in the section, we borrow information on relatedness between genes based on biological studies and derive a prior $\pi(G)$ from the ontology terms associated with the genes of interest.

2.2.3 GO-based Prior for G

As mentioned above, GO terms describe gene product characteristics in a controlled vocabulary. A pair of genes with the same or closely related ontology terms in biological process are thought to be potentially associated in signaling pathway or expression regulation. We measure the relatedness of all pairs of genes in terms of their associated ontology terms and derive the priors $p(e_{ij})$ that are proportional to the measurements. Here we use the functional semantic similarity as a measure of the relatedness of two genes.

The semantic similarity measures the similarity of two GO terms by evaluating how much information the two terms share. Here we use Wang et al.'s measure (Wang et al., 2007), which is based on the relative locations of the terms in the DAG structure of the GO graph and their semantic relations with the ascendant terms that subsume the two terms. For a GO term A, let T_A denote the set of all its ancestor terms including term A itself, and $S_A(t)$ be defined as the contribution of a term $t \in T_A$ to the semantics of A based on the relative locations of t and A in the GO graph. The semantic similarity score between two GO terms (A,B) is defined as follows:

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} \{S_A(t) + S_B(t)\}}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)},$$

which is within $(0,1)$. Usually one gene is annotated by many GO terms. The functional similarity between two genes G_1 and G_2, $Sim(G_1, G_2)$, is then calculated by averaging the semantic similarity scores for all pairs of their associated terms. The functional similarity score between any two genes (G_i, G_j) is within $(0,1)$, where a value close to 0 indicates the two genes unlikely to be related and a value near 1 indicates close relatedness of the two genes in cellular functioning. Hence, we consider the score as a natural prior probability of an edge between the two genes, i.e. $p(e_{ij} = 1) = Sim(G_i, G_j)$.

We derive the prior for the graph $G = (V, E)$ based on the GO similarity scores as:

$$\begin{aligned}
\pi(G) &= \prod_{(i,j) \in E} p(e_{ij} = 1) \prod_{(i,j) \notin E} p(e_{ij} = 0), \\
&= \prod_{(i,j) \in E} Sim(G_i, G_j) \prod_{(i,j) \notin E} \{1 - Sim(G_j, G_j)\}.
\end{aligned}$$

With the above prior on the graph G, we actually assign a prior probability to the existence or absence of each edge in the graph. As a high similarity score between two genes reflects their potential relatedness in gene regulation, the specified prior favors the graph model that includes edges between semantically similar genes.

2.3 Model Selection Using False Discovery Rates

The posterior sampling schemes we have outlined explore the model space and result in MCMC samples of graph G at each iteration. One method of summarizing the information in the samples is to pick the graph model that is visited mostly by the sampler. However, this particular graph may only appear in a very small proportion of MCMC samples. An alternative strategy is to utilize all of the MCMC samples and average over the various models visited by the sampler. This model averaging approach weighs the evidence of significance for each edge separately using all MCMC samples. We outline an approach to conduct Bayesian model selection based on controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

Suppose we have T posterior samples of graph $G = (V, E)$ from an MCMC computation, which are represented as T sets of edge indicators $\{e_{ij}^{(t)} : i < j\}$. If $(i, j) \in E^{(t)}$ we have $e_{ij}^{(t)} = 1$; else $e_{ij}^{(t)} = 0$. Let p_{ij} represent the posterior probability of including the edge (i, j) in the graph. We can estimate p_{ij} to be the relative number of times the edge (i, j) is present in the graph across the T MCMC samples:

$$p_{ij} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}\{e_{ij}^{(t)} = 1\},$$

where $\mathcal{I}(\,\cdot\,)$ is an indicator function.

We assume that for some significance threshold ϕ, any edge (i, j) with $p_{ij} > \phi$ is considered as significant and is included in the graph G. Then the graph with $E = \{(i, j) : p_{ij} > \phi\}$ includes all the edges considered to be significant. Note that $(1 - p_{ij})$'s can be interpreted as estimates of the local FDRs (Storey, 2003) as they measure the probability of a false positive if an edge (i, j) is significant but not in the true graph. The significance threshold ϕ can be determined based on classical Bayesian utility considerations, such as in Müller et al. (2004), based on the elicited relative costs of false positive and false negative errors, or can be set to control the overall average Bayesian FDR. (See Morris et al., 2008; Baladandayuthapani et al., 2010; and Bonato et al., 2011 for detailed expositions in other settings).

Thus given a global FDR bound $\alpha \in (0, 1)$, we are interested in finding the threshold value ϕ_α for flagging the set of edges $\{(i, j) : p_{ij} > \phi\}$ as potentially relevant and labeling them as *discoveries.* This implies that the threshold ϕ_α is a cut-off on the (model-based) posterior probabilities that corresponds to an expected Bayesian FDR of α, which means that $100\alpha\%$ of the edges identified as discoveries are expected to be false positives. The threshold ϕ_α is determined in the following way: For all $(i, j) : i < j$, we sort p_{ij} in descending order to yield $p_{(k)}, k = 1, ..., q(q-1)/2$. Then, $\phi_\alpha = p_{(\xi)}$, where $\xi = max\{(k^*) : \sum_{k=1}^{k^*}(1 - p_{(k)})/k^* \le \alpha\}$. The set of edges $\{(i, j) : p_{ij} > \phi_\alpha\}$ can be claimed to be positive in the graph based on an average Bayesian FDR of α.

2.4 Simulation Study

In this section, we conduct a simulation study to examine the performance of our method. We set $q = 25$ and consider three scenarios that portray different complexity levels of the networks in generating data. We assume that the discrete data matrix X is generated from the model,

$$\begin{aligned} X_{ij} &\sim Pois(e^{\theta_{ij}}), \\ \boldsymbol{\theta_i} &\sim \mathcal{N}_q(\boldsymbol{\mu}, \Sigma), \end{aligned}$$

where $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{iq})'$. The covariance matrix, Σ, or its corresponding precision matrix, $\Omega = \Sigma^{-1}$ is specified as follows:

- Model 1: We assume that Σ is the covariance matrix of a Gaussian AR(1) process with the element $\sigma_{ij} = 0.7^{|i-j|}$. The specification of Σ corresponds to a band-diagonal precision matrix Ω of bandwidth 1, where only 4% of the off-diagonal elements are zero.

- Model 2: We assume that Σ is the covariance matrix of a Gaussian AR(4) process with the element of $\Omega = \Sigma^{-1}$ to be $\omega_{ij} = 2I\{|i-j|=0\} - 0.5I\{|i-j| = 1\} - 0.8I\{|i-j|=2\} + 0.2I\{|i-j|=3\} + 0.3I\{|i-j|=4\}$. The precision matrix Ω is a band-diagonal matrix of bandwidth 4, where 30% of the off-diagonal elements are zero.

- Model 3: The true decomposable graph G is specified such that about 15% of the off-diagonal elements in the corresponding precision matrix are set to be zero. The true Σ is then generated from the HIW distribution HIW$(G, 3, \Phi)$ conditional on the graph G, where Φ is an arbitrary positive definite matrix.

The configurations of the nonzero off-diagonal elements in the precision matrices as specified in models 1, 2, and 3 are shown in Figure 2.1. For each model, datasets

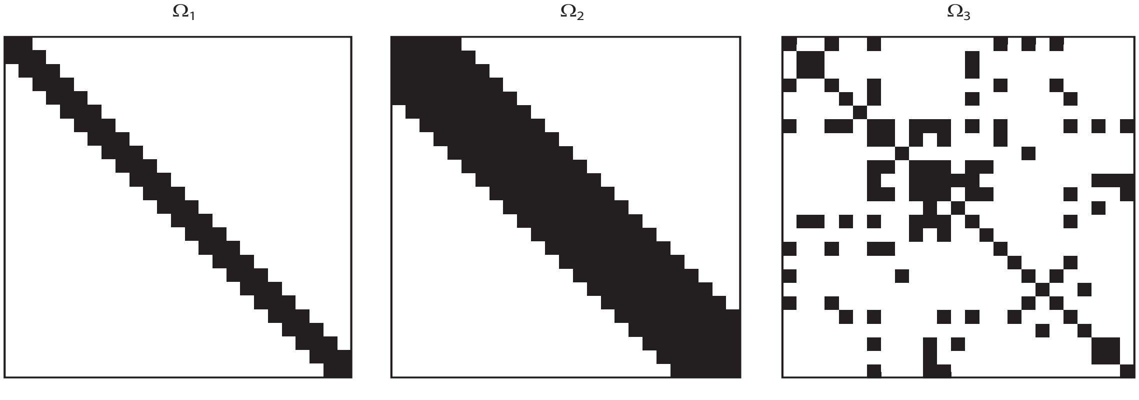


Fig. 2.1. The zero patterns of the precision matrices specified in the models 1 (left), 2 (middle), and 3 (right) in the simulation study. Zero entries in the precision matrices are indicated in white color, while nonzero entries are in black.

are generated of three sample sizes $n = 25, 50$, and 100. Our proposed Bayesian model for discrete data is used to estimate the network structure based on the posterior samples of 15,000 iterations after 5,000 burn-in iterations. The prior for the graph $\pi(G)$ is taken to be constant since no biological GO information is available for simulated data.

For comparison, we transform the simulated data into log ratios, and assume a Gaussian graphical model for the log-transformed data with the hierarchy:

$$
\begin{aligned}
\log(\mathbf{X_i}) &\sim \mathcal{N}_q(\boldsymbol{\mu}, \Sigma).,\\
p(\boldsymbol{\mu}|\Sigma, G) &\propto \text{constant},\\
\Sigma|\delta, r, G &\sim \text{HIW}(G, \delta, rI_q),\\
r &\sim \text{Unif}(0, c),\\
G &\sim \pi(G).
\end{aligned}
$$

The above Gaussian graphical model has the same hierarchical priors on $\boldsymbol{\mu}$ and Σ as in our discrete graphical model, but has a different likelihood. The posterior distributions and the MCMC sampling schemes of the parameters $\boldsymbol{\mu}, \Sigma, r, G$ are the same as described in Appendix A except that the $\log(\theta)$ is replaced with $\log(X)$.

To evaluate the performance of the methods, we calculate the true positive rates (TPRs) and the false positive rates (FPRs) defined as

$$
TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{TN + FP},
$$

where TP, FP, TN, and FN denote the number of true positive, false positive, true negative, and false negative edges, respectively. Figure 2.2 shows the plots of TPRs versus FPRs as we vary the decision threshold on the posterior probabilities of edge inclusion, which are called the receiver operating characteristic (ROC) curves, based on one simulation result under each of the three settings and $n = 50$. The figure shows that the ROC curves of the discrete graph models are closer to the upper left corner than those of the Gaussian graph models for the three simulation settings, indicating a better performance of our method in estimating the discrete graphs.

Table 2.1 summarizes the mean numbers of false positive edges and false negative edges over 20 replications. The graph models are selected as described in Section 2.3 based on thresholds corresponding to an FDR of $\alpha = 0.20$. In accordance with the ROC curves, the discrete graph models have significantly fewer false negatives than the Gaussian graph models, suggesting a higher sensitivity of our method to the edges in a graph. As the sample size increases, both the FPRs and the FNRs decrease obviously in our discrete graph models compared with those in the Gaussian graphical models. When the sample size increases to 100, our method estimates the discrete models optimally, especially for the sparse model, model 1, which has FPRs of less than 10% on average and FNRs of 0%.

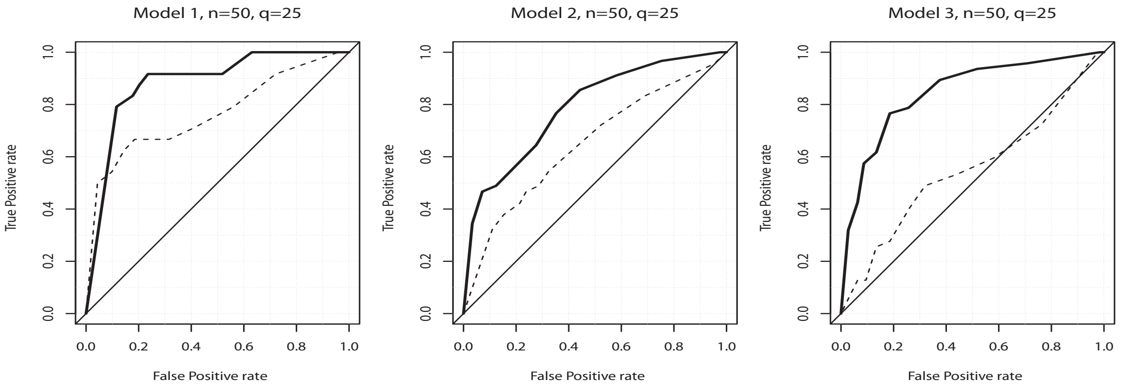


Fig. 2.2. Plots of true positive rates versus false positive rates as the threshold on the posterior probabilities of inclusion in the model is varied (i.e. ROC curves). The curves are based on one simulation run under each of the three scenarios described in Section 2.4 and of the sample size $n = 50$. The solid lines correspond to the ROC curves for the discrete graph models, while the dashed lines for the Gaussian graph models.

2.5 Real Analysis

2.5.1 SAGE Dataset and Pre-Processing

We apply our algorithm in modeling the gene expression network of 25 genes. These genes are identified to be differentially expressed in breast cancer cells by comparison of SAGE expression files followed by statistical tests (Allinen et al., 2004; Porter et al., 2001). The SAGE dataset of these 25 genes is composed of 50 SAGE libraries obtained from carcinoma breast tissue cells. These libraries are publicly available for sharing at the Human SAGE Genie website (http://cgap.nci.nih.gov/SAGE). This website, motivated by the Cancer Genome Anatomy Project (CGAP), is a platform where researchers share their SAGE datasets that are generated from diverse cancer and normal tissues in many laboratories. Sequencing resources vary across laboratories, so each SAGE library has a different total number of tags. As a consequence, the variances of errors are not in the same

Table 2.1
Simulation results under different network reconstruction methods for model 1, 2, and 3. The mean false positive edges and the mean false negative edges over 20 replications are presented in the table with the standard deviations in parentheses. FP: false positive; FN: false negative. See Section 2.4 for details about the models.

		true graph		discrete graph model		Gaussian graph model	
Model	n	No edges	Edges	FP edges	FN edges	FP edges	FN edges
Model 1	25	276	24	71.45 (14.84)	5.45 (1.43)	65.15 (12.98)	9.65 (1.98)
	50	276	24	49.55 (13.33)	0.65 (0.81)	60.70 (10.75)	6.55 (1.79)
	100	276	24	23.95 (5.06)	0.00 (0.00)	50.65 (8.91)	3.10 (1.55)
Model 2	25	210	90	64.00 (10.02)	45.10 (4.74)	54.30 (9.92)	52.95 (5.40)
	50	210	90	48.65 (14.41)	30.35 (4.43)	55.35 (8.66)	46.75 (4.51)
	100	210	90	26.60 (8.21)	11.40 (3.36)	52.05 (12.69)	37.55 (4.13)
Model 3	25	253	47	50.85 (10.58)	19.95 (2.80)	52.50 (11.47)	30.80 (4.11)
	50	253	47	43.60 (10.56)	14.70 (2.25)	48.30 (10.17)	26.60 (3.94)
	100	253	47	26.80 (8.47)	9.80 (2.86)	39.70 (8.35)	19.05 (3.91)

scale. Hence, we normalize the tag frequencies in each library by scaling them so that the total numbers of tags are 20,000 in all libraries.

2.5.2 The Inferred Gene Expression Network

The functional semantic similarity is calculated for each pair of the 25 genes of interest as discussed in Section 2.2.3 using the Bioconductor package *GOSemSim* (Yu et al., 2010). The semantic similarity scores are obtained for biological process GO terms for each pair of genes. Figure 2.3 shows the intensities of the calculated priors on the edges between all pairs of genes in a color map, where the darkness of a lattice is proportional to the prior probability of including the edge in a graph. The prior on a graph is then assigned as the product of the prior probabilities as discussed in Section 2.2.3.

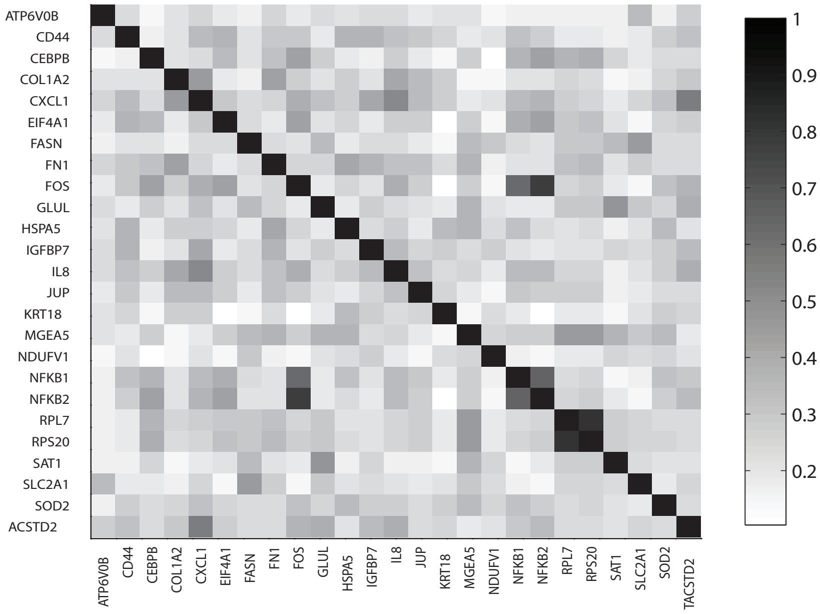


Fig. 2.3. The color map displaying the GO-derived prior probabilities on edges. The probabilities are indicated by the color of the rectangles: the darker the color, the closer is the prior probability to 1 for the corresponding edge.

The proposed Bayesian discrete graphical model is applied to the discrete SAGE data to estimate the network structure and model selection is based on the posterior samples of 50,000 iterations after 10,000 burn-in iterations. The Bayesian estimates of the graphs are obtained by averaging each edge separately and graph models are selected based on an FDR of $\alpha = 0.20$. The resulting gene network is displayed in Figure 2.4 (a), which is partially supported by biological studies.

The genes NFKB1 and NFKB2 encode the subunits of a transcription factor NF-κB, which is known in biological research to stimulate the expression of genes involved in a wide variety of biological functions. The target genes of NF-κB include CD44, COL1A2, CXCL1, FOS, FN1, HSPA5, SAT1, and TACSTD2, which are identified in our inferred network. The promoters of these target genes are all found to contain the NF-κB binding site for transcriptional regulation. Biological studies also find that TACSTD2, the gene encoding a cell surface receptor, transduces an

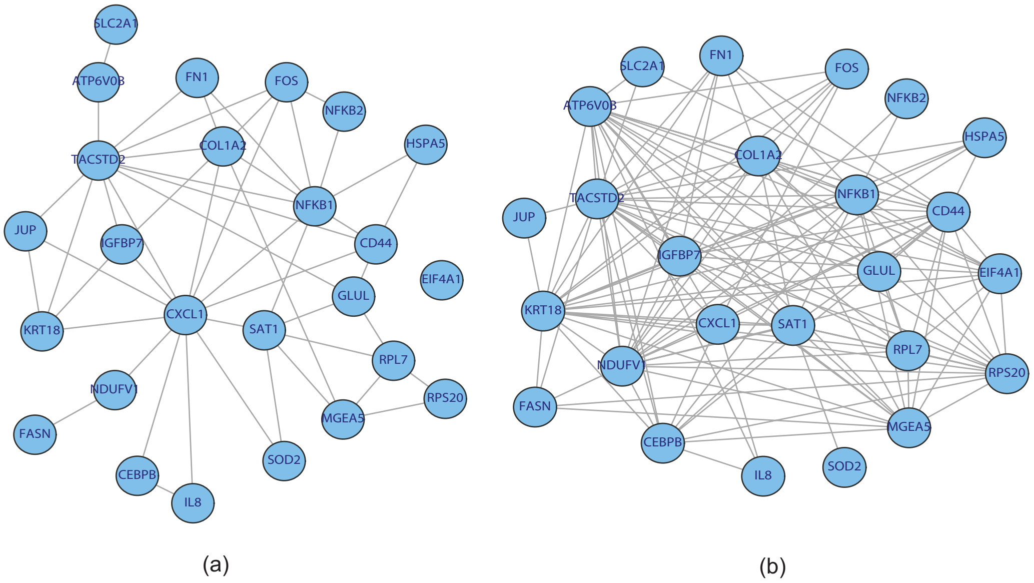


Fig. 2.4. The inferred networks in real SAGE data analysis using the discrete graphical model. (a) The prior for the graph G is derived from GO-based functional semantic similarities. (b) The prior for G is constant for all graphs.

intracellular calcium signal and contributes to tumor pathogenesis by activating the ERK/MAPK pathway (Cubas et al., 2010). This discovery is consistent with the constructed gene network in which TACSTD2 is connected to FOS, NF-κB, and IGFBP7, which are involved in the ERK/MAPK pathway.

The gene COL1A2 encodes one of the chains for type I collagen, the major component of extracellular matrix in skin and other tissues. Research in biology has found that the synthesis of the chain is highly regulated by different cytokines and transcription factors at the transcriptional level. These protein factors involved in COL1A2 regulation include AP1, a family of transcription factors containing the protein product of FOS; NF-κB, the protein product of NFKB1/2; ERK1/2, which can be activated by TACSTD2; and IGFs, which interact with IGFBP7-encoded proteins (Ghosh, 2002). The protein product of FN1 gene is also known to interact with the chain of type I collagen encoded by COL1A2 (Sipes et al., 1993). These findings directly support the dependency between COL1A2 and its neighbors in the inferred gene network. Some other biological discoveries also agree with our network such as the transcriptional regulation of IL-8 by CEBPB (Weber et al., 2003) and SOD2-related oxidative stress induced expression of CXCL1 (Wu et al., 2009).

To test the sensitivity of the network inference to the GO-derived prior probabilities, we reanalyze the real data without the priors induced based on GO information, i.e. the prior of each graph is constant. The resulting network as shown in Figure 2.4 (b) is much more complex than the GO-based network, including more than 100 edges between the 25 genes.

2.6 Discussion

In this section, we extend the Bayesian graphical model for gene networks to discrete expression data from SAGE experiments. We model the count data of mRNA transcripts with independent Poisson distributions, and assume that the log ratios of the Poisson means follow a multivariate normal distribution, whose inverse covariance matrix gives the conditional independence structure of the gene network. We utilize a conjugate prior for the covariance matrix and a collapsed Gibbs sampling algorithm for a fast graphical model search. In addition, we incorporate biological information on genes in our algorithm, by measuring the GO-based semantic similarity between each pair of genes as the prior for a graph. The derivation of GO-based priors is rooted in the biological characteristics of gene regulation. Regulation of gene expression usually occurs between two genes involved in the same metabolic or signaling pathways. Hence, two genes with unrelated gene functions are unlikely to have a direct regulation relationship.

Simulation studies show that our method of modeling discrete gene expression data estimates the network structures with lower FNRs and FPRs than the Gaussian graphical models that are applied to the log-transformed data. In addition, simulation results show that our discrete graph model performs obviously better as sample size increases, and leads to optimal predicted models for moderate sample-size/dimension ratios (=4). We also apply this algorithm to a real SAGE dataset of 25 genes. We show in the result that the derived gene network model with our method agrees with some discoveries in the traditional biological research, which partially supports our model.

3. BAYESIAN HIERARCHICAL STRUCTURED VARIABLE SELECTION METHODS WITH APPLICATION TO MIP STUDIES IN BREAST CANCER

3.1 Introduction

3.1.1 Molecular Inversion Probe-based Arrays for Copy Number Measurement

Changes in chromosomal copy numbers have been identified as important causes of cancer (Pinkel and Albertson, 2005). Chromosomal copy number alteration (CNA) can lead to over-expression of pro-oncogenes or silence of tumor suppressor genes, and affect cellular functions in cell division or programmed cell death (Guha et al., 2008). The accumulation of these DNA errors will eventually influence the development or progression of carcinogenesis; hence, chromosomal copy number analysis has the potential to elucidate tumor progression and identify genetic markers for cancer diagnosis and treatment. CNAs, as gains and losses, are frequent events in breast tumors and occur in patterns that are thought to distinguish genetic paths to tumorigenesis and influence the clinical behavior of the disease (Rennstam et al., 2003; van Beers and Nederlof, 2006).

Many techniques have been developed for the genome-wide detection of CNAs, such as array-based comparative genomic hybridization (CGH), bacterial artificial chromosome CGH, and oligonucleotide array-based CGH (Pinkel et al., 1998; Iafrate et al., 2004; Lucito et al., 2003). A technique that has recently been used for the measurement of allele copy numbers is the molecular inversion probe (MIP) (Hardenbol et al., 2003; Wang et al., 2007). Unlike other CNA measuring techniques such as the CGH methods, the MIP assay requires sequences at the ends to bind genomic DNA simultaneously and utilizes enzymatic steps (ligation) to capture specific loci. The circularization method ensures a high degree of specificity in identifying the loci of interest and reduces cross-talk between probes. The MIP assay generates genotype data as well as copy numbers, which can be used for sample tracking and data quality assessment. Most banked samples with clinical follow-up data are from formalin-fixed, paraffin-embedded (FFPE) tissues, which generally show degraded DNA in the cells. Wang et al. (2007) show that the MIP technology, which requires only a small (∼40bp) target binding site, is more accurate in measuring probe copy numbers for degraded FFPE-derived DNA. Other advantages of the MIP assay are the low amount of DNA sample required, its high levels of multiplexing, and its reproducibility. Refer to Hardenbol et al. (2003) and Wang et al. (2007) for more detailed descriptions of the MIP assay.

In this section, we focus on the analysis of a novel high-dimensional MIP dataset from 971 samples of early-stage breast cancer (stages I and II) collected through the Specialized Programs of Research Excellence (SPORE) in breast cancer at the University of Texas MD Anderson Cancer Center for the purpose of improving risk prediction for disease recurrence. The dataset includes full genome quantifications for 330,000 MIPs from tumor cells of patients using high-density Oncoscan$^{\text{TM}}$arrays from Affymetrix$^{\text{TM}}$. A detailed description of the dataset with regard to data collection, pre-processing and normalization is provided in Section 3.6.1. Briefly, the resulting (normalized) data for downstream statistical analysis consist of the $\log_2$ intensity ratios of the copy numbers in test samples to the copy numbers in normal reference cells for all probes. Hence, for a cell sample with the normal probe copy number $(= 2)$, the normalized value is $\log_2(2/2) = 0$; for a probe with a gain of measured copy numbers (> 2) the log ratio is positive, and for a probe with a loss of copy numbers (< 2) the log ratio is negative. The magnitudes of intensity ratios in the positive and negative direction are indicative of multiple probe-level gains and losses, respectively.

In addition to the MIP copy number profiles, we have data on a number of relevant clinical outcomes from these patient samples, such as the clinical subtype of breast cancer defined by tumor markers (i.e., hormone receptor status, HER2 status, Ki67), tumor size, and lymph node status, as well as other clinicopathological characteristics such as age, stage, tumor grade, histology, and type of treatment (Thompson et al., 2011). Our main focus in this section is to identify MIPs that are significantly associated with the clinical and pathologic characteristics of the tumors with an emphasis on clinical subtypes. Discovering and validating CNAs that correlate with tumor characteristics will identify regions of high interest for further investigation as clinically useful diagnostic and treatment biomarkers.

We assume that many of the acquired chromosomal events act jointly in mediating the biological effects. Thus it is of high interest to model the joint effects of CNAs detected using the MIP probes and discover regions of the genome that exhibit significant associations – in contrast to univariate single MIP analysis, which might miss regions with weak marginal but important joint effects on the clinical outcomes. However, inferential challenges for the MIP copy number dataset include not only its high-dimensionality but also that markers tend to be spatially correlated because the MIPs are indexed by genomic location. We propose a novel structured "hunting" approach to identify areas of the genome that are significantly associated with clinically relevant outcomes. We follow the two-level hierarchical structure induced by biology: a gene level and MIP-within-gene level architecture. Thus we group contiguous MIPs (as per their genomic location) by their unique gene annotation and treat the genes (group) as the first level of the hierarchy and the MIPs within the genes (subgroup) as the second level of the hierarchy.

To illustrate our main idea, in Figure 3.1 we show an example plot of the partial MIP copy number profile for a randomly selected patient sample, where the x-axis is the genomic location and each vertical line is the log-intensity ratio for an MIP probe. The different line type patterns correspond to the gene groups, indicating the uniquely annotated gene structures on the chromosome. There are several features exemplified in the plot. There exists substantial variability both within and between the genes, primarily due to different numbers of probes mapped to each gene and different probes within the same gene contributing differently, both positively and negatively. Also there exists serial correlation between the probes within the same gene, given their proximity by genomic location. Biologically, our gene-centric selection approach is of more interest to our scientific collaborators since there exists substantial knowledge about genes as functional units and the analytic result is more interpretable in terms of a medical diagnosis. In addition, for a specific gene, different probes may confer different factors to the gene's function and thus it is of equal interest to identify predictive probes within a selected gene. Therefore, in our study, we want to select both genes and within-gene probes that are significantly associated with clinically relevant outcomes – leading to a statistical formulation of *hierarchical structured variable selection.*

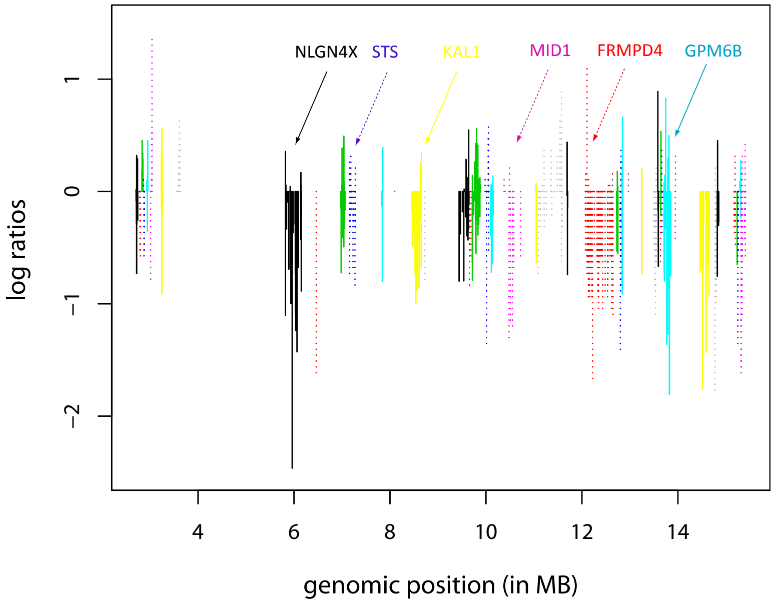


Fig. 3.1. Copy number profile from a tumor sample. The log-ratios are plotted on the vertical axis against their genomic position (in MB). The line type patterns indicate the gene structures on the chromosome.

3.1.2 Relevant Statistical Literature

Variable selection is a fundamental issue in statistical analysis and has been extensively studied. Penalized methods such as the bridge regression (Frank and Friedman, 1993), the lasso regression (Tibshirani, 1996), the SCAD regression (Fan and Li, 2001), the LARS regression (Efron et al., 2004) and the OSCAR regression (Bondell and Reich, 2008) have been proposed due to their relatively stable performance in model selection and prediction. The lasso method has especially gained much attention. It utilizes an L_1-norm penalty function to achieve estimation shrinkage and variable selection. In a Bayesian framework, the variable selection problem can be viewed as the identification of nonzero regression parameters based on posterior distributions. Different priors have been considered for this purpose. Mitchell and Beauchamp (1988) propose a "spike and slab" method that assumes the prior distribution of each regression coefficient to be a mixture of a point mass at 0 and a diffuse uniform distribution elsewhere; this is extended by George and McCulloch (1993; 1997), Kuo and Mallick (1998), and Ishwaran and Rao (2005) in different settings. Other methods specify absolutely continuous priors that approximate the "spike and slab" shape, shrinking the estimates toward zero (Xu, 2003; Bae and Mallick, 2004; Park and Casella, 2008; Griffin and Brown, 2007; 2010). In particular, Park and Casella (2008) extend the frequentist lasso with a full Bayesian method by assigning independent and identical Laplace priors to the regression parameters.

The above mentioned methods ignore the grouping structure that appears in many applications such as ours. The individual-level variable selection methods tend to select more groups than necessary when selection at group level is desired. To accommodate group-level selection, Yuan and Lin (2006) propose the group lasso method, in which a lasso penalty function is applied to the L_2-norm of the coefficients within each group. This method is subsequently extended by Raman et al. (2009) in a Bayesian setting. Zhao et al. (2009) generalize the group lasso method by replacing the L_2-norm of the coefficients in each group with the L_γ-norm for $1 < \gamma \leq \infty$. In the extreme case where $\gamma = \infty$, the coefficient estimates within a group are encouraged to be exactly the same. However, these grouped model selection methods carry out selection only at group level, not at within-group level; that is, they only allow for the variables within a group to be in or out of the model simultaneously. More recently, some frequentist methods have been developed for selection at both group level and within-group level. Wang et al. (2009) reparameterize predictor coefficients and selected variables by maximizing the penalized likelihood with two penalizing terms. Ma et al. (2010) propose a clustering threshold gradient-directed regularization (CTGDR) method for genetic association studies.

In this section, we propose a Bayesian method to perform the variable selection on hierarchically structured data given that the grouping structures are known. We propose a novel hierarchical structured variable selection (HSVS) prior that generalizes the traditional "spike and slab" selection priors of Mitchell and Beauchamp (1988) for grouped variable selection. Specifically, instead of the uniform or multivariate normal distribution of the traditional "spike and slab" methods, we let the "slab" part in the prior be a general robust shrinkage distribution such as a Laplace distribution, which leads to the well-developed lasso-type penalization formulations. Unlike other group selection methods, which usually utilize lasso penalties for group-level shrinkage and selection, our proposed prior uses selection priors for group-level selection that are combined with a Laplace "slab" to obtain Bayesian lasso estimates for within-group coefficients, thus achieving group selection and within-group shrinkage simultaneously. More advantageously, because the full conditionals of the model parameters are available in closed form, this formulation allows for efficient posterior computations, which greatly aid our analysis of high-dimensional datasets. Using full Markov chain Monte Carlo (MCMC) methods, we can obtain the posterior probability of a group's inclusion, upon which posterior inference can then be conducted using false discovery rate (FDR)-based methods, which are crucial in high-dimensional data. Our method thresholds the posterior probabilities for group selection by controlling the overall average FDR while within-group variable selection is conducted based on the posterior credible intervals of the within-group coefficients obtained from the MCMC samples. Furthermore, we propose extensions to account for the correlation between neighboring coefficients within a group by incorporating a Bayesian fused lasso for within-group variable selection. Due to the conjugate nature of model formulation, our method could also be easily extended to nonlinear regression problems for discrete response variables.

The rest of Section 3 is organized as follows. In Section 3.2 we propose our hierarchical Bayesian models for simultaneous variable selection at both group and within-group levels. In Section 3.3, we extend the Bayesian models for variable selection of generalized linear models. In Section 3.4, we show the FDR-based methods for group selection. Simulation studies are then carried out and discussed in Section 3.5. We apply the models to the real MIP data analysis in Section 3.6 and conclude with a discussion in Section 3.7. The technical details including the full conditional distributions and the posterior sampling algorithm are described in Appendix B.

3.2 Probability Model

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denote the clinical outcomes/responses of interest from n patients/samples and X denote the $n \times q$-dimensional covariate matrix of q probes from MIP measurements. For ease of exposition we present the model for the Gaussian case here and discuss generalized linear model extensions for discrete responses in Section 3.3. The model we posit on the clinical outcome is

$$\mathbf{Y} = U\mathbf{b} + X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where U denotes the fixed effects of non-genetic factors/confounders such as age at diagnosis, tumor size, and lymph node status with associated parameters $\mathbf{b}$. We further assume that the data matrix X and the coefficients $\boldsymbol{\beta}$ are known to be partitioned into G groups/genes, where the g^{th} group contains k_g elements for $g = 1, ..., G$. We assume that a given probe occurs in only one gene (group), which is trivially satisfied for these data since the probes are grouped by genomic location and mapped to a uniquely annotated gene. Thus, we write $X = (X_1, ..., X_G)$, with $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_G})$ denoting the group-level coefficients and $\boldsymbol{\beta_g} = (\beta_{g1}, ..., \beta_{gk_g})$ denoting the within-group coefficients. The error terms $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are assumed to be independently and identically distributed $\mathcal{N}(0, \sigma^2)$ for the Gaussian responses. Our key construct of interest is the q-dimensional coefficient vector $\boldsymbol{\beta}$, which captures the association between the probe measurements and the clinical outcome. Hereafter we propose a novel hierarchical prior construction based on the natural hierarchical structure of the probe measurements that simultaneously selects relevant genes and significant probes-within-genes. We present the independent case first, wherein we assume the within-group coefficients are independent and subsequently extend the method in Section 3.2.2 to account for within-group correlations.

3.2.1 Hierarchical Structured Variable Selection Model

At the group level, we employ a "selection" prior and introduce a latent binary indicator variable γ_g for each group g with the following interpretation: when $\gamma_g = 0$, the coefficients $\boldsymbol{\beta}_g$ of the g^{th} group have a point mass density at zero, reflecting that the predictors in the g^{th} group are not selected in the regression model; conversely, when $\gamma_g = 1$, the g^{th} group is selected in the model. At the within-group level, we assign a robust "shrinkage" prior and use the scale mixture normal distribution

(Andrews and Mallows, 1974; West, 1987) for each element in $\boldsymbol{\beta}_g$ conditional on $\gamma_g = 1$. Our hierarchical formulation of the prior can be succinctly written as

$$\begin{aligned}
\boldsymbol{\beta_g}|\gamma_g, \sigma^2, \boldsymbol{\tau_g^2} &\sim (1-\gamma_g)\delta_{\{\boldsymbol{\beta_g}=\mathbf{0_{k_g}}\}} + \gamma_g \mathcal{N}_{k_g}(\mathbf{0_{k_g}}, \sigma^2 D_{\tau_g}), \\
&\quad \text{where } D_{\tau_g} = \text{diag}(\ \tau^2_{\text{g1}}, ..., \tau^2_{\text{gk}_\text{g}}), \\
\gamma_g|p &\sim \text{Bernoulli}(p), \\
\tau^2_{gj}|\lambda_g &\sim \mathcal{G}(\bullet), \qquad (3.1)
\end{aligned}$$

where $\delta_\bullet$ represents the Dirac delta measure that places all its mass on zero, τ_{gj}'s are the Gaussian scaling parameters of the "slab" distribution, and $\mathcal{G}(\bullet)$ is a general mixing distribution. By setting $\mathcal{G}$ to different mixing distributions, various shrinkage properties can be obtained. In this dissertation, we let $\mathcal{G}(\bullet)$ be an exponential distribution, $\tau^2_{gj}|\lambda_g \sim \text{Exp}(\lambda_g^2/2)$, with a rate parameter, λ_g for g^{th} group. This prior leads to well-developed lasso formulations with a (group-specific) penalty/regularization parameter λ_g for the g^{th} group. Other formulations are possible as well, such as the normal-exponential-gamma prior of Griffin and Brown (2007) and normal-gamma prior of Griffin and Brown (2010), by using other families of scaling distributions. We call our prior in (2.1) the hierarchical structured variable selection (HSVS) prior, which has the following properties: (1) It generalizes the spike and slab mixture priors of Mitchell and Beauchamp (1988) to grouped settings, and accommodates robust shrinkage priors for the slab part of the prior replacing the uniform slab. (2) The within-group shrinkage follows the well-developed lasso formulation, which promotes sparseness within selected groups and automatically provides interval estimates for all coefficients. (3) The hierarchy allows for the simultaneous selection and shrinkage of grouped covariates as opposed to all-in or all-out group selection (Yuan and Lin, 2006) or two-stage methods (Ma et al., 2010; Wang et al., 2009). (4) Most importantly, it is computationally tractable for large datasets since all full conditionals are available in closed form, which greatly aids our MCMC computations and subsequent posterior inference, as we show hereafter.

In order to gain more intuition regarding this prior, Figure 3.2 (a) shows the schematic plot of an HSVS prior distribution versus a Bayesian group lasso prior distribution. In each plot, the density of the HSVS and the group lasso prior is imposed on a group composed of two individual variables with coefficients β_1 and β_2. The "spike" at zero in the HSVS prior introduces group-level sparsity by simultaneously forcing both variables in the group to zero when β_1 and β_2 are both small in value. The Laplace distribution elsewhere in the prior shrinks individual coefficients within a group toward zero, which in return influences the group selection. In contrast, the Bayesian group lasso prior simultaneously shrinks β_1 and β_2 and does not lead to within-group selection; whereas our HSVS prior results in both group and within-group variable shrinkage and selection, as is evidenced in Figure 3.2 (b) – which shows an example plot of the posterior distribution for the two coefficients in a group with an HSVS and a Bayesian group lasso prior, respectively.

To complete the prior specifications in the Gaussian case, we use a diffuse Gaussian prior $\mathcal{N}(0, cI)$ for the coefficients for fixed effects b, where c is some large value. For the parameter p that controls the group level selection, we use a conjugate Beta hyperprior: Beta(a, b) with (fixed) parameters a and b. We estimate the group-specific lasso parameters $\lambda_1^2, ..., \lambda_G^2$ and specify a common gamma mixing distribution Gamma(r, δ), ensuring their positivity. We use the improper prior density $\pi(\sigma^2) = 1/\sigma^2$ on the error variance, which leads to a closed form of the full conditional distribution. These hyperpriors result in conjugate full conditional distributions for all model parameters, allowing for an efficient Gibbs sampler. (See Appendix A in Supplementary Materials for the full conditional distributions and corresponding

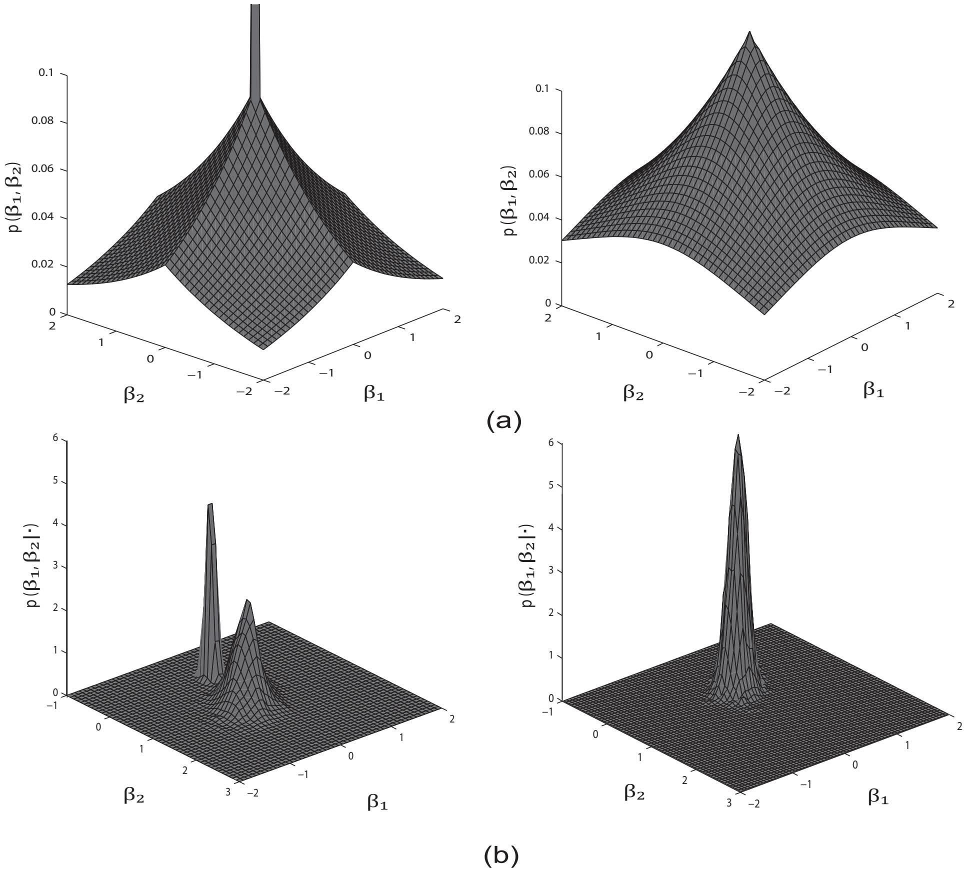


Fig. 3.2. Schematic plot of prior and posterior distribution of the hierarchical structured variable selection (HSVS) method. (a) Left: the density curve of an HSVS prior for a group with two variables; Right: a Bayesian lasso prior for a group with two variables. (b) Left: an example plot of the posterior distribution for a group with two variables when an HSVS prior is applied; Right: an example plot of the posterior distribution for the group of two variables when a Bayesian lasso prior is applied.

Gibbs sampling schemes.) Our full hierarchical model for the HSVS linear model can be succinctly written as

$$\begin{aligned}
\text{Likelihood: } & \mathbf{y}|U, X, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_q(U\mathbf{b} + X\boldsymbol{\beta}, \sigma^2 I_n), \\
\text{Priors: } & \mathbf{b} \sim \mathcal{N}(0, cI), \\
& \boldsymbol{\beta_g}|\gamma_g, \sigma^2, \boldsymbol{\tau_g^2} \sim (1-\gamma_g)\delta_{\{\boldsymbol{\beta_g}=\mathbf{0_{k_g}}\}} + \gamma_g \mathcal{N}_{k_g}(\mathbf{0_{k_g}}, \sigma^2 D_{\tau_g}), \\
& \text{ where } D_{\tau_g} = \text{diag}(\ \tau^2_{\mathrm{g}1}, ..., \tau^2_{\mathrm{gk_g}}), \\
\text{Hyperpriors: } & \gamma_g|p \sim \text{Bernoulli}(p), \quad \tau^2_{gj}|\lambda_g \sim \text{Exp}(\frac{\lambda_g^2}{2}), \quad p \sim \text{Beta}(a, b), \\
& \lambda_g^2 \sim \text{Gamma}(r, \delta), \quad \sigma^2 \sim 1/\sigma^2.
\end{aligned}$$

3.2.2 Fused Hierarchical Structured Variable Selection Model

In the proposed HSVS construction above, we utilize a group-specific binary indicator for group-level selection and a Bayesian lasso method via independent Laplace priors for within-group shrinkage, which is invariant to the permutation of the order of the group-specific variables. In situations where there exists a natural ordering of the variables, such as ours, where the probes within a gene have a natural order with respect to their genomic positions, it might be useful to account for the "serial" structure of such data. For this purpose, we extend our HSVS model by implementing a Bayesian fused lasso for within-group variable selection.

The (frequentist) fused lasso is defined by Tibshirani et al. (2005) using the following penalization framework:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \arg\min\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\}, & \text{ subject to} \\
\sum_{j=1}^{q} |\beta_j| \leq \lambda_1 \text{ and } & \sum_{j=2}^{q} |\beta_j - \beta_{j-1}| \leq \lambda_2,
\end{aligned}$$

where, in comparison with the regular lasso, the fused lasso utilizes two regulation parameters. The first parameter λ_1 encourages sparsity in the coefficient estimation and the second parameter λ_2 reduces the differences between neighboring coefficients, thus encouraging smoothness in the coefficient profiles β_j as a function of j and accounting for the adjacency structure of the data. Kyung et al. (2010) derive the Bayesian version of the fused lasso for non-grouped settings.

In our second proposed prior, the *fused-HSVS*, we incorporate the Bayesian fused lasso in the hierarchical model for within-group variable selection, as follows:

$$\boldsymbol{\beta_g}|\gamma_g, \sigma^2, \boldsymbol{\tau_g^2}, \boldsymbol{\omega_g^2} \sim (1-\gamma_g)\delta_{\{\boldsymbol{\beta_g}=\mathbf{0_{k_g}}\}} + \gamma_g \mathcal{N}_{k_g}\left(\mathbf{0_{k_g}}, \sigma^2 \varSigma_{\beta_g}\right),$$

$$\text{where } \Sigma_{\beta_g}^{-1} = \begin{bmatrix} \frac{1}{\tau_{g1}^2}+\frac{1}{\omega_{g1}^2} & -\frac{1}{\omega_{g1}^2} & \cdots & 0 \\ -\frac{1}{\omega_{g1}^2} & \frac{1}{\tau_{g2}^2}+\frac{1}{\omega_{g1}^2}+\frac{1}{\omega_{g2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\tau_{gk_g}^2}+\frac{1}{\omega_{g(k_g-1)}^2} \end{bmatrix},$$

$$\gamma_g|p \sim \text{Bernoulli}(p),$$

$$\tau_{gj}^2|\lambda_{1g} \sim \text{Exp}(\frac{\lambda_{1g}^2}{2}), \text{ for j} = 1, ..., k_g,$$

$$\omega_{gj}^2|\lambda_{2g} \sim \text{Exp}(\frac{\lambda_{2g}^2}{2}), \text{ for j} = 1, ..., k_g - 1,$$

where τ_{gj}'s are the variances of the individual coefficients within a group and ω_{gj}'s introduce correlations between neighboring coefficients in the prior. By using the exponential hyperpriors with the regularization parameters, λ_{1g}'s and λ_{2g}'s, the hierarchy shrinks the coefficient estimates and reduces the difference in neighboring coefficients.

As with the independent HSVS model, we can assign a beta hyperprior distribution to the parameter p and diffuse gamma hyperprior distributions Gamma(r_1, δ_1) and Gamma(r_2, δ_2) to the two sets of regularization parameters $\{\lambda_{1g} : g = 1, ..., G\}$ and $\{\lambda_{2g} : g = 1, ...G\}$, respectively. We use the same prior parameters for λ_{1g}'s and λ_{2g}'s. However, different values could be used for each set. These choices of hyperprior densities lead to conjugate conditional posterior distributions, which can then easily join the other parameters in the Gibbs sampler. The full hierarchical model with fused within-group priors is formulated as follows:

$$\begin{aligned}
\text{Likelihood: } & \mathbf{y}|U, X, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_q(U\mathbf{b} + X\boldsymbol{\beta}, \sigma^2 I_n), \\
\text{Priors: } & \mathbf{b} \sim \mathcal{N}(0, cI), \\
& \boldsymbol{\beta_g}|\gamma_g, \sigma^2, \boldsymbol{\tau_g^2}, \boldsymbol{\omega_g^2} \sim (1-\gamma_g)\delta_{\{\boldsymbol{\beta_g}=\mathbf{0_{k_g}}\}} + \gamma_g \mathcal{N}_{k_g}\left(\mathbf{0_{k_g}}, \sigma^2 \Sigma_{\beta_g}\right), \\
\text{Hyperpriors: } & \gamma_g|p \sim \text{Bernoulli}(p), \quad \tau_{gj}^2|\lambda_{1g} \sim \text{Exp}(\frac{\lambda_{1g}^2}{2}), \quad \omega_{gj}^2|\lambda_{2g} \sim \text{Exp}(\frac{\lambda_{2g}^2}{2}), \\
& p \sim \text{Beta}(a, b), \quad \lambda_{1g}^2 \sim \text{Gamma}(r_1, \delta_1), \quad \lambda_{2g}^2 \sim \text{Gamma}(r_2, \delta_2), \\
& \sigma^2 \sim 1/\sigma^2.
\end{aligned}$$

3.2.3 Choice of Hyperparameters

For the parameters of the beta prior on p in the HSVS and the fused-HSVS models, we set $(a, b) = (1, 1)$, which is a uniform prior. This choice of prior results in the prior probability of a model containing k groups out of G potential groups being $p(k) \propto \binom{G}{k}^{-1}$ (Scott and Berger, 2010), which encourages sparsity in model selection. More informative choices can be accommodated using appropriate specifications of these parameters. For the gamma priors of λ_g^2 in the HSVS model and of λ_{1g}^2, λ_{2g}^2 in the fused-HSVS model, we consider the shape parameters (r, r_1, r_2) to be 1, as in Kyung et al. (2010) and Park and Casella (2008), such that the prior densities approach 0 sufficiently fast, and we use the empirical Bayes estimator of the rate parameters $(\delta, \delta_1, \delta_2)$. For example, conditional on $r = 1$, the empirical Bayes estimator of δ in the HSVS model is $\delta^{(k)} = \frac{G}{\sum_g \mathbf{E}_{\delta^{(k-1)}}(\lambda_g^2|y)}$ at the k^{th} iteration.

3.3 Generalized Hierarchical Structured Variable Selection Model for Discrete Responses

Due to the conjugate construction of both the HSVS and fused-HSVS models, they can be extended to discrete responses using the latent variable formulations as in Albert and Chib (1993) and Holmes and Held (2006). We present the binary case and note that extensions to multinomial and ordinal responses can be dealt with in a similar manner.

Suppose that n binary responses, $Y_1, ..., Y_n$, are observed and that Y_i has a Bernoulli distribution with probability p_i. We then relate the predictor variables with the responses using a probit regression model

$$Pr(Y_i = 1|\boldsymbol{\beta}) = \Phi(X_i'\boldsymbol{\beta}),$$

where Φ is the normal cumulative distribution function. Following Albert and Chib (1993) we introduce n independent latent variables $Z_1, ..., Z_n$, such that,

$$Y_i = \begin{cases} 1, & Z_i \geq 0; \\ 0, & Z_i < 0, \end{cases}$$

$$Z_i = \mathbf{U}_i\mathbf{b} + \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1),$$

where the prior on $\boldsymbol{\beta}$ parallels the developments in Sections 2.1 and 2.2 with Y_i's replaced by Z_i's, giving rise to our generalized-HSVS model. The generalized-HSVS model leads to a truncated normal for the full conditional distribution of Z_i. Hence Z_i's can easily be embedded in the Gibbs sampling. The posterior distribution and Gibbs sampling of Z_i's are detailed in Appendix B.

3.4 Model Selection Using False Discovery Rates

The posterior sampling schemes we have outlined explore the model space and result in MCMC samples of both the group indicators and the corresponding within-group coefficients at each iteration. The groups that are significant predictors may appear in most of the posterior samples; whereas the others may appear less often. We summarize the information in the samples for conducting model selection by using the false discovery rate (FDR)-based model averaging approach as in Section 2.4.

The detail of the FDR-based model selection method and its theoretical basis are discussed in Section 2.4. Briefly, suppose we have T posterior samples of a parameter set from an MCMC computation. Recall that by our prior structures, for each MCMC iteration, a certain set of variable groups is included in the regression model whose group indicator $\gamma_g^{(t)} = 1$, where $\gamma_g^{(t)}$ is the value of γ_g at the t^{th} MCMC iteration. Let p_g represent the posterior probability of including the g^{th} group in the model, $g = 1, ..., G$, which can be estimated by:

$$p_g = \frac{1}{T}\sum_{t=1}^{T} \mathcal{I}\{\gamma_g^{(t)} = 1\}.$$

We assume that for some significance threshold ϕ, any variable group with $p_g > \phi$ is significant, and thus is included in the regression model. We choose the threshold ϕ in the following way to control the overall FDR at a pre-determined value $\alpha \in (0, 1)$: for all the groups $g = 1, ..., G$, we sort p_g in descending order to yield $p_{(g)}, g = 1, ..., G$. Then, $\phi_\alpha = p_{(\xi)}$, where $\xi = \max\{g* : \sum_{g=1}^{g*}(1 - p_{(g)}/g*) \leq \alpha\}$. Thus, the set of groups $\mathcal{X}_{\phi_\alpha} = \{g : p_g > \phi_\alpha\}$ can be claimed as significant in the regression model based on an average Bayesian FDR of α. For the within-group selection, we select individual variables (conditional on the significant groups) based on the posterior credible intervals of the coefficients.

3.5 Simulation Studies

We conducted two detailed simulation studies to evaluate the operating characteristics of our method in the context of a linear regression model and a probit regression model (closely mimicking our real MIPs data), as presented in Sections 3.5.1 and 3.5.2 respectively.

3.5.1 Simulations for Linear Regression Models

We first assumed a simple linear model,

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and considered five scenarios that portray different aspects of the data generating process, with the following specification of the covariate matrix, X.

- Model I: We first generated twenty-one latent random variables $Z_1, ..., Z_{20}$ and W from independent standard normal distributions. The covariates $X_1, ..., X_{20}$ were defined as $X_i = (Z_i + W)/\sqrt{2}$. We considered 20 variable groups for the regression model, where the i^{th} group, $i = 1, ..., 20$, is composed of all the terms in a fourth-degree polynomial of X_i. The datasets were simulated from the following true model

  $$Y = \underbrace{X_3 + \frac{1}{2}X_3^4} \quad \underbrace{-\frac{1}{2}X_6 + \frac{2}{3}X_6^4} \quad \underbrace{+2X_9 - \frac{3}{2}X_9^3} + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(0, 2^2)$. We collected 100 observations from each run. This model is similar to the settings used in Yuan and Lin (2006), where the predictors have a natural grouping structure. However, our study is different in that not all elements in a group are present in the true models, i.e., some of the within-group coefficients are set to zero. Hence, selections at both group and within-group levels are desired for the model.

- Model II: We generated the covariates $X_1, ..., X_{20}$ as in model I. We then considered 20 variable groups for the regression model, where the i^{th} group, $i = 1, ..., 20$, is composed of all the terms in a fourth-degree polynomial of X_i. However, the data were simulated from a true model with a total of 9 variable groups, each containing only 2 terms of the fourth-degree polynomial. We collected 100 observations from each run. This model has the same setting as model I except for the sparsity level in the true model, with model II being less sparse (having more variables) than model I.

- Model III: We generated twenty latent variables $Z_1, ..., Z_{20}$ independently from a standard normal distribution. We then considered 20 groups for the regression model, with each group composed of four variables, X_{ij} for $j = 1, ..., 4$. X_{ij}'s were generated as $X_{ij} = (Z_i + e_{ij})/\sqrt{2}$, where $e_{ij} \sim \mathcal{N}(0, 1)$. The data were simulated from the true model

$$Y = \underbrace{X_{31} + X_{32} + X_{33}}\quad \underbrace{+\frac{4}{3}X_{61} + \frac{1}{2}X_{62}}\quad \underbrace{+\frac{1}{3}X_{91} - X_{93} - 2X_{94}} + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(0, 2^2)$. We collected 100 observations from each run. In model III, the four candidate variables within the same group are correlated, with a correlation $r = 0.5$; whereas the variables between groups are independent. The true model includes partial elements within three groups. Hence, selections at both group and within-group levels are desired for the model.

- Model IV: We generated twenty latent variables $Z_1, ..., Z_{20}$ independently from a standard normal distribution. We then considered 20 groups for the regression model, with each group composed of four variables, X_{ij} for $j = 1, ..., 4$. X_{ij}'s were generated as $X_{ij} = (Z_i + e_{ij})/\sqrt{1.01}$, where $e_{ij} \sim \mathcal{N}(0, 0.1^2)$. The data were simulated from the same model as in model III. We collected 100 observations from each run. This model has the same setting as model III, except that the variables within the same group have a much higher correlation, $r = 0.99$.

- Model V: We generated ten latent variables $Z_1, ..., Z_{10}$ independently from a standard normal distribution. We then considered 10 groups for the regression model, with each group composed of 10 variables, $X_{ij}, j = 1, ..., 10$. X_{ij}'s were generated in the same fashion as in model III. The data were simulated from the true model

  $$Y = \underbrace{X_{31} + X_{32} + X_{33} + X_{34} + X_{35} + X_{36}} \quad \underbrace{-X_{61} - X_{62} - X_{63} - X_{64}} + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(0, 2^2)$. We collected 100 observations from each run. Thus, model IV includes two predictive groups, each group having a block of constant nonzero coefficients. We use model IV to compare the performance of the HSVS and fused-HSVS method when collinearity between neighboring coefficients is present in a group.

For each dataset generated from models I, II, III, or IV, the HSVS, the group lasso, the regular lasso, and the stepwise selection methods were used to estimate the coefficients. For each dataset generated from model V, the HSVS, the fused-HSVS, and the group lasso methods were used to estimate the coefficients. The Bayesian estimates were posterior medians using $10,000$ iterations of the Gibbs sampler after $1,000$ burn-in iterations. Significant groups were selected based on an FDR of $\alpha = 0.10$. The regular lasso, and the group lasso methods estimated coefficients using the *lars* (Efron et al., 2004) and *grpreg* (Breheny and Huang, 2009) packages respectively, with the tuning parameters selected using C_p-criterion and 5-fold cross validation. To evaluate the performance of each method, we use the true model error defined as

$$ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' X' X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{3.2}$$

Table 3.1 summarizes the average model errors over 200 runs, along with the number of false positive (FP) and false negative (FN) groups/individual variables selected for each method. The results show that the HSVS method has slightly smaller model errors than the group lasso method and significantly smaller model errors than the lasso and stepwise methods for models I, II and III; but it performs no better than the group lasso for model IV, where there are extremely high correlations within groups. For the group-level selection, the HSVS method is similar in performance to the group lasso method. However, the HSVS method has an obviously higher false negative rate than the group lasso when the number of nonzero groups increases, as indicated in model II. For the within-group-level selection, we use the 95% posterior credible intervals based on MCMC samples to select significant variables within the FDR-based significant groups. Table 3.1 shows that the method performs better overall than other methods, with lower false positive rates, although at the price of small false negative rates. This is expected since we use the Bayesian lasso formulation, which shrinks within-group coefficients toward zero. Hence, the model tends to deselect the within-group variables that have only weak effects on the response. In our simulation study, the model has higher probabilities of obtaining false negatives for those variables whose true coefficients are less than 0.5 in absolute value.

The results of the model V estimation show that the fused-HSVS method has lower model errors than the other two. In addition, the fused-HSVS method performs better than the HSVS in within-group-level selection, with both lower false positive and false negative rates. The results indicate that the fused-HSVS method is better when the variables within a group have similar effects on the response. Compared to the HSVS prior, the fused-HSVS prior leads to less varying coefficient estimates within a group, due to the constraint on the differences between neighboring coefficients.

3.5.2 Simulations Based on Real Data

In this section, we present the results from simulations for high-dimensional generalized linear models closely mimicking our real breast cancer data presented in Section 3.6.2. Specifically, we simulated data from a probit regression model,

$$Pr(Y_i = 1|\boldsymbol{\beta}) = \Phi(X_i'\boldsymbol{\beta}), \tag{3.3}$$

where Φ is the cumulative distribution function of a standard normal distribution. We considered 61 groups that were composed of 917 variables, with the grouping structure similar to the gene-probe architecture of the breast cancer data. The data matrix, X, was generated with the following two correlation structures:

- Model VI: The data matrix, X, were partitioned by columns where each submatrix X_g corresponds to the covariates of the g^{th} group. The grouping structures were the same as the breast cancer data. For each group g, the corresponding submatrix, X_g, was independently generated from $\mathcal{N}_{k_g}(\mathbf{0}, \Sigma)$ with the element of Σ to be, $\sigma_{ij} = 0.5^{|i-j|}$.

- Model VII: As with model VI, the data matrix, X, were partitioned by columns where each submatrix X_g corresponds to the covariates of the g^{th} group. The grouping structures were the same as the breast cancer data. For each group g, the corresponding submatrix, X_g, was independently generated from $\mathcal{N}_{k_g}(\mathbf{0}, \Sigma)$. Differently, the element of Σ was set to be $\sigma_{ij} = 0.9^{|i-j|}$. Hence the model has a higher level of within-group correlations in generating X than model VI.

For each model, we collected 900 observations for each run. The generalized HSVS, the generalized fused-HSVS, and the generalized group lasso methods were compared in estimating models VI and VII. As in Section 3.5, the Bayesian estimates were posterior medians using $10,000$ iterations of the Gibbs sampler after $1,000$ burn-in iterations. Significant groups were selected based on an FDR of $\alpha = 0.10$. The generalized group lasso method estimated coefficients using the *grpreg* (Breheny and Huang, 2009) package with the tuning parameters selected by 5-fold cross validation.

The average model errors over 40 runs are presented at the bottom of Table 3.1, along with the false positive (FP) and false negative (FN) numbers of selected groups/individual variables for each method. In accordance with the real analysis results, the HSVS and the fused-HSVS methods are similar in performance. Furthermore, the two HSVS-based methods have slightly lower model errors than the generalized group lasso method, as well as better performance in within-group variable selection as evidenced by the FP and FN rates.

3.6 Application to Genomic Studies of Breast Cancer Subtype

3.6.1 MIP Study on Breast Cancer

Breast tumor samples with complete clinical and follow-up data and adequate tumor DNA from FFPE tissue blocks were identified through the Specialized Programs of Research Excellence (SPORE) in breast cancer at MD Anderson Cancer Center. The samples were obtained from a retrospective study of $2,409$ women diagnosed with pathologic stage I or II breast cancer who were surgically treated at MD Anderson between 1985 and 2000. Clinical information, including the patient's age, stage, tumor size, lymph node status, nuclear grade, estrogen receptor (ER) status, and progesterone receptor (PR) status, was also collected. The breast tumor samples were classified into four subtypes based on immunohistochemical analysis of ER, PR, HER2, and Ki67: luminal A ($\mathrm{ER}^{+}\mathrm{Ki67}^{low}$), luminal B ($\mathrm{ER}^{+}\mathrm{Ki67}^{high}$), HER2neu+, and triple-negative breast cancer (TNBC) ($\mathrm{ER}^{-}\mathrm{PR}^{-}\mathrm{HER2}^{-}$). DNA extracts from FFPE tumor samples and 10% matched normal samples (from the same patients)

Table 3.1

Simulation results under different model specifications. The mean errors over 200 replications for models I to V and the mean errors over 40 replications for models VI and VII are presented in the table; standard deviations are shown in parentheses. FP: false positive; FN: false negative. See Section 3.5.1 and Section 3.5.2 for details about the models.

		Group of Variables		Within-group Variable	
	Model Error	FP	FN	FP	FN
Model I:					
HSVS	0.22 (0.09)	0.0 (0.0)	0.0 (0.0)	0.14 (0.37)	0.79 (0.55)
Group Lasso	0.29 (0.15)	0.02 (0.15)	0.0 (0.0)	6.06 (0.63)	0.0 (0.0)
Lasso	0.64 (0.21)	11.41 (3.23)	0.0 (0.0)	18.42 (7.35)	0.28 (0.45)
Stepwise	1.92 (0.52)	17 (0.0)	0.0 (0.0)	74 (0.0)	0.0 (0.0)
Model II:					
HSVS	0.78 (0.26)	0.0 (0.0)	0.26 (0.44)	0.65 (0.88)	4.39 (1.28)
Group Lasso	0.84 (0.25)	0.1 (0.36)	0.25 (0.16)	18.35 (1.49)	0.05 (0.31)
Lasso	1.02 (0.30)	9.19 (1.68)	0.0 (0.0	22.97 (7.38)	0.93 (0.89)
Stepwise	1.85 (0.48)	11 (0.0)	0.0 (0.0)	62 (0.0)	0.0 (0.0)
Model III:					
HSVS	0.47 (0.18)	0.02 (0.14)	0.0 (0.0)	0.14 (0.37)	0.71 (0.79)
Group Lasso	0.53 (0.36)	0.07 (0.33)	0.0 (0.0)	4.27 (1.31)	0.03 (0.3)
Lasso	1.36 (0.60)	10.35 (5.32)	0.0 (0.01)	19.95 (15.99)	1.13 (0.99)
Stepwise	3.33 (0.57)	17 (0.0)	0.1 (0.0)	72 (0.0)	0.0 (0.0)
Model IV:					
HSVS	0.48 (0.23)	0.07 (0.25)	0.0 (0.0)	0.0 (0.0)	7.99 (0.10)
Group Lasso	0.46 (0.26)	2.29 (1.54)	0.0 (0.0)	13.16 (6.15)	0.0 (0.0)
Lasso	0.78 (0.89)	6.13 (5.52)	0.02 (0.12)	10.28 (12.68)	4.34 (1.23)
Stepwise	3.30 (0.58)	17 (0.0)	0 (0.0)	72 (0.0)	0.0 (0.0)
Model V:					
HSVS	0.37 (0.13)	0.01 (0.07)	0.0 (0.0)	0.52 (0.68)	0.37 (0.60)
Fused-HSVS	0.29 (0.12)	0.02 (0.14)	0.0 (0.0)	0.35 (0.58)	0.21 (0.45)
Group Lasso	0.40 (0.16)	0.37 (0.48)	0.0 (0.0)	13.7 (4.84)	0.0 (0.0)
Model VI:					
Generalized HSVS	0.05 (0.03)	0.0 (0.0)	0.15 (0.36)	0.20 (0.41)	1.15 (0.80)
Generalized Fused-HSVS	0.06 (0.03)	0.0 (0.0)	0.15 (0.36)	0.28 (0.51)	0.85 (0.83)
Generalized Group Lasso	0.13 (0.05)	2.1 (1.75)	0.0 (0.0)	12.43 (11.41)	0.0 (0.0)
Model VII:					
Generalized HSVS	0.08 (0.13)	0.0 (0.0)	1.8 (0.85)	0.13 (0.68)	6.95 (1.54)
Generalized Fused-HSVS	0.08 (0.12)	0.0 (0.0)	1.68 (0.92)	0.15 (0.58)	6.28 (1.74)
Generalized Group Lasso	0.19 (0.16)	2.5 (1.81)	1.43 (0.59)	16.93 (17.18)	2.85 (1.19)

were prepared for copy number measurements in the Affymetrix$^{\text{TM}}$MIP laboratory, which was blinded to all sample and subject information. The copy numbers measured from the MIP assay were then pre-processed following the AsCNAR method (Yamamoto et al., 2007), with the normal samples used as reference samples. See Thompson et al. (2011) for more details regarding the pre-processing steps.

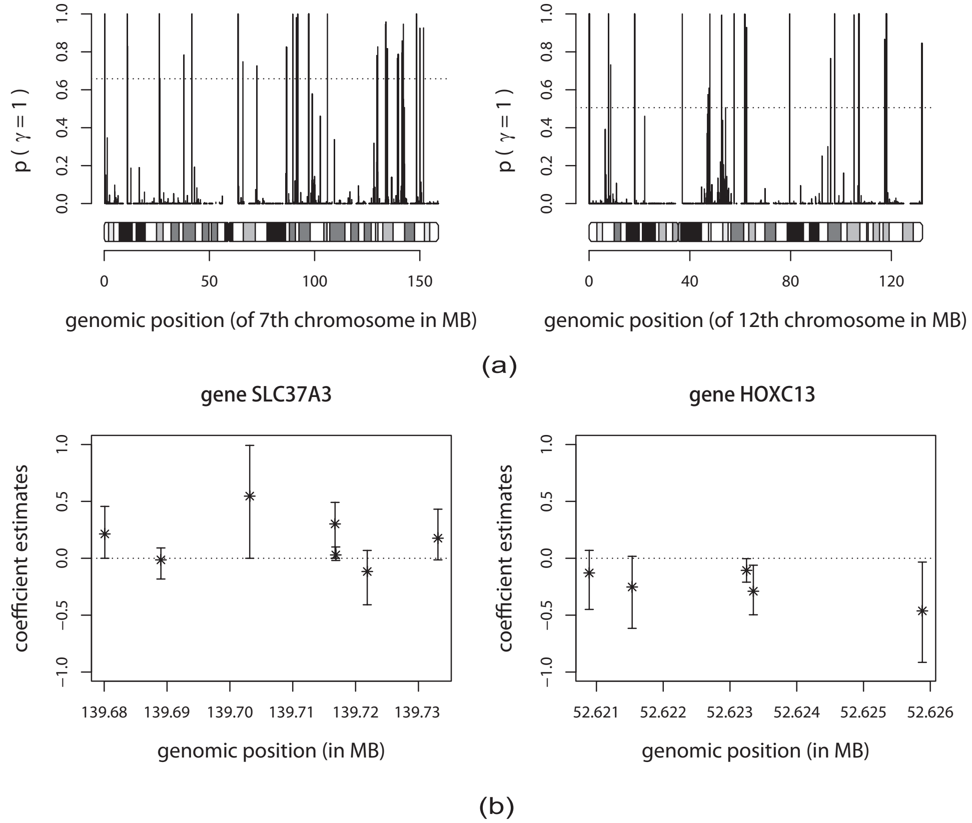
The MIP data contain the copy numbers of $330,000$ probes from tumor cells of 971 patients with breast cancer. More than $167,000$ probes were mapped to the coding regions of functional genes, with the number of probes located in the same gene ranging from 1 to more than 100. The dataset will be submitted to the NCBI database following the publication of Thompson et al. (2011).

3.6.2 Analysis Results

We applied our algorithm to the MIP assay dataset to identify genes as well as probes that are significantly associated with the clinically relevant subtypes of breast cancer. Among the 971 breast cancer samples, 389 are classified as luminal A, 156 as luminal B, 158 as HER2neu+, 184 as TNBC, and 84 as unclassified. For our proof of principle, we modeled the TNBC subtype using the MIP copy numbers. We elected to focus on the TNBC subtype as TNBC is among the more aggressive breast tumors for which there are no known treatment targets or prognostic factors; thus, results from our efforts would be of high interest. Hence, we have binary response variables, with $Y_i = 1$ if patient i has the TNBC subtype, and $Y_i = 0$ otherwise.

We modeled the binary response using the HSVS model for generalized linear models as discussed in Section 3.3. The candidate variables are the $167,574$ probes that are mapped to the coding regions of unique genes, with the probes in the same gene treated as a group. We ran our HSVS models for each chromosomal arm separately and used $10,000$ MCMC iterations with a burn-in of $1,000$ for inference. The convergence of the MCMC chains was assessed based on the Geweke diagnostic test (Geweke, 1992), which tests equality of the means of two nonoverlapping parts of a chain (the first 0.1 and the last 0.5 by default). The Geweke statistic asymptotically follows a standard normal distribution if the means are truly equal. The test on the MCMC samples on a random sample of model parameters indicated stationarity of the chains since the statistics were within $(-2, 2)$. Based on the posterior probabilities from the MCMC samples and an FDR of $\alpha = 0.10$, we selected 271 genes for the HSVS model. These genes were identified as significantly amplified (positive) or deleted (negative) in TNBC samples compared with other subtypes. Figure 3.3(a) shows the posterior probabilities of the genes on two chromosomes, with the dashed line indicating the FDR threshold. Genes are considered significant if their probabilities exceed the threshold. Figure 3.3(b) shows the posterior median coefficient estimates and the corresponding 95% credible intervals for the probes for two gene groups.

The Ingenuity System was used to perform a functional analysis based on the associated ontology terms of these selected genes and to generate Figure 3.4. Figure 3.4(a) shows that the immune response term is enriched in genes deleted from the TNBC subset. The cellular proliferation, development, and signaling-associated terms are enriched in both gene sets. Specifically, the genes amplified in the TNBC samples include the enzymes associated with oxidative phosphorylation (as seen in Figure 3.4(b)), gene RBBP8 (retinoblastoma binding protein 8) in the DNA damage response pathway, oncogenes such as PI3K (phosphoinositide-3-kinase) and SOS1 (son of sevenless homolog 1), and oncogenic transcription factor ETS1 (v-ets erythroblastosis virus E26 oncogene homolog 1-avian) (Chinnadurai, 2006; Dittmer, 2003). Other genes deleted in the TNBC samples are BTG2 (BTG family, member 2), which correlates with increased survival in breast cancer; PLK2 (polo-like kinase 2), which is associated with checkpoint-mediated cell cycle arrest; IRS1 (insulin receptor substrate 1), suppressor of metastasis in breast cancer; IL9 (interleukin-9) and IL13 (interleukin-13), which are associated with triggering immune response;



(a)

(b)

Fig. 3.3. Data analysis results: (a) The posterior probabilities of being included in the model in MCMC samples for the genes on chromosome 7 (left panel) and 12 (right panel). The dashed line indicates the FDR threshold where genes with probabilities above the line are considered significant; (b) The posterior median estimates with 95% credible intervals for the probes in two significant genes groups. The gene names are shown on the top of each plot.

and THBS1 (thrombospondin 1), an angiogenesis inhibiting factor (Eckerdt et al., 2005; Gibson et al., 2007; Lawler, 2002).

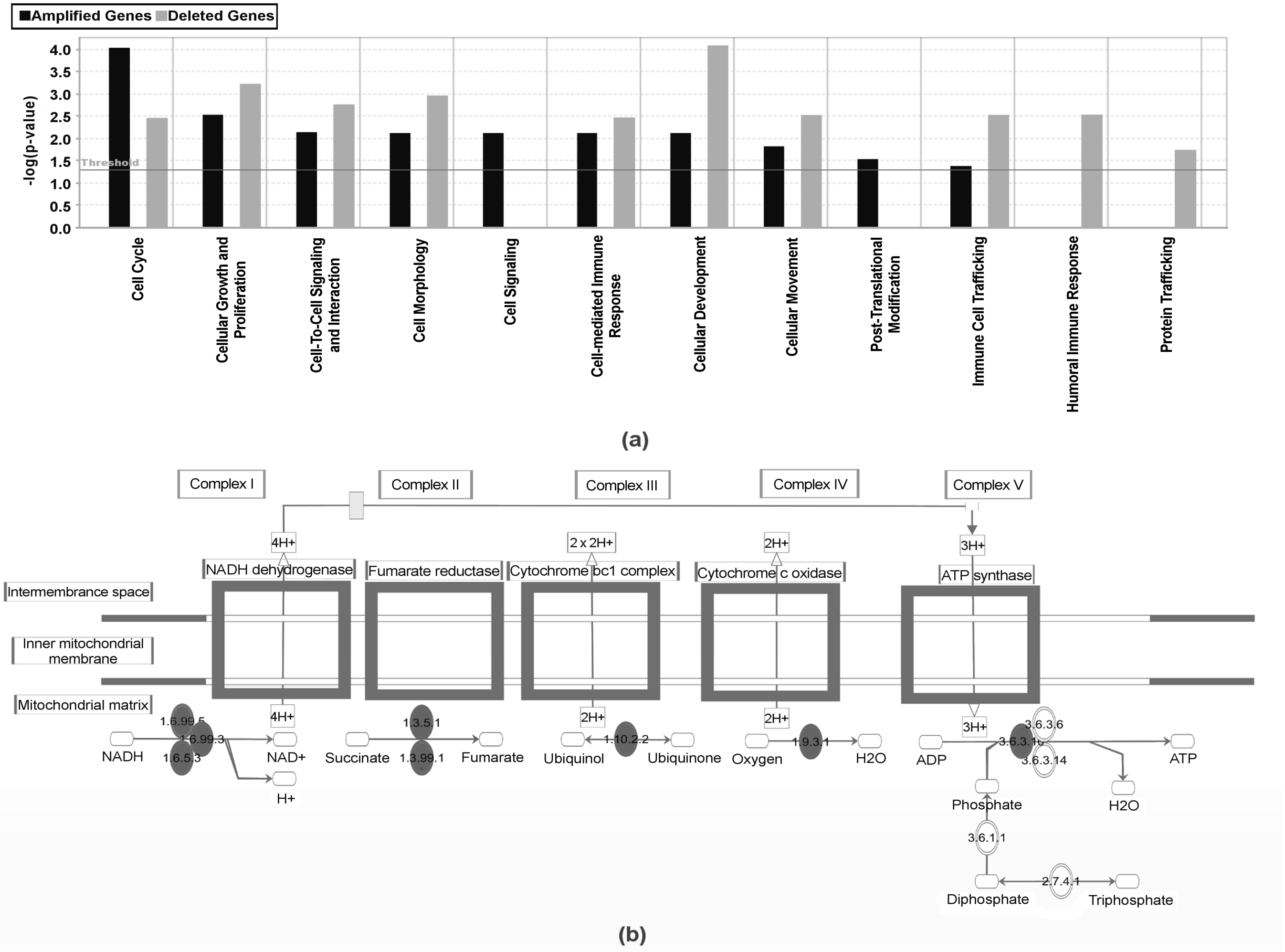


Fig. 3.4. Functional analysis of selected genes by the Ingenuity System. (a) Ontology terms associated with the genes that have a gain or loss of copy numbers in the TNBC data; (b) Ingenuity pathway depicting oxidative phosphorylation. The complexes denoted by the solid ellipses show the points at which each of the five genes (enriched in copy-number) plays a role in this pathway.

We also applied the fused-HSVS model to the MIP dataset. Based on the posterior probabilities from the MCMC samples and an FDR of 0.10, 294 genes were selected by the fused-HSVS model, with 232 genes the same as those identified by the HSVS model. A functional analysis shows that the basic cellular function, development, and signaling-associated terms are similar across the gene sets identified by the two methods (as seen in Figure 3.5(a)). Most of the genes of interest mentioned above, which were selected by the HSVS model, were also selected by the fused-HSVS model. Figure 3.5(b) shows the coefficient estimates for the two methods as well as the frequentist group lasso method when applied to a truncated MIP dataset of 1041 probes located in the coding region of 140 genes on chromosome 1. All the genes identified by the frequentist group lasso method also showed signals based on the HSVS methods. However, only two of them were considered significant with the FDR-based selection method. Comparing the HSVS and fused-HSVS models, the latter identified one more gene than the HSVS model, whose group members had very small coefficient estimates (0 to 0.20). The result further supports that the incorporation of the Bayesian fused lasso into our hierarchical model increases the performance of selecting large groups of variables with weak predictor members.

3.7 Discussion

In this section, we propose a novel Bayesian hierarchical method, HSVS, which performs both group-level and within-group-level variable selection simultaneously. We conducted simulation studies with various settings to evaluate the operating characteristics of our method. We found our HSVS method to be a strong variable selector at both group and within-group levels, which satisfies the need for parsimonious model selection. The method performs better overall than the group lasso and the regular lasso methods when both group-level and within-group-level selections are desired. However, the performance of the HSVS method decreases when the true model is less sparse or the variables have only weak effects on the response, due to the joint effect of the spike and slab and the lasso priors used in our method. In addition, the HSVS method performs slightly worse than the group lasso method when high correlations exist within groups, since we use the Bayesian lasso for within-group selection, which is not robust to such correlations.

We applied the method to a genetic association analysis of an MIP dataset collected from breast cancer patients, which gives new clues about genes that may be associated with TNBC. For example, it is generally accepted that cancer cells metabolize glucose by glycolysis rather than the more efficient oxidative phosphorylation. The identified copy-number gain of the genes associated with oxidative phosphorylation provides new information about the heterogeneous tumor group defined by TNBC. The copy-number gain of RBBP8 also indicates an effect of the oxidative stress caused by the enhanced oxidative phosphorylation in TNBC samples. Other genes that play important roles in regulating cell cycle, suppressing metastasis, and triggering immune response were identified as being deleted in TNBC patients, which may explain the aggressive property of the TNBC subtype.

Considering the natural ordering of probes within a gene in the MIP data, we extend the HSVS model by replacing the independent Laplace priors with the fused lasso priors for within-group-level selection. The implementation of the Bayesian fused lasso method encourages neighboring coefficients within a group to be close in value. This is expected in the genetic association study of the MIP data since the copy numbers of neighboring probes within a gene are thought to have similar effects on breast cancer development. The analysis suggests that the fused-HSVS prior tends to have a higher sensitivity than the HSVS prior for the genes whose probe variables have consistently weak regression coefficients.

There are several possible extensions of our HSVS-based models to more general settings in which variables have grouping structures. Examples of such applications include polynomial effects of the same factor, genes belonging to the same pathway, and proteins composing the same molecular complex. Another interesting extension would be in a survival context for time-to-event responses, which will address the more important biological question of finding prognostic markers for cancer progression. Finally, we can easily extend the hierarchical model by changing the "slab" part of the group prior for different purposes such as stronger within-group variable selection using various types of shrinkage priors. We leave these tasks for future consideration.

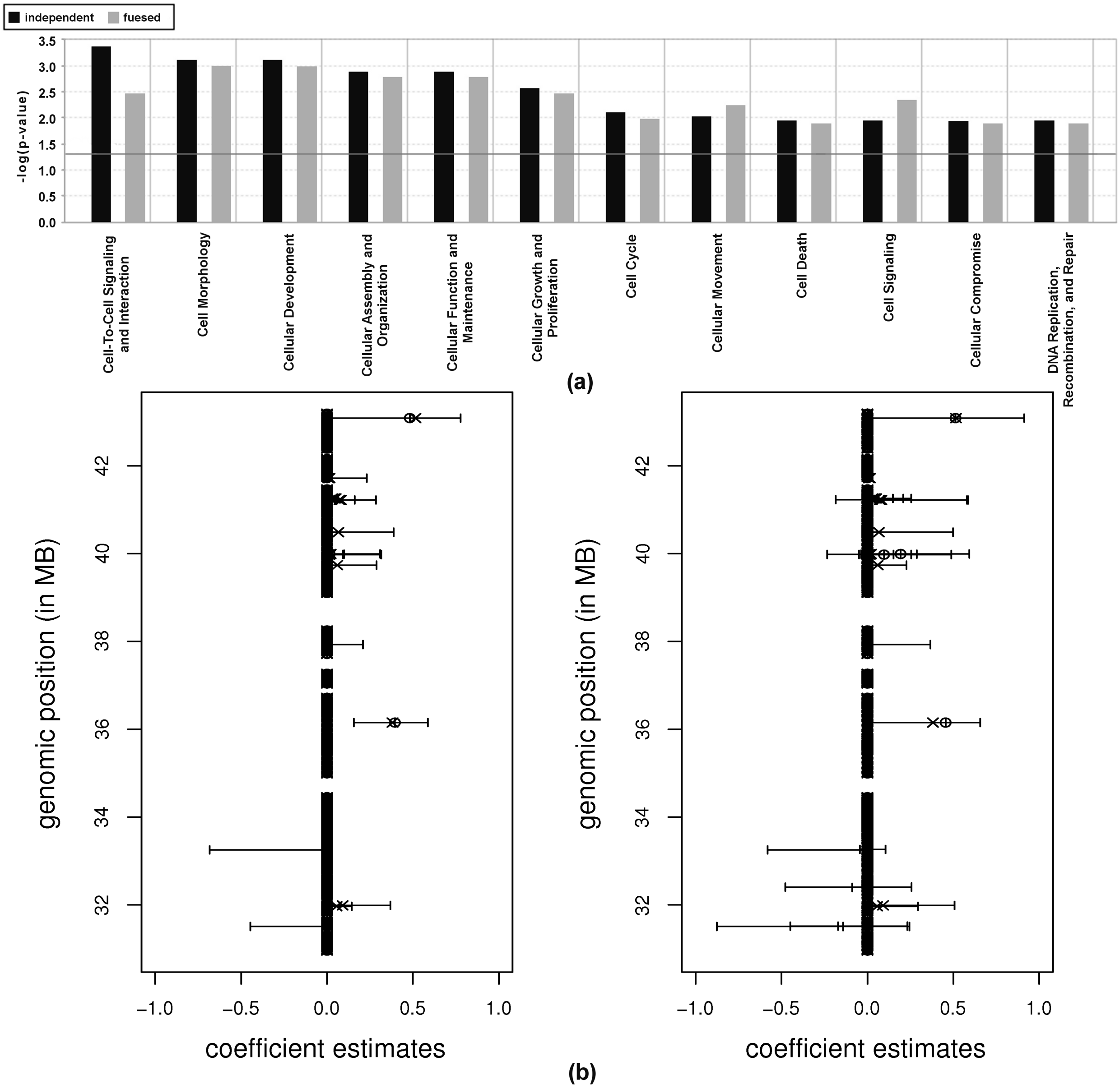


Fig. 3.5. Analysis results for the fused-HSVS model. (a) Comparison of the functional terms associated with the genes indicated by the HSVS (black color) and fused-HSVS (light grey color) methods. The plot is generated by the Ingenuity System; (b) Comparison of the coefficient estimates of a truncated MIP dataset for the HSVS model and fused-HSVS model. The left plot shows the posterior median estimates of the HSVS model with 95% credible intervals; the right plot shows the posterior median estimates of the fused-HSVS model with 95% credible intervals. The cross symbols in (b) are the coefficient estimates of the frequentist group lasso method.

4. BAYESIAN LOW RANK AND SPARSE COVARIANCE DECOMPOSITION

4.1 Introduction

Estimation of covariance matrices is a fundamental issue in multivariate analysis and many statistical applications including modeling genetic data, brain imaging, climate data, and many other fields. Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are q-dimensional random vectors which follow an independent and identical (iid) multivariate Gaussian distribution $\mathcal{N}_q(\boldsymbol{\mu}, \Sigma)$. It is well known that the sample covariance $\hat{\Sigma} = \sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'/(n-1)$ is not a stable estimator of the population covariance matrix, Σ, when the dimension of the covariance matrix is large relative to the sample size.

A number of approaches have been proposed for stable estimation of a high-dimensional covariance matrix efficiently. Pourahmadi (1999, 2000) and Huang et al. (2006) estimated the covariance matrix by parameterizing the Cholesky decomposition of its inverse. Leonard and Hsu (1992), Chiu et al. (1996), and Deng and Tsui (2010) modeled the matrix logarithm of the covariance matrix. Ledoit and Wolf (2004) constructed a shrinkage estimator which is a linear combination of the sample covariance matrix and a pre-chosen matrix. Others focused on parsimonious modeling by identifying zero off-diagonal elements in the covariance matrices or their inverse (Yuan and Lin, 2007; Friedman et al., 2008; Levina et al., 2008; Bien and Tibshirani, 2011; among others). In a Bayesian framework, Wong et al. (2003) used a selection prior for off-diagonal elements of the partial correlation matrix to identify zeros in an inverse covariance matrix. Talluri et al. (2011) and Wang (2012) used Byesian graphical lasso priors for sparse inverse covariance matrix estimation. Others employed a hyper-inverse Wishart (HIW) prior for covariance selection given a decomposable Gaussian graphical model (Lauritzen, 1996; Giudici and Green, 1999; and Armstrong et al., 2009), which was extended for nondecomposable graphical models by Giudici and Green (1999), Roverato (2002), Brooks et al. (2003), and Atay-Kayis and Massam (2005).

Luo (2011) introduced a different covariance structure for high-dimensional datasets, which can be decomposed into the summation of a low-rank and sparse matrix as

$$\Sigma = L + S, \tag{4.1}$$

where L is a low rank component and S is a sparse component. This decomposition of covariance matrices for dimension reduction has a wide range of applications including factor analytical model, random effects model and conditional covariance model, where the low rank component L indicates that the variation of the random vector can be explained by a small number of common factors or principal components, and the sparse part S displays the variance/covariance between the variables conditional on these latent common factors. Luo (2011) proposed a frequentist approach, LOREC, which regularizes $\hat{\Sigma}$ by the Frobenius norm and uses a composite penalty on the trace norm of L and the l_1 norm of S to achieve the low-rank and sparse component estimation respectively. In this dissertation, we propose a likelihood-based Bayesian approach to estimate a covariance matrix with the decomposition structure in equation (4.1), and extend our method for graphical factor analysis.

We represent the $q \times q$ low-rank matrix L utilizing a singular value decomposition (SVD) as follows

$$L = M D_\tau M^T, \tag{4.2}$$

where $M \in \mathbf{R}^{q \times r^*}$, the diagonal matrix $D_\tau = \text{diag}(\tau_1^2, ..., \tau_{r^*}^2) \in \mathbf{R}^{r^* \times r^*}$ consists of singular values of L, and r^* denotes the true rank of L. This decomposition ensures positive definiteness of Σ and determines the rank of L, which is given by the dimension of D_τ. When L is low-rank, $r^* << q$. This representation of the covariance matrix has the same structure as in a factor analytic model, where M could be viewed as the latent factor loadings matrix, the singular values τ_k^2's as the variances of the latent factors, and the sparse component S as the covariance matrix of the residuals. In our Bayesian method, we estimate L and S through the factor analytic model, which has an unknown number of factors and a sparse covariance matrix of the residuals. To estimate the rank of L, i.e. the number of true factors, we introduce a binary indicator for each factor separating factor selection and singular value estimation, and use a hierarchical prior strongly penalizing the rank of L, as is desired for the low rank property of L. To achieve sparsity in S estimation, we propose a Bayesian graphical lasso selection method, which combines a lasso-type prior for covariance element shrinkage and a variable selection prior enforcing weak elements to be exactly zero. We assign priors such that most of the parameters have closed-form conditional posterior distributions, which facilitate the Monte Carlo Markov chain (MCMC) computation and also allow an automatic choice of the regularization parameters.

In statistical applications such as gene expression and financial data analysis, we often are interested in the graphical model of the variables, which is flagged by zero pattern in the off-diagonal of S inverse. In these cases, the sparsity in $C = S^{-1}$ instead of S is desired for graphical model inference. For this purpose, we extend our method to a graphical factor analytic model so that it achieves sparsity in estimating C. We employ a conjugate HIW prior on S in the graphical factor analytic model when S is restricted to decomposable graphs, and use a Bayesian graphical lasso selection prior on C when the graph is unrestricted. We show through simulations that the extended model can recover both the number of latent factors and the graphical model of the residuals successfully when the sample size is sufficient relative to the dimension.

The rest of the section is organized as follows. In Section 4.2, we describe our proposed Bayesian model for low rank and sparse covariance matrix estimation. In Section 4.3 we report results from simulation studies to assess the operating characteristics of our method. A real data analysis of gene expression data is included in

Section 4.4. In Section 4.5 we extend our method to develop a graphical factor analytic model to handle data where selection of the latent factors as well as inference of the graphical model among variables are both of interest, and show the application of the graphical factor analytic model to a gene expression dataset. We provide a discussion and conclusion in Section 4.6. The full conditional distributions and the posterior sampling algorithm are described in Appendix C.

4.2 Proposed Bayesian Low Rank and Sparse Covariance Model

Consider a $q \times n$ data matrix y, with each column vector $\mathbf{y}_i$ for $i = 1, \ldots, n$ following an iid Gaussian distribution

$$\mathbf{y}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma). \tag{4.3}$$

We assume that the covariance matrix Σ could be decomposed as a sum of a low rank component L and a sparse component S, with L to be represented as a singular value decomposition $MD_\tau M^T$ as in equation 4.2. Hence we have

$$\Sigma = MD_\tau M^T + S. \tag{4.4}$$

Note that the representation of the covariance in equation (4.4) can be viewed from the standpoint of a latent factor analytic model:

$$\mathbf{Y} = M\mathbf{F} + \boldsymbol{\epsilon}, \tag{4.5}$$

where M is the $q \times r^*$ latent factor loadings matrix, the diagonal elements in D_τ are the variances of F, the r^*-dimensional vector of random factors, and S is the covariance matrix of the residuals $\boldsymbol{\epsilon}$. Grzebyk, Wild and Chouanière (2004) gave a sufficient condition for the identification of a multi-factor model with correlated residuals as in (4.4). However, our main focus here is on the inference of covariance matrices Σ. Thus, we do not restrict our method to a unique solution for Σ decomposition.

Since we do not know the rank of L, i.e. the number of latent factors r^* in the model, we need to both estimate the number of factors and the variances of the factors. To separate the two tasks, we introduce an extra binary indicator matrix Z. Our proposed Bayesian model is

$$\Sigma = M(ZD_\tau)M^T + S, \tag{4.6}$$

where D_τ is a diagonal matrix with positive diagonal elements τ_k^2 for $k = 1, \ldots, r$ for some $r > r^*$, and Z is a diagonal matrix with binary entries $z_k \in \{0, 1\}$ for $k = 1, \ldots, r$ along the diagonal. While τ_k gives the variance of the k^{th} latent factor, the indicator z_k determines if the latent factor is included in the model. In this way, we separate the recovery of rank and the estimation of the singular values of L. The rank of L, r^*, is only determined by the number of 1's in the diagonal of Z. Now the estimation of r^* is equivalent to selecting the true number of latent factors in a factor analytic model. In our method, we choose a relatively large integer $r \leq q$ which is supposed to be much larger than r^*, and expect that the diagonal entries of Z is sparse. If the estimates of Z diagonals are not sparse, we increase the value of r.

We can rewrite the likelihood model in (4.3) and (4.6) as the regression-type representation of a latent factor analytic model:

$$\begin{aligned} \mathbf{y}_i &= MZ\mathbf{f}_i + \boldsymbol{\epsilon}_i, \text{ for } i = 1, \ldots, n \\ f_{ki} &\sim \mathcal{N}(0, \tau_k^2), \text{ for } k = 1, \ldots, r \\ \epsilon_i &\sim \mathcal{N}_q(0, S), \end{aligned} \tag{4.7}$$

where $\mathbf{f}_i = (f_{1i}, \ldots, f_{ri})^T$ are the values of the r-dimensional latent factor vector in the i^{th} replication for $i = 1, \ldots, n$. If $z_k = 1$, the k^{th} factor, $k = 1, \ldots, r$, is a true factor of the variables $\mathbf{Y}$ with variance τ_k^2; otherwise, the k^{th} factor is not included in the factor model. By rewriting the model in the form of a linear regression problem, we can assign conjugate priors to M, Z and D, which leads to closed-form full conditional distributions of the parameters and facilitates the posterior sampling using a Gibbs algorithm.

4.2.1 Prior Specification for the Low Rank Component

To complete model specification, we need to assign priors to the set of parameters $\{M, Z, D, S\}$ in the hierarchical likelihood model (4.7), where M, Z, D give the low rank component L, and S is the sparse component. Let m_{jk} be the element of M in the j^{th} row and k^{th} column, and $\mathbf{M}_k = (m_{1k}, \ldots, m_{qk})'$ be the k^{th} column vector of M, which could be viewed as the loading vector of the factor k on the variables. We assume that M_k has a Gaussian prior:

$$\mathbf{M}_k \sim \mathcal{N}_q(\mathbf{0}, \frac{1}{q} I_q), \quad k = 1, \ldots, r,$$

where I_q is a $q \times q$ identity matrix. Note that for a large dimension q, the columns of M are approximately orthogonal. Furthermore, by assigning the prior variance of m_{ik} to be $1/q$, we reduce the variability of m_{ik}, shift the variability to the single element τ_k^2 in D_τ, and obtain a relatively stable estimate of L.

The binary diagonal matrix Z is modeled as

$$z_k \sim \text{Bernoulli}(p_k), \quad k = 1, \ldots, r,$$

where p_k is the prior probability of $z_k = 1$. The values of p_k determine the strength of the penalization that is assigned to the rank of L as $r^* = rank(L)$ is equivalent to the number of $z_k = 1$. Since L is assumed to be of low-rank, most of the prior probabilities are expected to be small or zero. We model these probabilities with the following hyper-prior distribution:

$$p_k \sim (1-\pi)\mathcal{I}\{p_k = 0\} + \pi \text{Beta}(a_p, b_p),$$

The hyper-prior of p_k is a Beta distribution mixed with a point mass at 0 with probability π, where π is drawn from a Beta prior distribution Beta(a_π, b_π). The sparseness of the diagonals of Z is explicitly imposed through the hyperparameters (a_p, b_p) and (a_π, b_π). When $a_\pi/(a_\pi + b_\pi) \ll 1$, p_k has a high prior probability to be zero; when $a_p/(a_p + b_p) \ll 1$, p_k is still likely to be close to zero if it is not zero. To impose high penalization on the rank of L, we choose $(a_\pi, b_\pi) = (1/q, (q-1)/q)$ and $(a_\pi, b_\pi) = (1, r)$.

Each diagonal entry τ_k^2 in D_τ corresponds to the variance of the k^{th} factor. Hence we can assign a conjugate Inverse-Gamma prior

$$\tau_k^2 \sim \text{IG}(a_\tau, b_\tau), \quad k = 1, \ldots, r,$$

which leads to a closed form of the posterior conditional distribution. Another option is to specify an improper prior distribution for τ_k^2, $p(\tau_k^2) \propto 1/\tau_k^2$.

4.2.2 Prior Specification for the Sparse Component

In order to achieve adaptive shrinkage of the sparse component S, we use a Bayesian graphical lasso prior for estimating the sparse matrix S. In the graphical lasso method, an l_1 penalty term is assigned to S, which, in a Bayesian framework, is equivalent to independent exponential priors on the diagonal elements S_{jj}, $j = 1, \ldots, q$, and double exponential priors on the off-diagonal elements $S_{jj'}$, $j < j'$. However, the Bayesian graphical lasso method does not set the off-diagonal elements to exact zeros, which is desired for the sparse S estimation. To this end, we modify the method by placing a point mass at 0 in the double exponential priors. The priors are detailed as follows:

$$
\begin{aligned}
S_{jj}|\lambda &\sim \text{Exp}(\frac{\lambda}{2}), \quad j = 1, \ldots, q, \\
S_{jj'}|\lambda, \rho_{jj'} &\sim (1 - \rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda), \quad j < j',
\end{aligned}
$$

where $\mathcal{I}(\,\cdot\,)$ is an indicator function. We choose a conjugate gamma hyper-prior distribution for the shrinkage parameter λ and a conjugate Beta hyper-prior distribution for the selection parameters $\rho_{jj'}$:

$$
\begin{aligned}
\lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \\
\rho_{jj'} &\sim \text{Beta}(a_\rho, b_\rho).
\end{aligned}
$$

In this construction, the hyper-parameter λ shrinks the covariance elements toward zero, while $\rho_{jj'}$ controls the probability that the (j, j') element will be enforced to be a zero. In our experiments, we specify $(a_\lambda, b_\lambda) = (1, 1)$ for a diffuse prior for λ, and $(a_\rho, b_\rho) = (0.5, 0.5)$ for a noninformative prior of $\rho_{jj'}$.

The complete hierarchical model can be succinctly summarized as

$$\left.\begin{aligned} \mathbf{y}_i &\sim \mathcal{N}_q(MZ\mathbf{f}_i, S), \quad i = 1, \ldots, n \\ f_{ki} &\sim \mathcal{N}(0, \tau_k^2), \quad k = 1, \ldots, r \end{aligned}\right\} \tag{4.8}$$

$$\left.\begin{aligned} \mathbf{M}_k &\sim \frac{1}{q}\mathcal{N}_q(\mathbf{0}, \frac{1}{q}I_q), \\ z_k &\sim \text{Bernoulli}(p_k), \\ p_k &\sim (1-\pi)\mathcal{I}\{p_k = 0\} + \pi\text{Beta}(a_p, b_p), \\ \pi &\sim \text{Beta}(a_\pi, b_\pi), \\ \tau_k^2 &\sim \text{IG}(a_\tau, b_\tau), \end{aligned}\right\} \tag{4.9}$$

$$\left.\begin{aligned} S_{jj} &\sim \text{Exp}(\frac{\lambda}{2}), \quad j = 1, \ldots, q \\ S_{jj'} &\sim (1-\rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda), \quad j < j' \\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \\ \rho_{jj'} &\sim \text{Beta}(a_\rho, b_\rho), \end{aligned}\right\} \tag{4.10}$$

where i denotes the sample, k denotes the latent factor, $i = 1, \ldots, n$, and $k = 1, \ldots, r$.

4.3 Simulation Studies

We conducted a detailed simulation study to evaluate the operating characteristics of our method. We considered three covariance models to generate the data:

- Model 1: $\Sigma = UDU^T + I$, where $U \in \mathbf{R}^{q \times 3}$ with orthonormal columns, and $D = \text{diag}(8, 8, 8) \cdot (q/n)$. This covariance model comes from a factor model with independent residuals.

- Model 2: $\Sigma = 0.3\mathbf{1}\mathbf{1}^T + S$, where S is block diagonal with each square block matrix B of dimension 5, and $B = 0.7\mathbf{1}\mathbf{1}^T + 0.3I$. This covariance matrix simulates a random effect model, with the covariance of the residuals to be block diagonal.

- Model 3: $\Sigma = UDU^T + S$, where U and D are the same as in model 1, and S is a block diagonal matrix as in model 2. This covariance model comes from a factor model with the covariance of the residuals to be block diagonal.

For each model, 50 observations were generated from the multivariate Gaussian distribution $\mathcal{N}_q(\mathbf{0}, \Sigma)$ with varying dimensions $q = 50$, 100, and 200. We compared our proposed Bayesian model for low-rank and sparse covariance decomposition with the frequentist LOREC method (Luo, 2011) in estimating the covariance matrices as well as recovering the rank of L and sparsity of S. The estimates of the parameters using the Bayesian method were based on the posterior samples of 5000 iterations after 1000 burn-in iterations. The tuning parameters for the LOREC estimators were picked by 5-fold cross validation using the Bregman divergence loss as in Luo (2011).

Table 4.1 compares the performance of covariance estimation with our Bayesian method, Luo's LOREC method, and the sample covariance over 20 replicates measured by the l_1 norm and the Frobenius norm. The two matrix norms are defined as follows: Let $X = (X_{ij})$ be any matrix; $|X|_1 = \sum_i \sum_j |X_{ij}|$ gives the l_1 norm, and $|X|_F = \sqrt{\sum_i \sum_j X_{ij}^2}$ gives the Frobenius norm. While the LOREC estimator performs better for the random effect model, our Bayesian estimator has lower losses for the factor model with independent residuals. The two methods have similar losses in term of the norms for the factor model with correlated residuals. Both of them are better than the sample covariance in all models.

Table 4.2 summarizes the inference results in the recovery of rank of L and sparsity of S. The table shows that the Bayesian estimator can recover the true rank of the low rank components with high frequencies for all the three models, with the successful recovery rates close to the LOREC estimator. Furthermore, our method has much lower false positive rates in support recovery of S when S is non-diagonal, at the price of a little higher false negative rates. The above results indicate that our method can recover both the rank and the sparsity of the two components with high frequencies.

Table 4.1

Simulation results for Bayesian low rank and sparse matrix decomposition for model 1, 2 and 3. The mean results over 20 replications are presented in the table with the standard deviations in parentheses. See Section 4.5.5 for details about the models.

Losses of Covariance Estimators

		model 1		model 2		model 3	
		L_1 norm	Frobenius	L_1 norm	Frobenius	L_1 norm	Frobenius
q=50	Bayesian	11.90 (1.82)	8.46 (0.82)	12.89 (1.63)	11.62 (1.46)	13.40 (2.02)	9.78 (0.70)
	LOREC	13.64 (1.76)	9.15 (0.68)	11.84 (1.77)	9.12 (0.98)	13.55 (1.73)	9.75 (0.70)
	Sample	15.18 (2.12)	11.63 (0.78)	13.98 (2.95)	11.06 (1.27)	13.69 (1.99)	10.75 (0.86)
q=100	Bayesian	15.28 (2.45)	10.00 (0.78)	25.48 (2.74)	16.13 (1.17)	20.89 (3.81)	15.93 (1.00)
	LOREC	16.54 (2.19)	10.37 (0.96)	23.67 (2.98)	14.18 (1.68)	21.97 (3.18)	15.40 (0.96)
	Sample	20.74 (1.90)	17.36 (0.52)	26.18 (4.64)	20.08 (1.80)	25.39 (5.00)	19.51 (1.43)
q=200	Bayesian	29.90 (3.98)	18.16 (1.15)	49.83 (6.28)	32.71 (10.94)	35.58 (4.10)	25.23 (1.21)
	LOREC	32.79 (4.06)	19.31 (1.77)	48.96 (7.51)	28.02 (6.09)	39.66 (3.37)	23.68 (1.04)
	Sample	42.58 (2.56)	35.42 (0.97)	54.49 (4.96)	37.6 (1.66)	45.16 (5.25)	35.76 (1.75)

Table 4.2

Simulation results for Bayesian low rank and sparse matrix decomposition for model 1, 2 and 3. The mean results over 20 replications are presented in the table with the standard deviations in parentheses. FP: false positive discoveries; FN: false negative discoveries. See Section 4.5.5 for details about the models.

		Rank Recovery					
		model 1		model 2		model 3	
		%(3 factors)	mean(se)	%(1 factors)	mean(se)	%(3 factors)	mean(se)
q=50	Bayesian	95	2.95 (0.22)	90	0.90 (0.31)	80	2.80 (0.41)
	LOREC	100	3.00 (0.00)	100	1.00 (0.00)	40	2.05 (1.00)
q=100	Bayesian	90	2.90 (0.31)	80	1.20 (0.41)	85	2.85 (0.37)
	LOREC	95	2.95 (0.22)	100	1.00 (0.00)	50	2.35 (0.93)
q=200	Bayesian	100	3.00 (0.00)	90	0.92 (0.28)	90	2.9 (0.31)
	LOREC	90	2.90 (0.31)	100	1.00 (0.00)	80	2.7 (0.66)
		Sparsity Recovery					
		model 1		model 2		model 3	
		FN	FP	FN	FP	FN	FP
q=50	Bayesian	0 (0)	7.75 (4.06)	2.60 (1.90)	40.95 (11.38)	13.45 (7.96)	21.7 (8.16)
	LOREC	0 (0)	6.00 (15.75)	0.00 (0.00)	188.25 (50.49)	6.20 (3.07)	518.25 (93.35)
q=100	Bayesian	0 (0)	13.70 (6.33)	15.90 (4.28)	68.90 (15.48)	33.40 (6.06)	37.95 (11.33)
	LOREC	0 (0)	3.00 (8.05)	0.05 (0.22)	508.25 (97.80)	5.85 (3.10)	1011.6 (304.4)
q=200	Bayesian	0 (0)	2.0 (1.73)	103.0 (11.91)	34.6 (41.13)	157.0 (15.19)	10.7 (3.01)
	LOREC	0 (0)	0.8 (1.47)	0.6 (1.26)	1175.6 (265.0)	6.4 (4.03)	1848.3 (495.0)

4.4 Covariance Estimation on Gene Expression Dataset

In this section, we applied the Bayesian low rank and sparse decomposition model to estimating the covariance of a gene expression dataset from Stranger et al. (2007). The dataset was composed of 60 unrelated individuals of Northern and Western European ancestry from Utah (CEU). There were four replicates for each individual. The raw data were background corrected, quantile normalized across replicates of each individual, and then median normalized across all individuals. We considered 100 genes in our dataset which are most variable among all the genes available in the gene expression profile. Thus we had $n = 60$ and $q = 100$ in our dataset.

We estimated the covariance matrix using our Bayesian method with the LOREC estimator and the sample covariance as comparison. Figure 4.1 displays a heatmap showing the absolute intensities of three covariance matrix estimates. Compared to the sample covariance matrix, the LOREC estimator regulates the sample covariance estimate by shrinking all the off-diagonal elements uniformly, whereas the Bayesian decomposition model shrinks more of the elements with strong signals on the top left corner while keeps the abundant elements with moderate signals at the same time.

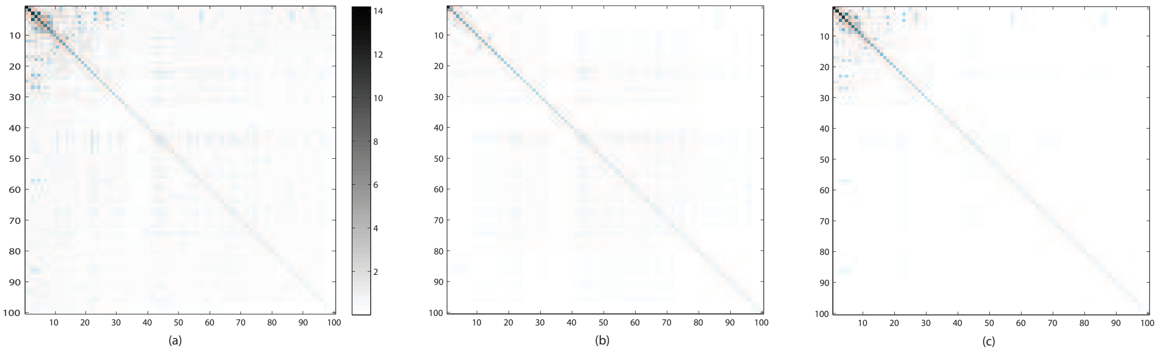


Fig. 4.1. Heatmaps of the absolute of covariance estimates by (a) sample covariance (b) Bayesian decomposition method (c) LOREC estimator.

The Bayesian method and the LOREC estimator also decompose the covariance matrix into low rank and sparse components in the gene expression data. The LOREC estimator identifies a rank 1 component, and our Bayesian method identifies a low rank component of rank 2. The singular vector of a rank 1 component is equivalent (up to a multiplying constant) to the loading in a single factor model, and therefore we obtain the single loading vector from the rank 1 component of the LOREC estimator. We also obtain the loadings matrix corresponding to the two random factors identified by our Bayesian decomposition model. Figure 4.2 shows the scatter plots of the two loadings by the Bayesian model versus the single loading by the frequentist LOREC method, one of which corresponds to a correlation close to 1. It suggests that the Bayesian model identifies one latent factor with a similar loading as the LOREC estimator.

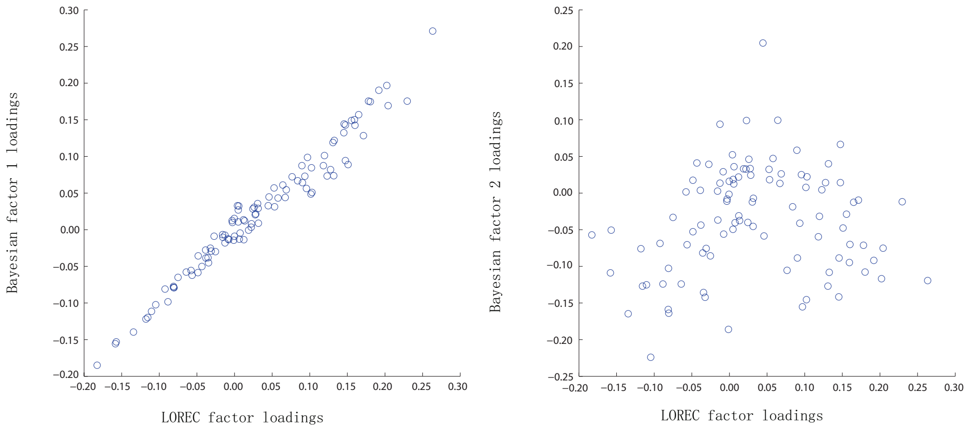


Fig. 4.2. Scatter plots of the single factor loading identified by LOREC versus the two factors loadings identified by the Bayesian decomposition model. The correlation between the two loading vectors on the left subplot is 0.98, and the correlation between the two loading vectors on the right subplot in 0.12.

Figure 4.3 displays the sparse support of the residual covariance component obtained by the Bayesian decomposition method as well as the LOREC estimator. The

LOREC estimator identifies nonzero correlations predominantly on the left corner of the sparse component, whereas the Bayesian decomposition model detects nonzero correlations overspreading the sparse matrix. This difference in the support of the sparse component agrees with the patterns in the covariance estimators plotted in Figure 4.1.

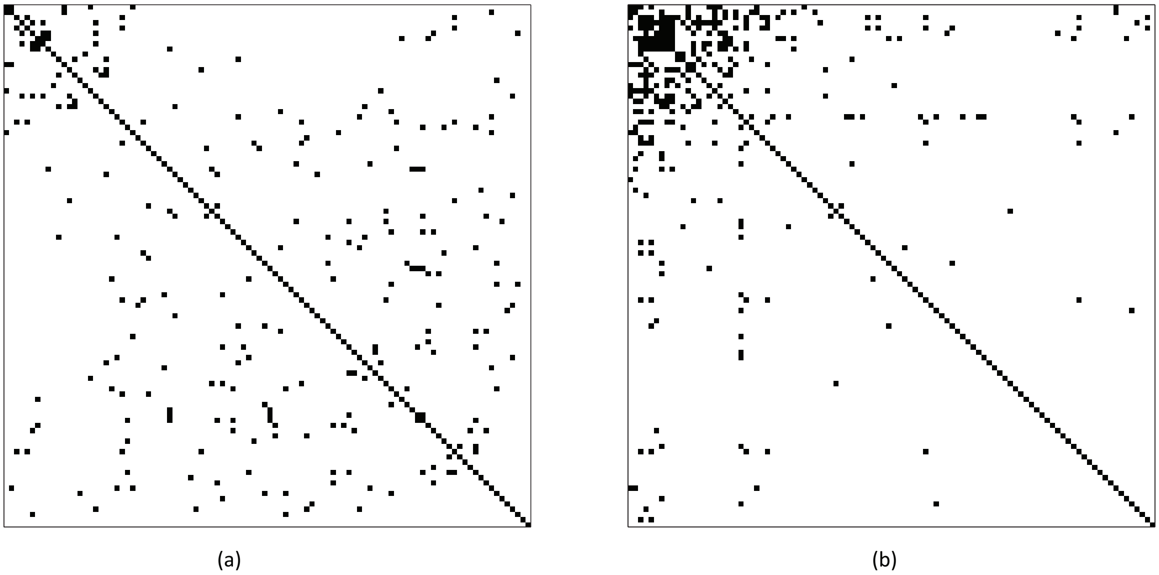


Fig. 4.3. Matrix plot indicating the sparse support of the residual covariance component by (a) Bayesian decomposition method (b) LOREC estimator.

4.5 Bayesian Graphical Latent Model

4.5.1 Introduction

As shown in the hierarchical likelihood model (4.7), the covariance model could be represented as a latent factor model. We assume in our proposed hierarchical model in (4.8)-(4.10) that the covariance matrix S of the residuals ϵ_i in the latent factor model is sparse, and use a Bayesian graphical lasso selection prior to achieve sparsity in S. However, in many applications of factor analysis including gene expression data and financial data analysis, the graphical Markov models are of more interest as they represent the conditional dependence among a set of observed variables. In this section, we extend the latent factor model in (4.8) by assuming that the inverse covariance matrix of the residuals, $C = S^{-1}$, is sparse, whose nonzero pattern in the off-diagonal elements gives the conditional dependence arising out of a graphical model. The extended latent factor model with sparse S^{-1} is a sparse graphical factor model with the number of factors unknown.

Factor analytic models have been extensively studied for summarizing the variance and covariance patterns in multivariate data. With advances in computational tools such as MCMC algorithms, Bayesian methods for factor analysis have been rapidly developed as seen in Geweke and Zhou (1996), Aguilar and West (2000), and Rowe (2003) among others. Lopes and West (2004) explored the inference on the number of latent factors in a factor model with a reversible jump MCMC method. Other recent Bayesian factor analysis incorporated different modeling structures through the columns of the factor loadings matrix (Lopes and Carvalho, 2007; Carvalho et al., 2008). However, all of the methods assume that the covariance matrix of the residuals is diagonal. That is, all the associations among the observed variables are exclusively contributed to the latent factors. Giudici (2001) induced the concept of a graphical factor analytic model, which generalizes factor analytic models by allowing the concentration matrix of the residuals to have non-zero off-diagonal elements. He used an HIW prior (Dawid and Lauritzen, 1993) for inference on the concentration matrices restricted to decomposable graphical models, and assigned a uniform prior on all decomposable graphs.

We make the following contributions in our extended model. First, we recover the number of factors as well as the graphical models in a graphical factor analytic model. Second, we propose a novel prior on the decomposable graphs for the HIW method, which induces adaptive sparsity in the inferred graphical models. Finally, we extend the method from modeling only decomposable graphs with HIW priors to unrestricted graphs by using a Bayesian graphical lasso selection method. Hence, this framework allows for additional flexibility both in the aspect of the analysis of common factors and the modeling of graphical models.

4.5.2 The Graphical Factor Model with Unknown Number of Factors

Before introducing the graphical factor analytic model, we first describe the notations in a graphical model. Let $\mathbf{Y}$ be a $q-$dimensional vector of random variables. A conditional independence graph is a pair of $G = (V, E)$ with the vertex set $V = \{1, ..., q\}$ and the edge set $E \subseteq V \times V$. Nodes j and j' are adjacent or connected in G if $(j, j') \in E$, whereas j and j' are conditionally independent if $(j, j') \notin E$. A graph G with $E = V \times V$ is called a complete graph. Complete subgraphs $P \subset V$ are called cliques; the joint subset of two cliques is called a separator denoted by Q. If a graph G could be partitioned into a sequence of subgraphs $(P_1, Q_2, P_2, ..., P_K)$ such that $V = \bigcup_k P_k$ and $Q_k = P_{k-1} \bigcap P_k$ are complete for all $k = 1, ..., K$, G is called a decomposable graph (Lauritzen, 1996). For a covariance matrix S of the variables $\mathbf{Y}$, let $C = S^{-1}$ be the inverse covariance matrix, or the precision matrix. Nodes j and j' are conditionally independent, given other nodes, if and only if $C_{jj'} = 0$. Thus, the graph G is given by the configuration of nonzero off-diagonal elements of C: $E = \{(j, j') : C_{jj'} \neq 0\}$.

The standard factor model relates each sample of size q, $\mathbf{y}_i$, to an underlying r^*-dimensional vector of common random factor $\mathbf{f}_i$ via the linear regression model

$$\mathbf{y}_i = M\mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

where M is the $q \times r^*$ factor loadings matrix, $\mathbf{f}_i = (f_{1i}, \ldots, f_{r^*i})^T$ are the values of the factors in i^{th} replication, and $\boldsymbol{\epsilon}_i$ are the residuals independent of the latent factors. In classical factor models with pre-specified number of factors, the factors $\mathbf{f}_i$ are assumed to follow an independent normal distribution $\mathcal{N}_{r^*}(0, I_{r^*})$, and the residuals $\boldsymbol{\epsilon}_i$ are from an independent normal $\mathcal{N}_q(0, S)$ with $S = \text{diag}(s_{11}, \ldots, s_{qq})$.

We relax the assumptions in our model in the following two aspects: (1) the number of underlying factors, r^*, is unknown, and is thought to be much smaller than some pre-specified integer r. Since the number of common factors is usually small, we pick a moderate to large value of r and expect only a small fraction of the factors are selected. (2) $\boldsymbol{\epsilon}_i \sim \mathcal{N}_q(0, S)$ with $C = S^{-1}$ to be a sparse concentration matrix. That is, we allow nonzero off-diagonal elements in the concentration matrix C so that the unobserved variables could be dependent on each other conditional on the latent factors.

A sufficient condition for identification of a graphical factor model with a single factor and multiple factors is given in Stanghellini (1997) and Guidici (2001) respectively. However, from a Bayesian viewpoint, identification is of less theoretical concern but more important for posterior computation as discussed in Guidici (2001). When the graphical factor model is unidentifiable (with more than one solutions), the likelihood would be flat, and the posterior distribution of parameters would be multimodal except for extremely informative priors.

Our objective is to select the true number of factors out of r candidate factors, and to recover the sparse graphical model of the variables represented by the nonzero pattern in C as well. For factor selection, we introduce a binary indicator z_k for each candidate factor k and assume the factor model in the following linear regression form

$$\mathbf{y}_i = MZ\mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

where Z is an $r \times r$ diagonal matrix with $z_k \in \{0, 1\}$ to be the k^{th} diagonal element, and f_{ki} is distributed as $\mathcal{N}(0, \tau_k^2)$. In words, a random factor k is a common factor of the observed variables with variance τ_k^2 if $z_k = 1$; otherwise, it will be excluded from the factor model. Hence, the hierarchical specification of the graphical factor model is

$$\begin{aligned}
\mathbf{y}_i &= MZ\mathbf{f}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \\
f_{ki} &\sim \mathcal{N}(0, \tau_k^2), \\
\boldsymbol{\epsilon}_i &\sim \mathcal{N}(0, S),
\end{aligned}$$

where $C = S^{-1}$ is sparse. Notice that this graphical factor model parallels the Bayesian covariance decomposition model in equations (4.8), except that now the inverse covariance matrix of $\boldsymbol{\epsilon}_i$ is modeled to be sparse.

4.5.3 Bayesian Hierarchical Model for Decomposable Graphs

For factor selection and loadings estimation in the graphical factor model, we use the same modeling method as in the Bayesian covariance decomposition model. That is, we assign the same priors for M ,Z, and D_τ as in equations (4.9). Hence the graphical factor model has the same likelihood as the Bayesian covariance decomposition model in equations (4.8), and the same hierarchy for factor estimation as in equations (4.9). The only difference lies in the assumption and the modeling method for S, the covariance matrix of the residuals. In this section, we focus on the modeling method when the graphical models of the residuals are restricted to be decomposable. The Bayesian model for nondecomposable or unrestricted graphs will be discussed in the next section.

When the graphical models of the residuals are restricted to be decomposable, we allow the covariance matrix of the residuals S to follow a mixture of HIW priors over decomposable graphs as

$$\begin{aligned} S \sim & \text{HIW}(G, \delta, \Phi), \\ G \sim & p(G), \end{aligned}$$

where $\delta \in \mathbf{R}^+$ is a fixed degree-of-freedom parameter, Φ is a symmetric positive-definite scale matrix, and $p(G)$ is the mixing prior over decomposable graphs. The HIW distribution was introduced by Dawid and Lauritzen (1993) with the probability density function (pdf) given by

$$p(S|G, \delta, \Phi) = \frac{\prod_{k=1}^{K} p(S_{P_k}|\delta, \Phi_{P_k})}{\prod_{k=2}^{K} p(S_{Q_k}|\delta, \Phi_{Q_k})},$$

where P_k and Q_k are the cliques and separators of the graph G respectively. The terms $p(S_{P_k}|\delta, \Phi_{Q_k})$ denote the inverse Wishart (IW) density of $S_{P_k} \sim \text{IW}(\delta, \Phi_{P_k})$ with the pdf

$$p(S_{P_k}|\delta, \Phi_{P_k}) \propto |S_{P_k}|^{-(\delta/2+|P_k|)} \exp\Big\{ -\frac{1}{2} tr(S_{P_k}^{-1}\Phi_{P_k})\Big\}.$$

The HIW distribution is a conjugate prior distribution for the covariance matrix S. Specifically, if q-dimensional random variables $\mathbf{x_i}$ follow an iid multivariate normal distribution $\mathcal{N}_q(\mathbf{0}, S)$ for $i = 1, \ldots, n$, and S follows $\text{HIW}(G, \delta, \Phi)$, the posterior of S is $S|\mathbf{x}, G \sim \text{HIW}(G, \delta + n, \Phi + \mathbf{x}^T\mathbf{x})$. The closed form of the posterior distribution for S plays a key part in the posterior inference based on an MCMC algorithm. In our model, we consider $\delta = 3$ as reflecting the lack of prior information on S, and specify $\Phi = I_q$.

Let e_{ij} be a binary indicator denoting whether the edge (i, j) is included or excluded from the graphical model. One option for the mixing prior $p(G)$ is to assign an independent prior probability of an edge, $p(e_{ij}) \in (0,1)$, to each pair of nodes (i,j), so that $\pi(G) = \prod_{(i,j)\in E} p(e_{ij} = 1) \cdot \prod_{(i,j)\notin E} p(e_{ij} = 0)$. However, such priors with a uniform probability controlling the sparsity of the graph G are not satisfactory in performance. In this dissertation, we propose a new prior of G which induces adaptive sparsity in the graphical models as:

$$G \propto \exp(-|G|^{\xi}),$$

where ξ is a positive value penalizing on the size of the graph G. Varying ξ penalizes a graph size with different strength. A large value of ξ (> 3) constrains the graph to be extremely sparse, while a value of ξ near zero approximates a uniform prior on all graphs. In the Bayesian method, we assign a uniform prior between 0 and a large value (e.g. 5) on ξ, and estimate ξ using an MCMC algorithm. Such choice of prior lets the data choose the intensity of the penalization on the graph size and leads to adaptive sparsity in the inferred graph G. When the graph is restricted to be decomposable, the hierarchy for modeling the covariance of residuals in the graphical factor model can be summarized as follows:

$$\left.\begin{aligned} S &\sim \mathrm{HIW}(G,\delta,\Phi), \\ G &\sim \exp(-|G|^{\xi}), \\ \xi &\sim \mathrm{Unif}(0,5). \end{aligned}\right\} \tag{4.11}$$

4.5.4 Bayesian Hierarchical Model for Unrestricted Graphs

In this section, we assume the graphical model of the residuals is unrestricted. In Section 4.2.2, we employ a Bayesian graphical lasso selection prior to achieve sparse estimation of S. We now apply the Bayesian graphical lasso selection prior on C, the inverse of S. As mentioned above, the graphical lasso prior on a matrix C is equivalent to independent exponential priors on the diagonal elements C_{jj},

$j = 1, \ldots, q$, and double exponential priors on the off-diagonal elements $C_{jj'}$, $j < j'$. However, the Bayesian graphical lasso method does not set the off-diagonal elements in C to exact zeros. Hence similarly, we add a point mass at 0 in the double exponential priors. The priors are detailed as follows:

$$\begin{aligned} C_{jj}|\lambda^C &\sim \text{Exp}(\frac{\lambda^C}{2}), \quad j = 1, \ldots, q, \\ C_{jj'}|\lambda^C, \rho^C_{jj'} &\sim (1 - \rho^C_{jj'})\mathcal{I}\{C_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda^C), \quad j < j'. \end{aligned}$$

Again, we choose a conjugate Gamma hyper-prior distribution for the shrinkage parameter λ^C and a conjugate Beta hyper-prior distribution for the selection parameters $\rho^C_{jj'}$:

$$\begin{aligned} \lambda^C &\sim \text{Gamma}(a_\lambda, b_\lambda), \\ \rho^C_{jj'} &\sim \text{Beta}(a_\rho, b_\rho), \quad j < j'. \end{aligned}$$

As mentioned above, the hyper-parameter λ^C shrinks the covariance elements toward zero, while $\rho^C_{jj'}$ controls the probability that the (j, j') element will be enforced to be a zero. Based on our experiments, we find that noninformative priors on $\rho^C_{jj'}$ would lead to significant inaccuracy in estimating S, and influence the factor selection and loadings estimation. In this case, we specify $(a_\lambda, b_\lambda) = (1, 1)$ for a diffuse prior for λ^C, and $(a_\rho, b_\rho) = (1, q)$ for a sparse prior of $\rho^C_{jj'}$.

The hierarchical model for a sparse graphical factor model with unrestricted graphs can be summarized as follows:

$$\left.\begin{aligned} C_{jj} &\sim \text{Exp}(\frac{\lambda^C}{2}), \\ C_{jj'} &\sim (1 - \rho^C_{jj'})\mathcal{I}\{C_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda^C), \\ \lambda^C &\sim \text{Gamma}(a_\lambda, b_\lambda), \\ \rho^C_{jj'} &\sim \text{Beta}(a_\rho, b_\rho). \end{aligned}\right\} \tag{4.12}$$

4.5.5 Application

In this section, we applied the graphical factor analytic model to the analysis of simulated datasets as well as a real dataset. In the simulation, we considered three graphical factor models to generate datasets.

- Model 4: An initial simulation study considered a one-factor model for a 30-dimensional random vector. The factor loading vector M was randomly generated with $\|\mathbf{M}\| = 1$, the variance of the factor was set to be $\tau^2 = 4$, and the covariance matrix of the residuals S corresponded to an AR(1) model with $S_{jj'} = 0.7^{|j-j'|}$.

- Model 5: The second simulation considered a two-factor model for a 30-dimensional random vector. The $q \times 2$ factors loadings matrix $M = (\mathbf{M}_1, \mathbf{M}_2)$ with $|\{j : M_{j1} \neq 0\}| = q/2$, $\{j : M_{j2} \neq 0\} = \{1, \ldots, q\} \backslash \{j' : M_{1j'} \neq 0\}$, and $\|\mathbf{M}_1\| = \|\mathbf{M}_2\| = 1$. The variances of the factors were $\tau^2 = (4, 4)$, and the covariance matrix of the residuals S was a block diagonal matrix with each square block matrix B of dimension 5, and $B = 0.7\mathbf{1}\mathbf{1}^T + 0.3I$.

- Model 6: The third simulation considered a one-factor model for a 30-dimensional random vector. The factor loading vector $\mathbf{M}$ was randomly generated with $\|M\| = 1$, the variance of the factor was set to be $\tau^2 = 4$,, and the covariance matrix of the residuals S corresponds to a nondecomposable graphical model depicted in Figure 4.4.

We generated datasets with varying sample size $n = 100, 300$ for each model. The proposed graphical factor model with HIW prior method was used to recover the number of factors and the decomposable graph in model 4 and 5, and the graphical factor model with Bayesian graphical lasso method was used for model 6. The estimates of the Bayesian method for model 4 and 5 were based on the posterior samples of 10000 iterations after 5000 burn-in iterations, and the estimates of the

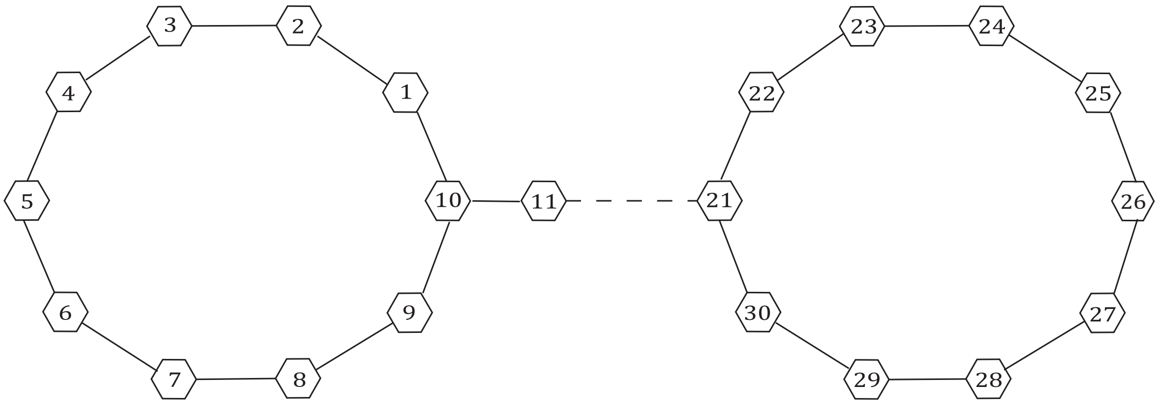


Fig. 4.4. Graphical structure of model 6 in the simulations in Section 4.5.5

Bayesian method for model 6 were based on the posterior samples of 5000 iterations after 1000 burn-in iterations. The estimates of the factor number were the posterior modes of the number of selected factors, and the graphical model estimates were based on an overall FDR rate of 0.20.

Table 4.3 summarizes the results in estimating the number of factors and the graphical model of the residuals over 20 replications. The results indicate that our Bayesian graphical factor analytic model can recover the number of true latent factors most of the times. Besides, it can recover the graphical model of the residuals with both low rates of false positives and false negatives when the sample size is sufficient relative to the dimension.

We also applied the graphical factor models to a microarray gene expression dataset from Liu et al. (2011), which was generated from 176 primary breast cancer patients. We focused on 26 mRNA transcripts whose coding genes are known to lie in the estrogen receptor (ER) pathway. The estrogen pathway regulates a variety of genes and plays key roles in the development or progression of breast carcinogenesis.

We analyzed the data with the two graphical factor models. The graphical factor model for decomposable graphs identifies two latent factors, and the model for

Table 4.3
Simulation results for Bayesian graphical factor analytic models for model 4, 5 and 6. The mean results over 20 replications are presented in the table with the standard deviations in parentheses. FP: false positive; FN: false negative. See Section 4.5.5 for details about the models.

Rank Recovery						
	model 4		model 5		model 6	
Sample Size	%(1 factor)	mean(se)	%(2 factor)	mean(se)	%(1 factor)	mean(se)
n=100	95	1.05 (0.22)	70	2.40 (0.68)	75	1.25 (0.44)
n=300	100	1.00 (0.00)	95	1.95 (0.22)	100	1.00 (0.00)
Graph of Residuals						
	model 4		model 5		model 6	
Sample Size	FN(se)	FP(se)	FN(se)	FP(se)	FN(se)	FP(se)
n=100	0.05 (0.22)	7.15 (0.49)	17.35 (3.69)	6.00 (1.68)	0.45 (1.05)	15.8 (2.75)
n=300	0.00 (0.00)	6.90 (0.31)	3.30 (1.59)	9.00 (1.86)	0.00 (0.00)	17.25 (1.80)

unrestricted graphs does not detect any latent factors. The adjacency matrices corresponding to the inferred graphical models are plotted in Figure 4.5, depicting the conditional dependency relationship among the variables. Some of the genes have multiple sets of oligonuleotide sequences on the microarray, and hence have multiple appearances including the estrogen receptor gene (ESR1a, ESR1b), MYBL1 (MYBL1a, MYBL1b, MYBL1c, MYBLd), TFF3 (TFF3a, TFF3b, TFF3c, TFF3d), XBP (XBP1a, XBP1b), and IGF1R (IGF1Ra, IGF1Rb).

The graph in Figure 4.5(a) is decomposable derived by the HIW method, and the graph Figure 4.5(b) is nondecomposable derived by the Bayesian graphical lasso method. The decomposable graph is much sparser than the nondecomposable graph, which is probably due to the restriction of decomposition using the HIW method. This difference in sparsity level resulting from the graph restriction also explains the different numbers of selected latent factors by the two methods. Comparison of the two graphs also indicates a lot of shared features, especially in the positive linking of multiple probe sets for one gene, which partially supports the inference results.

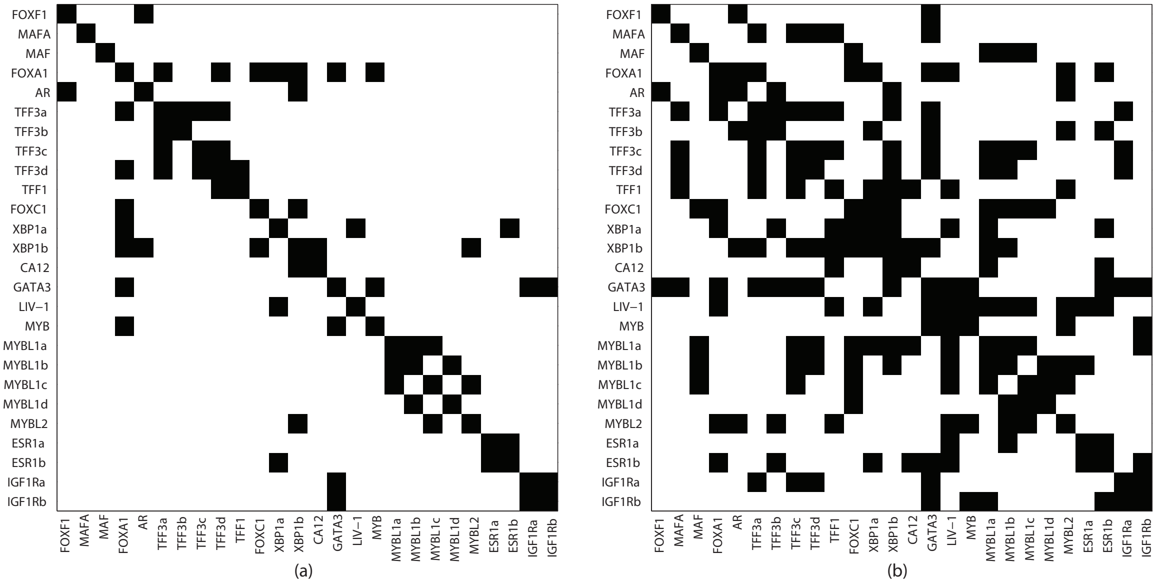


Fig. 4.5. Adjacency matrix of the genes involved in ER pathway depicting the graphical model of the residuals inferred by (a) Bayesian graphical factor model for decomposable graph (b) Bayesian graphical factor model for unrestricted graph. Some genes have multiple sets of oligonuleotide sequences on the microarray, and hence the appearance of multiples of some genes: estrogen receptor itself (ESR1a, ESR1b), MYBL1 (MYBL1a, MYBL1b, MYBL1c, MYBL1d), TF-F3 (TFF3a, TFF3b, TFF3c, TFF3d), XBP (XBP1a, XBP1b), and IGF1R (IGF1Ra, IGF1Rb).

4.6 Discussion

In this report, we propose a Bayesian method of estimating covariance matrices of a particular structure, which is a summation of a low rank and a sparse component. Different from the frequentist LOREC method, which is based on the sample covariance estimate only, our Bayesian method of covariance decomposition is likelihood-based. Hence, it takes the variability of the variables into consideration in case of $q >> n$ and shrinks the covariance elements of varying intensities differently, as indicated in the real data analysis. We model the low rank and sparse component in the form of a latent factor model with correlated residuals. The representation of the decomposable covariance facilitates a Bayesian inference by using conjugate priors on all the parameters except for the off-diagonal elements in the sparse component. Simulations indicate that such representation favors the covariance estimation for a latent factor model but does not perform as good as the frequentist LOREC method for a random effect model.

We further extend our method to a graphical factor analytic model, in which we perform inference on both the number of factors and the sparse graphical model of the residuals. Simulation studies show that the methods can successfully recover the number of factors as well as the graphical model when the sample size is sufficient relative to the dimension. However, simulations (not presented here) also indicate that in the case of $q >> n$, the methods tend to choose over-sparse graphical models of the residuals. This is reasonable: when the sample size is small, the estimate of the covariance of the residuals S would be inaccurate resulting in significant change in S inverse estimate. In high-dimensional low-sample-size condition, the Bayesian methods would choose the most sparse estimate of S inverse that fits the data.

5. CONCLUSIONS

We have considered three aspects of statistical analysis of genetic data, graphical network modeling, feature selection, and covariance estimation. We propose new statistical approaches to derive information from the datasets while accounting for different natures of the genetic data.

In Section 2, we focus on developing methods for graphical modeling of count data from SAGE experiments. We propose a generalized linear model to fit the count data with the log ratios of the mean levels assumed to follow a Gaussian distribution. We model the underlying gene network with the prior on the covariance matrix to be a mixture of HIW distributions, and derive graph structures with an FDR-based method, which thresholds posterior inclusion probabilities of edges in MCMC sampling. A prior on graphs is derived based on gene relatedness measured by shared GO terms, which takes existing biological information as a priori.

In the next section, we concern feature selection for chromosomal copy number data from MIP experiments, which consist of variables with natural grouping structures. The goal is to select genes at group level as well as probes at within-group level that are correlated to clinically relevant outcomes simultaneously. We propose the HSVS model, which utilizes a discrete mixture prior distribution for group selection and group-specific Bayesian lasso hierarchies for variable selection within groups. Considering the potential serial correlations among the probes with a gene, we further provide method incorporating Bayesian fused lasso methods for within-group selection.

In Section 5, we consider covariance estimation of Gaussian-distributed variables, whose variation can be partially explained by a small number of common factors or principal components, and the covariance conditional on these latent common factors is sparse. In this case, the covariance matrix can be decomposed into a low rank and sparse component. We model such structured covariance in the form of a factor analytic model where the number of latent factors is unknown, and introduce binary indicators with a strong selection prior to estimate the rank of the low rank component combined with a Bayesian graphical lasso method for the sparse component estimation. The Bayesian covariance decomposition method is further extended to a graphical factor analytic model where sparsity in inverse covariance of the residuals is desired. We use the method to model the conditional independence structures among the variables while accounting for potential latent factors underlying the variables. We consider both decomposable graph modeling with hyper-inverse Wishart prior methods and nondecomposable graph modeling with Bayesian hierarchical graphical lasso methods.

We can further develop the proposed models accounting for varying characteristic of data. For example, we may model the gene network of the count data from SAGE with negative binomial or overdispersed Poisson instead of Poisson distribution to adjust for the overdispersion existing in SAGE data. In the copy number data analysis, we can extend the HSVS method for the existence of overlapping coding regions in the hierarchical structured data. We leave these tasks for future studies.

REFERENCES

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

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669-679.

Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., Richardson, A., Schnitt, S., Sellers, W. R., and Polyak, K. (2004). Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* **6(1),** 17-32.

Andrew, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distribution. *Journal of the Royal Statistical Society: Series B* **36,** 99-102.

Armstrong, H., Carter, C., Wong, K., and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistical Computation* **19(3),** 303-316.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P. et al. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25,** 25-29.

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

Bae, K. and Mallick, B. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20,** 3423-3430.

Baladandayuthapani, V., Ji, Y., Talluri, R., Neito-Barajas, L. E., and Morris J. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *Journal of the American Statistical Association* **105,** 390-400.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57(1),** 289-300.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807-820.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27,** 359-367.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* **64,** 115-123.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2,** 369-380.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9-25.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B* **65,** 39.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* **57,** 473-484.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438-1456.

Chinnadurai, G. (2006). CtIP, a candidate tumor susceptibility gene is a team player with luminaries. *Biochimica et Biophysica Acta* **1765,** 67-73.

Chiu, T. Y. M., Leonard, T., and Tsui, K. W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association* **91**, 198-210.

Cubas, R., Zhang, S., Li, M., Chen, C., and Yao, Q. (2010). Trop2 expression contributes to tumor pathogenesis by activating the ERK MAPK pathway. *Molecular Cancer* **9,** 253.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21 (3),** 1272-1317.

Deng, X., and Tsui, K. W. (2010). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics.* Accepted.

Dittmer, J. (2003). The biology of the Ets1 proto-oncogene. *Molecular Cancer* **2,** 29.

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

Eckerdt, F., Yuan, J., Strebhardt, K. (2005). Polo-like kinases and oncogenesis. *Oncogene* **24,** 267-76.

Efron, B., Hastie, T., Johnstone I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **2,** 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348-1360.

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35,** 109-148.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9(3)**, 432-441.

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,** 339-374.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics* 4, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith). Clarendon Press, Oxford.

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

Gibson, S. L., Ma, Z., Shaw, L. M. (2007). Divergent roles for IRS-1 and IRS-2 in breast cancer metastasis. *Cell Cycle* **6,** 631-637.

Giudici, P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics* 5, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith). Oxford University Press, Oxford.

Giudici P. (2001). Bayesian inference for graphical factor analysis models. *Psychometrika* **66(4)**, 577-592.

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

Ghosh, A. K. (2002). Factors involved in the regulation of type I collagen gene expression: implication in fibrosis. *Experimental Biology and Medicine* **227,** 301-314.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82(4),** 711-732.

Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. Technical Report, Department of Statistics, University of Warwick, Coventry, UK.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5,** 171-188.

Grzebyk, M., Wild, P. and Chouanière, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91(1)**, 141-151.

Guha S., Li Y., and Neuberg D. (2008). Bayesian hidden markov modeling of array CGH data. *Journal of the American Statistical Association* **103,** 485-497.

Hardenbol, P., Ban è r, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad, H. et al. (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* **21,** 673-678.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14,** 382-417.

Holmes, C. C. and Held, L. (2006).Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1,** 145-168.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85-98.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36,** 949-951.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100,** 764-780.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya Series B* **60,** 65-81.

Kustra, R., Zagdanski, A. (2006). Incorporating gene ontology in clustering gene expression data. *19th IEEE Symposium on Computer-Based Medical Systems, IEEE Computer Society* 555-563.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5,** 369-412.

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

Lawler, J. (2002). Thrombospondin-1 as an endogenous inhibitor of angiogenesis and tumor growth. *Journal of Cellular Molecular Medicine* **6,** 1-12.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88(2)**, 365-411.

Leonard, T. and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics* **20**, 1669-1696.

Levina, E., Rothman, A. J., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics* **2(1)**, 245-263.

Liu, R. Z., Graham, K., Glubrecht, D. D., Germain, D. R., Mackey, J. R., and Godbout, R. (2011). Association of FABP5 expression with poor survival in triple-negative breast cancer: Implication for retinoic acid therapy. *The American Journal of Pathology* **178(3)**, 997-1008.

Lopes, H. F. and Carvalho, C. M. (2007). Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference* **137**, 3082-91.

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

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brandy, A., Sebat, J. et al. (2003). Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Research* **13,** 2291-2305.

Luo, X. (2011). High dimensional low rank and sparse covariance matrix estimation via convex minimization. arXiv:1111.1133v1

Ma, S., Zhang, Y., Huang, J., Han, X., Holford, T., Lan, Q., Rothman, N., Boyle, P., and Zheng, T. (2010). Identification of non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. *Bioinformatics* **26,** 15-21.

Mitchell, T. J., and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83,** 1023-1032.

Morris, J. S., Brown, P. J., Herrick, R. C., Beggerly, K. A., and Coombes K. R. (2008). Bayesian analysis of mass spectrometry data using wavelet-based functional mixed models. *Biometrics* **64,** 479-489.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association* **99,** 990-1001.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103,** 681-686.

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W. L., Chen C., Zhai Y. et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20,** 207-211.

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37,** Suppl: S11-7.

Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. (2001). A SAGE view of breast tumor progression. *Cancer Rearch* **61**.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677-90.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425-35.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92,** 179-191.

Raman, S., Fuchs, T., Wild, P., Dahl, E., and Roth, V. (2009). The Bayesian group-lasso for analyzing contingency tables. *Proceedings of the 26th International Conference on Machine Learning.* 881-888.

Rennstam, K., Ahlstedt-Soini, M., Baldetorp, B., Bendahl P. O., Borg, A., Karhu, R., Tanner, M., Tirkkonen, M., and Isola, J. (2003). Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization. *Cancer Research* **63,** 8861-8868.

Resnik, R. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* **11,** 95-130.

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.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics.* Chapman and Hall/CRC Press, London.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38,** 2587-2619.

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

Sipes, J. M., Guo, N., Nègre, E., Vogel, T., Krutzsch, H. C., and Roberts, D. D. (1993). Inhibition of fibronectin binding and fibronectin-mediated cell adhesion to collagen by a peptide from the second type I repeat of thrombospondin. *Journal of Cell Biology* **121(2),** 469-477.

Stanghellini, E. (1997). Identification of a single-factor model using graphical Gaussian rules. *Biometrika* **84(1)**, 241-244.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31,** 2013-2035.

Talluri, R., Baladandayuthapani, V. and Mallick, B. K. (2011). Bayesian sparse graphical models and their mixtures using lasso selection priors. *Journal of the American Statistical Association.* Submitted.

Thompson, P. A., Brewster, A., Broom, B., Do, K-A, Baladandayuthapani, B., Edgerton, M., Hahn, K., Murray, J., Sahin, A., Tsavachidis, S., Wang, Y., Zhang, L., Hortobagyi, G., Mills, G. and Bondy, M. (2011). Selective genomic copy number imbalances and probability of recurrence in early-stage breast cancer. *PLos One* **6(8),** e23543.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58,** 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67,** 91-108.

van Beers E. H., and Nederlof, P. M. (2006). Array-CGH and breast cancer. *Breast Cancer Research* **8,** 210.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270(5235),** 484-487.

Wang, H. (2012) Bayesian graphical lasso models and efficient posterior computation. Working. Department of Statistics, University of South Carolina, Columbia, South Carolina.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23,** 1274-1281.

Wang, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression for censored data with grouped variables. *Biometrika* **96,** 307-322.

Wang, Y., Moorhead, M., Karlin-Neumann, G., Wang, N. J., Ireland, J., Lin, S., Chen, C., Heiser, L. M., Chin, K. et al. (2007). Analysis of molecular inversion probe performance for allele copy number determination. *Genome Biology* **8,** R246.

Weber, M., Sydlik, C., Quirling, M., Nothdurfter, C., Zwergal, A., Heiss, P., Bell, S., Neumeier, D., Ziegler-Heitbrock, H. W., and Brand, K. (2003). Transcriptional inhibition of Interleukin-8 expression in tumor necrosis factor-tolerant cells: evidence for involvement of C/EBPβ. *Journal of Biological Chemistry* **278(26),** 23586-23593.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74,** 646-648.

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

Wu, J., Mei, C., Vlassara, H., Striker, G. E., and Zheng, F. (2009). Oxidative stress-induced JNK activation contributes to proinflammatory phenotype of aging diabetic mesangial cells. *American Journal of Physiology: Renal Physiology* **297(6),** F1622-F1631.

Wu, X., Ye, Y., and Subramanian, K. R. (2003). Interactive analysis of gene interactions using graphical Gaussian model. *ACM SIGKDD Workshop on Data Mining in Bioinformatics* **3,** 63-69.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163,** 789-801.

Yamamoto, G., Nannya, Y., Kato, M., Sanada, M., Levine, R. L., Kawamata, N., Hangaishi, A., Kurokawa, M., Chiba, S. et al. (2007). Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *American Journal of Human Genetics* **81,** 114-126.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26(7),** 976-978.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68,** 49-67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94(1)**, 19-35.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association* **86,** 79-86.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics* **37,** 3468-3497.

APPENDIX A
POSTERIOR INFERENCE FOR GRAPHICAL MODEL INFERENCE FOR DISCRETE GENE EXPRESSION DATA

In this section, we present the full conditional posterior distributions and the MCMC algorithm to generate posterior samples of the parameters in the Bayesian hierarchical model for discrete gene network inference. With the hierarchical model specified in Section 2.2.2, the full conditional posterior distributions for Gibbs sampling are as follows:

$$
\begin{aligned}
\theta_{ij}|X, \boldsymbol{\mu}, \Sigma \quad &\propto \quad \exp\left\{-\frac{(\theta_{ij}-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2} - e^{\theta_{ij}} + t_{ij}X_{ij}\right\}, \\
& \qquad \tilde{\mu}_j = \mu_j + \Sigma_{j(-j)}\Sigma^{-1}_{(-j)(-j)}(\boldsymbol{\theta}_{i(-j)} - \boldsymbol{\mu}_{(-j)}), \\
& \qquad \tilde{\sigma}_j^2 = \sigma_j^2 - \Sigma_{j(-j)}\Sigma^{-1}_{(-j)(-j)}\Sigma_{(-j)j}, \\
G|\theta, r \quad &\propto \quad \frac{h(G,\delta,rI_q)}{h(G,\delta+n-1,rI_q+S_\theta)} \cdot \pi(G), \\
& \qquad S_\theta = \sum_{i=1}^{n}(\boldsymbol{\theta_i} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta_i} - \bar{\boldsymbol{\theta}})', \\
r|\theta, G \quad &\propto \quad \frac{h(G,\delta,rI_q)}{h(G,\delta+n-1,rI_q+S_\theta)}, \\
\boldsymbol{\mu}|\theta, \Sigma \quad &\sim \quad \mathcal{N}_q(\bar{\boldsymbol{\theta}}, \Sigma/n), \\
\Sigma|\theta, r, G \quad &\sim \quad \text{HIW}(G, \delta+n-1, rI_q+S_\theta),
\end{aligned}
$$

where $\Sigma_{j(-j)}$ is the j^{th} row vector of Σ excluding the j^{th} element, and likewise with $\Sigma^{-1}_{(-j)(-j)}$, $\boldsymbol{\theta}_{i(-j)}$, and $\boldsymbol{\mu}_{(-j)}$.

The parameters $\boldsymbol{\mu}$ and Σ can be sampled from the conditional distributions in closed form, while the parameters θ_{ij}'s, r, and G are generated using a random walk Metropolis-Hastings (MH) algorithm within the Gibbs sampling. The details of the MH algorithm for θ_{ij}'s, r and G are as follows:

- ◇ MCMC for θ_{ij}: Given the current state θ_{ij}^c, generate a proposal θ_{ij}^p from a normal distribution $\mathcal{N}(\theta_{ij}^c, \sigma_\theta^2)$, with the standard deviation σ_θ chosen properly for fast moving. The proposal θ_{ij}^p is accepted with the probability

$$\alpha(\theta_{ij}^c, \theta_{ij}^p) = \min\left\{1, \frac{f(\theta_{ij}^p|X, \boldsymbol{\mu}^c, \Sigma^c)}{f(\theta_{ij}^c|X, \boldsymbol{\mu}^c, \Sigma^c)}\right\},$$

  where f denotes the full conditional distribution of θ_{ij}.

- ◇ MCMC for r: Given the current state r^c, generate a proposal $\log(r^p)$ from a normal distribution $\mathcal{N}(\log(r^c), \sigma_r^2)$, with the standard deviation σ_r chosen properly. We generate the MCMC samples of r in the log scale to ensure positivity. The proposal r^p is then accepted with the probability

$$\alpha(r^c, r^p) = \min\left\{1, \frac{f(r^p|\theta^c, G^c)}{f(r^c|\theta^c, G^c)}\right\},$$

  where f denotes the full conditional distribution of r.

- ◇ MCMC for G: Let $\{e_{ij} : i < j\}$ be the set of edge indicators where $e_{ij} = 1$ if $(i, j) \in E$ and $e_{ij} = 0$ otherwise. In an iteration with the current state of graph $G^c = (V, E^c)$, we choose a pair (i, j) at random and change the state of the edge, i.e. $e_{ij}^p = 1 - e_{ij}^c$. If the proposed state $G^p = (V, E^p)$ is decomposable, the proposal is accepted as a new state with the probability

$$\alpha(G^c, G^p) = \min\left\{1, \frac{f(G^c|\theta^c, r^c)}{f(G^p|\theta^c, r^c)}\right\},$$

  where f denotes the full conditional distribution of G. If the proposed state $G^p = (V, E^p)$ is not decomposable, then choose another pair (i, j) until the proposal graph is decomposable.

Notice from the above formulations that the conditional distributions of G and r for Gibbs sampling is marginalized over $\boldsymbol{\mu}$ and Σ. Hence $\boldsymbol{\mu}$ and Σ are not necessary for sampling the graph model G, resulting in a more efficient collapsed sampler.

However, posterior Σ samples are still generated in our MCMC algorithm to facilitate the sampling of θ_{ij}'s.

APPENDIX B
POSTERIOR INFERENCE FOR BAYESIAN HIERARCHICAL STRUCTURED VARIABLE SELECTION METHODS

In this section we present the full conditional distributions for the Bayesian HSVS methods and the Gibbs sampling algorithm used to update the parameters.

B.1 The HSVS method

With the hyperpriors assigned as described in Section 2.1, the full conditional posterior distributions are in closed form as follows:

- $\beta_g | \cdot \sim (1-\gamma_g)\delta_{\{\boldsymbol{\beta}_g = \mathbf{0}_{k_g}\}} + \gamma_g \mathcal{N}_{k_g}(\tilde{\sigma}_g^2 \tilde{\mu}_g, \tilde{\sigma}_g^2 \sigma^2)$, where

$$\tilde{\mu}_g = X_g^T(\mathbf{Y} - X_{(g)}\boldsymbol{\beta}_{(g)})$$

$$\tilde{\sigma}_g^2 = \left\{ \begin{bmatrix} \tau_{g1}^{-2} & \cdots & 0 \\ \vdots & 0 \ddots & \vdots \\ 0 & \cdots & \tau_{g_{k_g}}^{-2} \end{bmatrix} + X_g^T X_g \right\}^{-1}$$

- $\gamma_g | \cdot \sim \text{Bernoulli}(p_g)$, where

  $p_g = 1 - (1-p)/\{1 - p + p \cdot \frac{|\tilde{\sigma}_g^2|^{0.5}}{\tau_{g1} \cdots \tau_{gk_g}} \cdot \exp(\frac{\tilde{\mu}_g^T \tilde{\sigma}_g^2 \tilde{\mu}_g}{2\sigma^2})\}$

- $p | \cdot \sim \text{Beta}(\sum_g \gamma_g + a, \sum_g (1-\gamma_g) + b)$,
- $\tau_{gj}^2 | \cdot \sim (1-\gamma_g)\text{Exp}(\frac{\lambda_g^2}{2}) + \gamma_g \text{GIG}(\frac{1}{2}, \lambda_g^2, \frac{\beta_{gj}^2}{\sigma^2})$,
- $\lambda^2 | \cdot \sim \text{Gamma}(r + k_g, \delta + \sum_j \tau_{gj}^2)$,
- $\sigma^2 | \cdot \sim \text{Inv} - \text{Gamma}(\frac{n-1+p}{2}, \frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta}) + \frac{1}{2}\sum_g \boldsymbol{\beta_g}' D_{\tau_g} \boldsymbol{\beta_g})$.

In the conditional distribution of τ^2_{gj}, GIG is the abbreviation for "generalized inverse Gaussian" distribution. A random variable x following a GIG distribution with the parameter set (p, a, b) has a probability density function as follows:

$$f(x|p, a, b) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\{-\frac{ax + b/x}{2}\},$$

where K_p is a modified Bessel function.

The closed form of full conditional distributions allows for an efficient Gibbs sampler, with block updating of $(\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_G})$, σ^2, $(\gamma_1, ..., \gamma_G)$, p , $(\tau^2_{11}, ..., \tau^2_{Gk_G})$, and $(\lambda_1, ..., \lambda_G)$ in sequence. Particularly, in the t^{th} iteration, $\tau^{2(t)}_{gj} (j = 1, ..., k_g)$ is drawn from GIG$(\frac{1}{2}, \lambda^{2(t-1)}_g, \beta^{2(t)}_{gj}/\sigma^{2(t)})$ if $\gamma^{(t)}_g = 1$; else, $\tau^{2(t)}_{gj}$ is drawn from Exp$(\lambda^{2(0)}_g/2)$. This is a case of the reversible jump MCMC (Green 1995), where the number of parameters changes when the value of γ_g switches between 0 and 1. Here we use a pseudo-prior algorithm, introduced in Carlin and Chib (1995), by sampling $\tau^{2(t)}_{gj}$ from Exp$(\lambda^{2(0)}_g/2)$.

B.2 The fused-HSVS method

For the fused-HSVS model, the resulting full conditional posterior distributions are also in closed form:

- $\beta_g| \cdot \sim (1 - \gamma_g)\delta_{\{\boldsymbol{\beta}_g = 0_{k_g}\}} + \gamma_g \mathcal{N}_{k_g}(\tilde{\sigma}^2_g \tilde{\mu}_g, \tilde{\sigma}^2_g \sigma^2)$, where

$$\tilde{\mu}_g = X^T_g(\mathbf{Y} - X_{(g)}\boldsymbol{\beta}_{(g)})$$
$$\tilde{\sigma}^2_g = \left\{\Sigma^{-1}_{\beta_g} + X^T_g X_g\right\}^{-1}$$

- $\gamma_g| \cdot \sim$ Bernoulli(p_g), where

$$p_g = 1 - (1 - p)/\{1 - p + p \cdot |\Sigma_{\beta_g}|^{-0.5} |\tilde{\sigma}^2_g|^{0.5} \exp(\tfrac{\tilde{\mu}^T_g \tilde{\sigma}^2_g \tilde{\mu}_g}{2\sigma^2})\}$$

- $\tau^2_{gj}| \cdot \sim (1 - \gamma_g)\text{Exp}(\frac{\lambda^2_{1g}}{2}) + \gamma_g \text{GIG}(\frac{1}{2}, \lambda^2_{1g}, \frac{\beta^2_{gj}}{\sigma^2}),$ for $j = 1, ..., k_g$
- $\omega^2_{gj}| \cdot \sim (1 - \gamma_g)\text{Exp}(\frac{\lambda^2_{2g}}{2}) + \gamma_g \text{GIG}(\frac{1}{2}, \lambda^2_{2g}, \frac{(\beta_{gj} - \beta_{g(j-1)})^2}{\sigma^2}),$ for $j = 1, ..., k_g - 1$.

With the above full conditional distributions, we use the Gibbs algorithm, with block updating of $(\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_G})$, σ^2, $(\gamma_1, ..., \gamma_G)$, p , $(\tau_{11}^2, ..., \tau_{Gk_G}^2)$, and $(\omega_{11}^2, ..., \omega_{G,k_G-1}^2)$ in sequence. As with the HSVS model, we impose a Beta(a,b) prior on p, a Gamma(r_1, δ_1) prior on λ_{1g}'s, and a Gamma(r_2, δ_2) prior on λ_{2g}'s. These hyperpriors lead to conjugate full conditional distributions Beta$(\sum_g \gamma_g + a, \sum_g (1 - \gamma_g) + b)$, Gamma$(r_1 + k_g, \delta_1 + \sum_j \tau_{gj}^2)$, and Gamma$(r_2 + k_g - 1, \delta_2 + \sum_j \omega_{gj}^2)$, respectively. Thus $(p, \lambda_{1g}, \lambda_{2g})$ can simply join the other parameters in the Gibbs sampler.

B.3 The generalized-HSVS method

In a generalized linear model framework as specified in Section 3, we have the full conditional posterior distribution of the latent variables Z_i, $i = 1, ..., n$ as follows:

$$\begin{aligned} Z_i|\boldsymbol{\beta}, Y_i = 1 \quad &\propto \mathcal{N}(X'\boldsymbol{\beta}, 1) \quad \text{truncated at the left by 0,} \\ Z_i|\boldsymbol{\beta}, Y_i = 0 \quad &\propto \mathcal{N}(X'\boldsymbol{\beta}, 1) \quad \text{truncated at the right by 0.} \end{aligned}$$

The full conditional distribution of Z_i is a truncated normal. Hence Z_i's can participate in the Gibbs sampling described in Sections B.1 and B.2.

APPENDIX C
POSTERIOR INFERENCE FOR BAYESIAN COVARIANCE DECOMPOSITION

C.1 Posterior sampling schemes for Bayesian covariance decomposition

In this section, we present the full conditional posterior distributions and a Gibbs sampling algorithm to generate posterior samples of the parameters. Note that all the parameters have closed-form full conditional distributions except for the elements in the sparse matrix S. In this case, we employ the Metropolis Hastings (MH) algorithm to simulate the parameters in S.

Sampling the factor loadings matrix M:

Let $\mathbf{M}_k = (m_{1k}, \ldots, m_{qk}))^T$ be the k^{th} column vector of M, and $M_{(-k)}$ be the matrix of M excluding the k^{th} column. The full conditional distribution of $\mathbf{M}_k$ for $k = 1, \ldots, r$ is

$$\mathbf{M}_k|y, M_{(-k)}, f, S \sim \mathcal{N}_q(\boldsymbol{\mu}_k^M, \Sigma_k^M),$$

where $\Sigma_k^M = \left\{S^{-1}(\sum_{i=1}^n f_{ki}^2)+qI_q\right\}^{-1}$, and $\boldsymbol{\mu}_k^M = \Sigma_k^M S^{-1}(y-M_{(-k)}f_{(-k)\cdot})\mathbf{f}_{k\cdot}^T$. Hence we can draw samples of each column of M from a multivariate Gaussian distribution.

Sampling the random factors f:

Let f be an $r \times n$ matrix with f_{ki} to be the value of k^{th} factor in i^{th} replicate. Then $\mathbf{f}_{k\cdot} = (f_{k1}, \ldots, , f_{kn})$ is the k^{th} row vector of f, and $f_{(-k)\cdot}$ denotes the matrix of f excluding the k^{th} row. Note $\mathbf{f}_{k\cdot}$ could be viewed as the unobserved values of the random factor k in the n replicates. The full conditional distribution of the transpose of $\mathbf{f}_{k\cdot}$ for $k = 1, \ldots, r$ is

$$\mathbf{f}_{k\cdot}^T|y, M, f_{(-k)\cdot}, S, z_k, \tau_k^2 \sim (1-z_k)\mathcal{N}_q(0, \tau_k^2 I_n) + z_k\mathcal{N}_n(\boldsymbol{\mu}_k^f, \sigma_k^f I_n),$$

where $\sigma_k^f = \Big(M_k^T S^{-1} M_k + \tau_k^{-2}\Big)^{-1}$, and $\boldsymbol{\mu}_k^f = \sigma_k^f \Big\{ M_k^T S^{-1}(y - M_{(-k)} f_{(k)\cdot}) \Big\}^T$. Hence we can draw samples of each row of f from a mixed multivariate Gaussian distribution conditional on the binary indicator z_k.

Sampling the binary diagonal matrix Z:

The full conditional of each diagonal element of Z, z_k, is a Bernoulli distribution:

$$z_k | y, M, f, S, \tau_k^2, p_k \sim \text{Bernoulli}(p_k^*),$$

where $p_k^* = 1 - (1-p_k)/\Big\{1 - p_k + p_k \cdot \Big(\frac{\sigma_k^f}{\tau_k^2}\Big)^{\frac{n}{2}} \cdot \exp\Big(-\frac{1}{2}\frac{(\boldsymbol{\mu}_k^f)^T \boldsymbol{\mu}_k^f}{\sigma_k^f}\Big)\Big\}$.

Sampling probabilities p_k and π:

The full conditional of p_k for $k = 1, \ldots, r$ is

$$\begin{aligned} &p_k | z_k = 1 \sim \text{Beta}(a_p + 1, b_p), \\ &p_k | z_k = 0 \sim (1 - \pi^*)\mathcal{I}\{p_k = 0\} + \pi^* \text{Beta}(a_p, b_p + 1), \end{aligned}$$

where $\pi^* = \frac{\pi b_p}{a_p + b_p - \pi a_p}$.

The full conditional of π is

$$\pi \sim \text{Beta}\left(a_\pi + \sum_k \mathcal{I}\{p_k = 0\}, b_\pi + \sum_k \mathcal{I}\{p_k \neq 0\}\right).$$

Sampling the positive diagonal matrix D_τ:

The full conditional of each diagonal element of D_τ, τ_k^2, for $k = 1, \ldots, r$ is

$$\tau_k^2 | z_k, \mathbf{f}_{k\cdot} \sim (1 - z_k)\text{IG}(a_\tau, b_\tau) + z_k \text{IG}\left(a_\tau + \frac{n}{2}, b_\tau + \frac{\mathbf{f}_{k\cdot}\mathbf{f}_{k\cdot}^T}{2}\right).$$

Sampling the diagonal elements S_{jj}:

The full posterior conditional density of the diagonal element S_{jj}, for $j = 1, \dots, q$, is

$$p(S_{jj}| \cdot\,) \propto \det(S)^{-n/2} \exp\{-\frac{1}{2}tr(S^{-1}\Lambda) - \frac{\lambda}{2}S_{jj}\}.$$

Without loss of generality, suppose that $j = p$. Let $S = \begin{bmatrix} S_{-j,-j} & S_{-j,j} \\ S_{j,-j} & S_{jj} \end{bmatrix}$.

With the property of matrices, we have

$$\begin{aligned} \det(S) =& \det(S_{-j}) \cdot \det(S_{jj} - S_{j,-j}S^{-1}_{-j,-j}S_{-j,j}), \\ \propto& (S_{jj} - c), \text{ where } c = S_{j,-j}S^{-1}_{-j,-j}S_{-j,j}. \end{aligned}$$

$$\begin{aligned} S^{-1} =& \begin{bmatrix} S_{-j} + S^{-1}_{-j,-j}S_{-j,j}(S_{jj} - c)^{-1}S_{j,-j}S^{-1}_{-j,-j} & -S^{-1}_{-j,-j}S_{-j,j}(S_{jj} - c)^{-1} \\ -(S_{jj} - c)^{-1}S_{j,-j}S^{-1}_{-j,-j} & (S_{jj} - c)^{-1} \end{bmatrix}, \\ =& \begin{bmatrix} S^{-1}_{-j,-j} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} S^{-1}_{-j,-j}S_{-j,j} \\ -1 \end{bmatrix} (S_{jj} - c)^{-1} \begin{bmatrix} S_{j,-j}S^{-1}_{-j,-j} & -1 \end{bmatrix}. \end{aligned}$$

$$\begin{aligned} tr(S^{-1}\Lambda) =& tr\Big(\begin{bmatrix} S^{-1}_{-j,-j} & 0 \\ 0 & 0 \end{bmatrix} \Lambda\Big) + tr\Big(\begin{bmatrix} S^{-1}_{-j,-j}S_{-j,j} \\ -1 \end{bmatrix} (S_{jj} - c)^{-1} \begin{bmatrix} S_{j,-j}S^{-1}_{-j,-j} & -1 \end{bmatrix} \Lambda\Big), \\ =& tr\Big(\begin{bmatrix} S^{-1}_{-j,-j} & 0 \\ 0 & 0 \end{bmatrix} \Lambda\Big) + d(S_{jj} - c)^{-1}, \\ & \text{where } d = \begin{bmatrix} S_{j,-j}S^{-1}_{-j} & -1 \end{bmatrix} \Lambda \begin{bmatrix} S^{-1}_{-j,-j}S_{-j,j} \\ -1 \end{bmatrix}. \end{aligned}$$

Hence, we have

$$p(S_{jj}| \cdot\,) \propto (S_{jj} - c)^{-n/2} \exp\{-\frac{d}{2}(S_{jj} - c)^{-1} - \frac{\lambda}{2}S_{jj}\}\mathcal{I}\{S_{jj} > c\}.$$

The above distribution does not seem to be in a closed form. We now transform S_{jj} to $\nu = S_{jj} - c$, then the conditional density of ν is

$$p(\nu|\cdot) \propto \nu^{-n/2} \exp\{-(d/\nu + \lambda\nu)/2\}\mathcal{I}\{\nu > 0\},$$

which is a generalized inverse Gaussian distribution with parameters $(1 - n/2, d, \lambda)$. Therefore, at each MCMC iteration, we can draw a sample of ν from the generalized inverse Gaussian distribution, and obtain $S_{jj} = \nu + c$.

Sampling the off-diagonal elements $S_{jj'}$:

The full posterior conditional density of the off-diagonal element $S_{jj'}$, $j < j'$ is

$$p(S_{jj'}|\cdot) \propto \det(S)^{-n/2} \exp\{-\frac{1}{2}tr(S^{-1}\Lambda)\}((1-\rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'}\frac{\lambda}{2}\exp(-\lambda|S_{jj'}|)).$$

Without loss of generality, suppose that $j = p-1$ and $j' = p$. Let $S = \begin{bmatrix} S_{-(jj')} & S_{-(jj'),jj'} \\ S_{jj',-(jj')} & S_{jj',jj'} \end{bmatrix}$, where $S_{jj',jj'} = \begin{bmatrix} S_{jj} & S_{jj'} \\ S_{jj'} & S_{j'j'} \end{bmatrix}$.

With the property of matrices, we have

$$\det(S) \propto \det(S_{jj',jj'} - B), \text{ where } B = S_{jj',-(jj')}S^{-1}_{-(jj')}S_{-(jj'),jj'},$$

$$S^{-1} = \begin{bmatrix} S^{-1}_{-(jj')} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} S^{-1}_{-(jj')}S_{-(jj'),jj'} \\ -I_2 \end{bmatrix} (S_{jj'} - B)^{-1} \begin{bmatrix} S_{jj',-(jj')}S^{-1}_{-(jj')} & -I_2 \end{bmatrix},$$

$$tr(S^{-1}\Lambda) = tr(\begin{bmatrix} S^{-1}_{-(jj')} & 0 \\ 0 & 0 \end{bmatrix} S) + tr((S_{jj',jj'} - B)^{-1}D),$$

$$\text{where } D = \begin{bmatrix} S_{jj',-(jj')}S^{-1}_{-(jj')} & -I_2 \end{bmatrix} \Lambda \begin{bmatrix} S^{-1}_{-(jj')}S_{-(jj'),jj'} \\ -I_2 \end{bmatrix}.$$

Hence, we have

$$
\begin{aligned}
p(S_{jj'}|\cdot) \propto & \det(S_{jj',jj'} - B)^{-n/2} exp\{-\frac{1}{2}tr((S_{jj'} - B)^{-1}D)\} \times p(S_{ij}|\lambda, \rho_{jj'}), \\
\propto & \exp\left\{-\frac{1}{2}\frac{(S_{j'j'} - B_{22})D_{11} + (S_{jj} - B_{11})D_{22} - 2(S_{jj'} - B_{12})D_{12})}{(S_{jj} - B_{11})(S_{j'j'} - B_{22}) - (S_{jj'} - B_{12})^2}\right\} \\
& \cdot \left\{1 - \frac{(S_{jj'} - B_{12})^2}{(jj'_{jj} - B_{11})(S_{j'j'} - B_{22})}\right\}^{-n/2} \cdot \{(1 - \rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'}\frac{\lambda}{2}\exp(-\lambda|S_{jj'}|)\}.
\end{aligned}
$$

We transform $S_{jj'}$ to $\nu = S_{jj'} - B_{12}$ and let $a = S_{jj} - B_{11}$, $b = S_{j'j'} - B_{22}$, then the conditional density of ν is

$$
p(\nu|\cdot) \propto (1 - \rho_{jj'})g(\nu)\mathcal{I}\{\nu = -B_{12}, \nu^2 < ab\} + \frac{\rho_{jj'}\lambda}{2}g(\nu),
$$

where $g(\nu) = (1 - \frac{\nu^2}{ab})^{-n/2} \exp\left\{-\frac{bD_{11}+aD_{22}-2D_{12}\nu}{2(ab-\nu^2)} - \lambda|\nu + B_{12}|\right\}\mathcal{I}\{\nu^2 < ab\}$.

The continuous part of the conditional distribution of ν, $g(\nu)$, cannot be sampled directly. Furthermore, $g(\nu)$ is nonconcave and therefore the sampler may be trapped locally if we use the random-walk MH algorithm within the Gibbs sampling. Since $g(\nu)$ only has density over $(-\sqrt{ab}, \sqrt{ab})$ and is zero elsewhere, we construct a piecewise uniform proposal distribution approximating $g(\nu)$.

We choose $\kappa - 1$ equally spaced grids between $(-\sqrt{ab}, \sqrt{ab})$, $-\sqrt{ab} = \nu_0 < \nu_1 < \cdots < \nu_\kappa = \sqrt{ab}$, which divide the domain of ν into κ intervals of width $2\sqrt{ab}/\kappa$. The piecewise uniform is as follows:

$$
g_a(\nu) = \begin{cases} g(\frac{\nu_0+\nu_1}{2}) & \text{if } \nu_0 < \nu \leq \nu_1 \\ g(\frac{\nu_1+\nu_2}{2}) & \text{if } \nu_1 < \nu \leq \nu_2 \\ \cdots & \\ g(\frac{\nu_{\kappa-1}+\nu_\kappa}{2}) & \text{if } \nu_{\kappa-1} < \nu \leq \nu_\kappa \end{cases}
$$

The independent MH proposal for generating $\nu = S_{jj'} - B_{12}$ is given by

$$q(\nu|\cdot) \propto (1-\rho_{jj'})g(\nu)I\{\nu = -B_{12}, \nu^2 < ab\} + \frac{\rho_{jj'}\lambda}{2} g_a(\nu).$$

Samples of ν could be generated using an inverse-CDF method from the proposal distribution and the proposal ν^* is accepted with the probability

$$\alpha = \min\left\{1, \frac{p(\nu^*)}{q(\nu^*)} / \frac{p(\nu^c)}{q(\nu^c)}\right\},$$

where $\nu^c = S^c_{jj'} - B_{12}$ denotes the current state of ν. The piecewise uniform proposal distribution avoids the local-trap problem and can be sampled easily using an inverse-CDF method. Furthermore, $q(\nu|\cdot)$ approximates the distribution $p(\nu|\cdot)$ more accurately with the increases of the number of grids. Based on our simulations, 100 grids are enough for a fast convergence of $S_{jj'}$.

C.2 Posterior sampling schemes for graphical factor model

We derive the full conditionals for all the parameters and perform the posterior inference using a Gibbs sampling algorithm. Note that the prior specification for the parameters $\{M, z_k, \tau_k\}$ in the graphical factor analytic models parallels the hierarchical model in equations 4.9, so the full conditionals of these parameters for a Gibbs algorithm are the same as in Section C.1. In this section, we just present the sampling algorithm of the parameter set $\{S, G, \xi\}$ for a decomposable graph of the residuals, and $\{C, \lambda^C, \rho^C\}$ for an unrestricted graph of the residuals, in sequence.

Sampling $\{S, G, \xi\}$ for decomposable graphical models

- Sampling of S: The full conditional distribution of S is

$$S|G, y, M, f \sim \text{HIW}(G, \delta + n, \Phi + \Lambda),$$

where $\Lambda = (y - Mf)(y - Mf)^T$. Hence we can generate posterior samples of S directly from the HIW distribution conditional on other parameters.

- Sampling of G: The conditional distribution of G is

$$G|y, M, f, \xi \propto \frac{h(G, \delta, \Phi)}{h(G, \delta + n - 1, \Phi + \Lambda)} \cdot \exp\{-|G|^{\xi}\}.$$

The term $h(G, \delta, \Phi)$ is the normalizing constant for the HIW(G, δ, Φ) distribution given by

$$h(G, \delta, \Phi) = \frac{\prod_{k=1}^{K} |\frac{\Phi_{P_k}}{2}|^{(\frac{\delta+|P_k|-1}{2})} \Gamma_{|P_k|}\left(\frac{\delta+|P_k|-1}{2}\right)^{-1}}{\prod_{k=2}^{K} |\frac{\Phi_{Q_k}}{2}|^{(\frac{\delta+|Q_k|-1}{2})} \Gamma_{|Q_k|}\left(\frac{\delta+|Q_k|-1}{2}\right)^{-1}},$$

where $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(x + (1 - j)/2)$ is the multivariate gamma function. Note that the conditional distribution of G is marginalized over S and hence only dependent on M, f and ξ. This marginalized posterior conditional of G leads to a collapsed Gibbs algorithm in sampling G, accelerating the graphical model search task. To sample the graph G from the conditional distribution, we use a random walk MH algorithm within the Gibbs sampling.

Let $\{e_{jj'} : j < j'\}$ be the set of edge indicators where $e_{jj'} = 1$ if $(j, j') \in E$ and $e_{j,j'} = 0$ otherwise. In an iteration with the current state of graph $G^c = (V, E^c)$, we choose a pair (j, j') at random and change the state of the edge, i.e. $e^p_{jj'} = 1 - e^c_{jj'}$. If the proposed state $G^p = (V, E^p)$ is decomposable, the proposal is accepted as a new state with the probability

$$\alpha(G^c, G^p) = \min\left\{1, \frac{p(G^c|y, M^c, f^c, \xi^c)}{p(G^p|y, M^c, f^c, \xi^c)}\right\},$$

where $p(\,\cdot\,)$ denotes the posterior conditional distribution of G. If the proposed state $G^p = (V, E^p)$ is not decomposable, then choose another pair (j, j') until the proposal graph is decomposable.

- Sampling of ξ: The conditional distribution of ξ is

$$\xi | G \propto \frac{\exp\{-|G|^{\xi}\}}{\sum_{G^*}\exp\{-|G^*|^{\xi}\}}\mathcal{I}\{\xi \in (0,5)\}.$$

  We use a random-walk MH algorithm to generate posterior samples of ξ. Given the current state ξ^c, generate a proposal $\log(\xi^p)$ from a normal distribution $\mathcal{N}(\log(\xi^c), \sigma_\xi^2)$, with the standard deviation σ_ξ chosen properly. We generate the MCMC samples of ξ in the log scale to ensure positivity. The proposal ξ^p is then accepted with the probability

$$\alpha(\xi^c, \xi^p) = \min\left\{1, \frac{p(\xi^p|G^c)}{p(\xi^c|G^c)}\right\},$$

  where $p(\,\cdot\,)$ denotes the full conditional distribution of ξ.

Sampling $\{C, \lambda^C, \rho^C\}$ for unrestricted graphical models

For convenience, let $\Lambda = (y - Mf)(y - Mf)^T$.

- Sampling of the diagonal elements of C, C_{jj}, for $j = 1, \dots, q$: The full conditional density of C_{jj} is

$$p(C_{jj}|\,\cdot\,) \propto (detC)^{n/2}\exp\Big(-\frac{1}{2}\Lambda_{jj}C_{jj} - \frac{\lambda^C}{2}C_{jj}\Big)I_{C\in M^+}.$$

  Without loss of generality, suppose that $j = q$. Let $C = R'R$ be the Cholesky decomposition of C where the matrix $R = (R_{jj'})$ is upper triangular. Then

$$p(C_{jj}|\,\cdot\,) \propto (C_{jj} - c)^{n/2}\exp\Big\{-(\frac{\Lambda_{jj}}{2} + \frac{\lambda^C}{2})C_{jj}\Big\}I_{C_{jj}>c},$$

  where $c = \sum_{j=1}^{q-1} R_{j,q}^2$ does not depend on C_{jj}. Let $\nu = (C_{jj} - c)$, then the conditional distribution of ν is

$$p(\nu|\,\cdot\,) \propto \nu^{n/2}\exp\Big\{-(\frac{\Lambda_{jj}}{2} + \frac{\lambda^C}{2})\nu\Big\}I_{\nu>0},$$

which follows Gamma($\frac{n}{2}+1, \frac{\Lambda_{jj}+\lambda^C}{2}$). Hence, we can draw samples of ν from the Gamma distribution first, and obtain $C_{jj} = \nu + c$.

- Sampling of the off-diagonal elements of C, $C_{jj'}$, for $j < j'$: The complete conditional density of $C_{jj'}$ is

$$p(C_{jj'}|\cdot) \propto (detC)^{n/2} \exp(-\Lambda_{jj'}C_{jj'}) I_{C \in M^+} \cdot p(C_{jj'}|\rho^C_{jj'}, \lambda^C).$$

Without loss of generality, suppose that $j = q-1$ and $j' = q$. Then using Lemma 2 of Wong et al.(2003),

$$\begin{aligned} p(C_{jj'}|\cdot) \propto & \Big\{1 - \frac{(C_{jj'}-a)^2}{cb^2}\Big\}^{n/2} \exp(-\Lambda_{jj'}C_{jj'}) I_{|C_{jj'}-a|<b\sqrt{c}} \cdot p(C_{jj'}|\rho^C_{jj'}, \lambda^C), \\ \propto & I_{|C_{jj'}-a|<b\sqrt{c}} \cdot \Big[(1-\rho^C_{jj'})\Big\{1 - \frac{(C_{jj'}-a)^2}{cb^2}\Big\}^{n/2} I_{C_{jj'}=0} \\ & + \rho^C_{jj'}\Big\{1 - \frac{(C_{jj'}-a)^2}{cb^2}\Big\}^{n/2} \exp\{-\Lambda_{jj'}C_{jj'} - \lambda^C|C_{jj'}|\} I_{C_{jj'}\neq 0}, \end{aligned}$$

where $a = \sum_{j=1}^{q-2} R_{j,q-1}R_{j,q}$, $b = R_{q-1,q-1}$, and $c = R^2_{q-1,q} + R^2_{q,q}$ do not depend on $C_{jj'}$. Now transform $C_{jj'}$ to $\nu = (C_{jj'} - a)/(b\sqrt{c})$, and let $\kappa = -a/(b\sqrt{c})$. The full conditional density of ν is

$$p(\nu|\cdot) \propto (1-\rho^C_{jj'})(1-\nu^2)^{n/2} I_{\nu=\kappa, |\nu|<1} + \frac{\rho^C_{jj'}\lambda^C}{2} g(\nu),$$

where $g(\nu) = (1-\nu^2)^{n/2} \exp\{-\Lambda_{jj'}(\nu b\sqrt{c}+a) - \lambda^C|\nu b\sqrt{c}+a|\}$. The continuous part of the conditional distribution of ν, $g(\nu)$, cannot be sampled directly. Since $g(\nu)$ only has density over $(-1, 1)$ and is zero elsewhere, we can use an independent MH algorithm as the sampling algorithm for $S_{jj'}$ in Section C.1. The details of the independent MH algorithm are explained in Section C.1. Briefly, we choose $\kappa - 1$ equally spaced grids between (-1,1), $-1 = \nu_0 < \nu_1 < \cdots < \nu_\kappa = 1$, which divide the domain of ν into κ intervals of width $2/\kappa$, and construct a piecewise uniform distribution $g_a(\nu)$ approximating $g(\nu)$. The independent MH proposal for generating ν is then given by

$$q(\nu|\cdot\,) \propto (1-\rho^C_{jj'})g(\nu)I_{\{}\nu=\kappa, |\nu|<1\} + \frac{\rho^C_{jj'}\lambda^C}{2}g_a(\nu).$$

Samples of ν could be generated using an inverse-CDF method from the proposal distribution, and the proposal ν^* is accepted with the probability

$$\alpha = \min\left\{1, \frac{p(\nu^*)}{q(\nu^*)}/\frac{p(\nu^c)}{q(\nu^c)}\right\},$$

where $\nu^c = (C^c_{jj'}-a)/(b\sqrt{c})$ denotes the current state of ν. Samples of $C_{jj'}$ are obtained as $C_{jj'} = \nu b\sqrt{c}+a$.