# BAYESIAN GAUSSIAN GRAPHICAL MODELS USING SPARSE SELECTION PRIORS AND THEIR MIXTURES

A Dissertation

by

RAJESH TALLURI

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

August 2011

Major Subject: Statistics

# BAYESIAN GAUSSIAN GRAPHICAL MODELS USING SPARSE SELECTION PRIORS AND THEIR MIXTURES

A Dissertation

by

RAJESH TALLURI

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 |
| | Veerabhadran Baladandayuthapani |
| Committee Members, | Jeffrey D. Hart |
| | Aniruddha Datta |
| Head of Department, | Simon J. Sheather |

August 2011

Major Subject: Statistics

# ABSTRACT

Bayesian Gaussian Graphical Models Using Sparse Selection Priors and Their Mixtures.
(August 2011)

Rajesh Talluri, B.Tech., Indian Institute of Technology, Guwahati
Co-Chairs of Advisory Committee: Dr. Bani K. Mallick
Dr. Veerabhadran Baladandayuthapani

We propose Bayesian methods for estimating the precision matrix in Gaussian graphical models. The methods lead to sparse and adaptively shrunk estimators of the precision matrix, and thus conduct model selection and estimation simultaneously. Our methods are based on selection and shrinkage priors leading to parsimonious parameterization of the precision (inverse covariance) matrix, which is essential in several applications in learning relationships among the variables. In Chapter I, we employ the Laplace prior on the off-diagonal element of the precision matrix, which is similar to the lasso model in a regression context. This type of prior encourages sparsity while providing shrinkage estimates. Secondly we introduce a novel type of selection prior that develops a sparse structure of the precision matrix by making most of the elements exactly zero, ensuring positive-definiteness.

In Chapter II we extend the above methods to perform classification. Reverse-phase protein array (RPPA) analysis is a powerful, relatively new platform that allows for high-throughput, quantitative analysis of protein networks. One of the challenges that currently limits the potential of this technology is the lack of methods that allows for accurate data modeling and identification of related networks and samples. Such models may improve the accuracy of biological sample classification based on patterns of protein network activation, and provide insight into the distinct biological relationships underlying different cancers. We propose a Bayesian sparse graphical modeling

approach motivated by RPPA data using selection priors on the conditional relationships in the presence of class information. We apply our methodology to an RPPA data set generated from panels of human breast cancer and ovarian cancer cell lines. We demonstrate that the model is able to distinguish the different cancer cell types more accurately than several existing models and to identify differential regulation of components of a critical signaling network (the PI3K-AKT pathway) between these cancers. This approach represents a powerful new tool that can be used to improve our understanding of protein networks in cancer.

In Chapter III we extend these methods to mixtures of Gaussian graphical models for clustered data, with each mixture component being assumed Gaussian with an adaptive covariance structure. We model the data using Dirichlet processes and finite mixture models and discuss appropriate posterior simulation schemes to implement posterior inference in the proposed models, including the evaluation of normalizing constants that are functions of parameters of interest which are a result of the restrictions on the correlation matrix. We evaluate the operating characteristics of our method via simulations, as well as discuss examples based on several real data sets.

To my parents

# ACKNOWLEDGMENTS

It was a treatise to work with Dr. Bani Mallick as he has an uncanny ability to envision cutting-edge problems. As a result, I got an opportunity to work on problems of practical significance with applications in diverse areas. His wisdom on conducting research and research management has surely benefited my research. I consider myself extremely fortunate to have worked under his supervision, and his pleasant personality always made the work environment a home away from home.

My collaboration with Dr. Veera started with a summer internship. During our time together, his style of functioning, his way of translating ideas into algorithms, his planning and scheduling of work and his views of using statistics from application point of view have greatly influenced my work. I thank him for being on my dissertation committee.

Any graduate student in the department agrees in no uncertain terms, that Dr. Mike Longnecker is a role model for any aspiring teacher. He corrected me whenever I was off track and made me realize how serious graduate education is in the US. I thank my lab mate Soma, for inspiring some new ideas and making my grad life filled with fun.

TABLE OF CONTENTS

CHAPTER Page

# LIST OF TABLES

# LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

# CHAPTER I

## INTRODUCTION TO BAYESIAN ADAPTIVE GAUSSIAN GRAPHICAL MODELS

### A. Introduction

Consider the $p$ dimensional random vector $\boldsymbol{Y} = (Y^{(1)}, \cdots, Y^{(p)})$, which follows a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where both the mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ are unknown. Flexible modelling of the covariance matrix, $\boldsymbol{\Sigma}$, or equivalently the precision matrix, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, is one of the most important tasks in analysing Gaussian multivariate data. Furthermore, it has a direct relationship to constructing Gaussian graphical models (GGMs) by identifying the significant edges. Of particular interest in this structure is the identification of zero entries in the precision matrix $\boldsymbol{\Omega}$. An off-diagonal zero entry $\Omega_{ij} = 0$ indicates conditional independence between the two random variables $Y^{(i)}$ and $Y^{(j)}$, given all other variables. This is the covariance selection problem or the model selection problem in the Gaussian graphical models (Dempster, 1972; Speed and Kiiveri, 1986; Wong et al., 2003; Yuan and Lin, 2007), which provides a framework for the exploration of multivariate dependence patterns.

GGMs are tools for modelling conditional independence relationships. Among the practical advantages of using GGMs in high-dimensional problems is their ability to (i) make computations more efficient by alleviating the need to handle large matrices, (ii) yield better predictions by fitting sparser models, and (iii) aid scientific understanding by breaking down a global model into a collection of local models that are easier to search. Estimating the precision matrix efficiently and understanding its graphical structure is challenging, however, due to a variety of reasons that we discuss hereafter.

A GGM for a random vector $\boldsymbol{Y}$ can be represented by an undirected graph $G =$

---

This dissertation follows the style of *Journal of the American Statistical Association.*

$(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ contains $p$ vertices corresponding to the $p$ variates and the edges $\boldsymbol{E} = (e_{ij})_{(1\leq i<j\leq p)}$ describe the conditional independence relationships among $Y^{(1)}, \ldots, Y^{(p)}$. The edge between $Y^{(i)}$ and $Y^{(j)}$ is absent if and only if $Y^{(i)}$ and $Y^{(j)}$ are independent, conditional on the other variables, which corresponds to $\Omega_{ij} = 0$. Thus, parameter estimation and model selection in the Gaussian graphical model are equivalent to estimating parameters and identifying zeros in the precision matrix. The two main difficulties are that the number of unknown elements in the covariance matrix increases quadratically with $p$, and that it is difficult to deal directly with individual elements of the covariance matrix because it is necessary to keep the estimated matrix positive definite. Yang and Berger (1994) and Dempster (1969) pointed out that estimators based on scalar multiples of the sample covariance matrix tend to distort the eigenstructure of the true covariance matrix unless $p/n$ is small. In this paper, we address these modelling and inferential challenges as we explore methods to adaptively estimate the precision matrix in a Gaussian graphical model setting.

There have been many approaches to Gaussian graphical modelling. In a Bayesian setting, modelling is based on hierarchical specifications for the covariance matrix (or precision matrix) using global priors on the space of positive-definite matrices, such as an inverse Wishart prior or its equivalents. Dawid and Lauritzen (1993) introduced an equivalent form as the hyper-inverse Wishart distribution. Although that construction enjoys many advantages, such as computational efficiency due to its conjugate formulation and exact calculation of marginal likelihoods (Scott and Carvalho, 2008), it is sometimes inflexible due to its restrictive form. Unrestricted graphical model determination is challenging unless the search space is restricted to decomposable graphs, where the marginal likelihoods are available up to the overall normalizing constants (Giudici, 1996; Roverato, 2000). The marginal likelihoods are used to calculate the posterior probability of each graph, which gives an exact solution for small datasets, but a prohibitively large number of graphs for a moderately large $p$. Moreover, extension to a nondecomposable graph is nontrivial and computation-

ally expensive using reversible-jump algorithms (Giudici and Green, 1999; Brooks et al., 2003). There have been several attempts to shrink the covariance/precision matrix via matrix factorizations for unrestricted search over the space of both decomposable and non-decomposable graphs. Barnard et al. (2000) factorized the covariance matrix in terms of standard deviations and correlations, proposed several shrinkage estimators and discussed suitable priors. Wong et al. (2003) expressed the inverse covariance matrix as a product of the inverse partial variances and the matrix of partial correlations, then used reversible-jump-based Markov chain Monte Carlo (MCMC) algorithms to identify the zeros among the diagonal elements. Liechty et al. (2004) proposed flexible modelling schemes using decompositions of the correlation matrix.

Alternate approaches for more adaptive estimation and/or selection of the graphical models are based on priors/penalties that enforce sparsity. In a regression context for variable selection problems such priors have been proposed by George and McCulloch (1993, 1997); Kuo and Mallick (1998); Dellaportas et al. (2000, 2002). However the context of covariance selection in graphical models is inherently a different problem with additional complexity arising due to the additional constraints of positive definiteness and the number of parameters to estimate being on the the order of $p^2$ instead of $p$. An alternate class of penalties that have received considerable attention in recent times have been lasso-type penalties (Tibshirani (1996)) that have the ability to promote sparseness, and have been used for variable selection in regression problems. In a frequentist graphical model context, Meinshausen and Búhlmann (2006), Yuan and Lin (2007) and Friedman et al. (2008) proposed methods to estimate the precision or covariance matrix based on lasso-type penalties that yield only point estimates of the precision matrix. Lasso-based penalties are equivalent to Laplace priors in a Bayesian setting (Figueiredo, 2003; Bae and Mallick, 2004; Park and Casella, 2008). However, in a Bayesian setting, lasso penalties do not produce absolute zeros as the estimates of the precision matrix, and thus cannot be used to conduct model

selection simultaneously in such settings.

In this paper, we propose novel Bayesian methods for GGMs that allow for simultaneous model selection and parameter estimation. We introduce a novel type of prior in Subsection C that can be decomposed into selection and shrinkage components in which lasso-type priors are used to accomplish shrinkage and variable selection priors are used for selection. We allow for local exploration of graphical dependencies that leads to a sparse structure of the precision matrix by enforcing most of the non-required elements to be exactly zero with positive probability while ensuring the estimate of the precision matrix is positive definite. More importantly, as a significant methodological innovation, we extend these methods to mixtures of GGMs for clustered data, with each mixture component assumed to be Gaussian with an adaptive covariance structure. For some kinds of data, it is reasonable to assume that the variables can be clustered or grouped based on sharing similar connectivity or graphs. Our motivation for this model arises from a high-throughput gene expression data set, for which it is of interest not only to cluster the patients (samples) into the correct subtype of cancer but also to learn about the underlying characteristics of the cancer subtypes. Of interest is differentiating the structure of the gene networks in the cancer subtypes as a means of identifying biologically significant differences that explain the variations between the subtypes. The modelling and inferential challenges are related to determining the number of components, as well as estimating the underlying graph for each component. We present a hierarchical extension of our adaptive methods for such settings, which, to the best of our knowledge, has not been addressed previously in the literature.

In this chapter, we propose novel Bayesian methods using shrinkage and selection priors for Gaussian graphical models that allow model selection and parameter estimation simultaneously. In Subsection B, we employ the Laplace prior on the off-diagonal element of the precision matrix, which is similar to the lasso model in a regression context. This type of prior encourages sparsity while providing shrinkage estimates. We introduce a novel

type of selection prior in Subsection C which will develop a sparse structure of the precision matrix by making most of the elements exactly zero, ensuring the estimate of the precision matrix is positive-definite. In Subsection D we describe about a naive Bayesian model for precision selection. In Subsection E we perform simulations to assess the operating characteristics of our methods and apply the model to real datasets.

## B. The Bayesian Lasso Model for Sparse Graphical Models

Let $\boldsymbol{Y}_{p\times n} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a $p \times n$ matrix with $n$ independent samples and $p$ variates, where each sample $\boldsymbol{Y}_i = (Y_i^{(1)}, \ldots, Y_i^{(p)})$ is a $p$ dimensional vector corresponding to the $p$ variates. We assume $\boldsymbol{Y}$ follows a matrix normal distribution $\boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2\boldsymbol{I}_n)$ with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}$ between the $p$ variates $(Y^{(1)}, \ldots, Y^{(p)})$ and $\sigma^2$ works as a scaling factor for the covariance matrix which without loss of generality can be assumed to be equal to one. Given a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ , we wish to estimate the precision/concentration matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The maximum likelihood estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{A}})$ where $\bar{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T$. The commonly used sample covariance matrix is $\hat{\boldsymbol{S}} = n\bar{\boldsymbol{A}}/(n-1)$. The concentration matrix $\boldsymbol{\Omega}$ can be estimated by $\bar{\boldsymbol{A}}^{-1}$ or $\hat{\boldsymbol{S}}^{-1}$. However, if the dimension is $p$, we need to estimate $p(p+1)/2$ numbers of unknown parameters, which even for a moderate size $p$, might lead to unstable estimates of $\boldsymbol{\Omega}$. In addition, given our main aim is to explore the conditional relationships among the variables, our main interest is the identification of zero entries in the concentration matrix, because a zero entry $\Omega_{ij} = 0$ indicates the conditional independence between the two covariates $Y^{(i)}$ and $Y^{(j)}$ given all other covariates. We propose different kinds of priors over $\boldsymbol{\Omega}$ to explore these zero entries. Here and throughout the paper we follow the notation, $\theta_1|\theta_2$ to represent the conditional distribution of the random variable $\theta_1$ given $\theta_2$. The likelihood of

the Gaussian graphical model is written as

$$\begin{aligned} \boldsymbol{Y}|G &\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Omega^{-1}}, \sigma^2 \boldsymbol{I}_n) \\ &= (2\pi\sigma^2)^{-\frac{np}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} exp\{-\frac{1}{2\sigma^2} tr\{\boldsymbol{\Omega Y Y^T}\}\}. \end{aligned}$$

Modeling the entire $p \times p$ covariance matrix is more complicated, so it is helpful to start by breaking it down into components. In our modeling framework, we directly work with standard deviations and a correlation matrix (Barnard et al. (2000)), which do not correspond to any type of parameterization (e.g. Cholesky, etc). This separation has a strong practical motivation as most practitioners are trained to think in terms of standard deviations and correlations. In this procedure, we would like to use partial correlations and the inverse of partial standard deviations to model the precision matrix instead of modeling the covariance matrix (Wong et al. (2003)).

To this end, we can parameterize the precision matrix as $\boldsymbol{\Omega} = \boldsymbol{S} \times \boldsymbol{C} \times \boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix and $\boldsymbol{C}$ is a correlation matrix. The partial correlation coefficients are related to $C_{ij}$ as

$$\rho_{ij} = \frac{-\Omega_{ij}}{(\Omega_{ii}\Omega_{jj})^{\frac{1}{2}}} = -C_{ij}.$$

To develop the Bayesian lasso (Blasso) model, we assign a Laplace prior on $C_{ij}$, $i < j$. We need an additional constraint that $\boldsymbol{C} \in \mathbb{C}_p$, where $\mathbb{C}_p$ is the space of all correlation matrices of dimension $p$, leading to the prior for $C_{ij}$ as,

$$C_{ij} \sim \text{Laplace}(0, \tau_{ij}) I(\boldsymbol{C} \in \mathbb{C}_p), i < j$$

where the indicator function $I(\bullet)$ ensures that the correlation matrix is positive-definite and introduces dependence among the $C_{ij}$'s.

Laplace priors have the ability to promote sparseness and have been used for variable selection in regression problems (Figueiredo (2003); Yuan and Lin (2005); Park and Casella

(2008)) and especially in high-dimensional settings (Bae and Mallick (2004)). It is well-known that the MAP estimates using the Laplace prior are the same as those produced by applying the lasso algorithm that minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. We induce sparsity in our model by using this Laplace prior where the prior on $\tau_{ij}$ tunes the level of sparsity. To complete the hierarchical formulation, we choose inverse gamma ($IG$) priors for the inverse of the partial standard deviations $S_i$ , Laplace shrinkage parameter $\tau_{ij}$ and $\sigma^2$.

The hierarchical model can be summarized as follows:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\Omega}, \sigma^2 &\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}, \sigma^2 \boldsymbol{I}_n) \\
\boldsymbol{\Omega} &= \boldsymbol{SCS} \\
C_{ij} &\sim \text{Laplace}(0, \tau_{ij}) I(\boldsymbol{C} \in \mathbb{C}_p),\ i < j \\
\tau_{ij} &\sim IG(e, f),\ i < j \\
S_i &\sim IG(g, h) \\
\sigma^2 &\sim IG(k, l)
\end{aligned}$$

for $i = 1, \ldots, p$, $j = 1, \ldots, p$.

1. Posterior inference and conditionals for the Bayesian Lasso Model

In this model, as the posterior is not of explicit form, we perform the posterior inference using MCMC methods. We derive the full conditionals for all the parameters, and as they are not of closed form, we employ the Metropolis-Hastings (MH) algorithm to draw those parameters.

The joint distribution of all parameters $\boldsymbol{C}, \boldsymbol{\tau}, \boldsymbol{S}, \sigma^2 | \boldsymbol{Y} \propto$

$$(2\pi\sigma^2)^{-\frac{np}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} exp\{-\frac{1}{2\sigma^2} tr\{\boldsymbol{\Omega Y Y}^T\}\} \times \prod_{i<j} K(\tau_{ij}) \frac{1}{2\tau_{ij}} exp(-\frac{|C_{ij}|}{\tau_{ij}}) I(\boldsymbol{C} \in \mathbb{C}_p)$$

$$\times \prod_{i<j} \tau_{ij}^{-e-1} exp(-\frac{f}{\tau_{ij}}) \times \prod_{i=1}^{p} S_i^{-g-1} exp(\frac{-h}{S_i}) \times (\sigma^2)^{-k-1} exp(\frac{-l}{\sigma^2}).$$

The unnormalized joint posterior can be computed using the above expression. For each MCMC run we can compute the unnormalized joint posterior by evaluating the expression by substituting the values of the parameters at that particular MCMC iteration. Here $\boldsymbol{\Omega} = \boldsymbol{SCS}$ and $K(\tau_{ij})$ is the normalizing constant for $\tau_{ij}$, which has a complicated expression due to the truncated range of $\boldsymbol{C}$ and constraint of positive definiteness. If $\mathbb{C}_p$ is the space of all correlation matrices of dimension $p$, then $I(\boldsymbol{C} \in \mathbb{C}_p)$ ensures that $\boldsymbol{C}$ is a correlation matrix which is an additional constraint on the lasso solution. Subsequently, we derive the conditional distribution of all the parameters to pursue our MCMC algorithm.

**Sampling of** $C_{ij}$:

The full conditional for $C_{ij}$ is

$$C_{ij} | \boldsymbol{C}_{-ij}, \sigma^2, \tau_{ij} \propto |\boldsymbol{\Omega}|^{n/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega Y Y}^T\} - \frac{1}{\tau_{ij}} |C_{ij}|\} I(\boldsymbol{C} \in \mathbb{C}_p).$$

where $\boldsymbol{C}_{-ij}$ contains all other off diagonal elements of $\boldsymbol{C}$ except the $ij^{th}$ one. While drawing each $C_{ij}$, we have to ensure the positive definiteness of the matrix $\boldsymbol{C}$. We choose to use the approach proposed by Barnard et al. (2000). We compute the range from which $C_{ij}$ should be sampled so that $\boldsymbol{C}$ is positive-definite. Details of this procedure are given in the Appendix. The range can be found out from the roots of a simple quadratic equation as outlined in Barnard et al. (2000). These roots depend only on $\boldsymbol{C}_{-ij}$. Hence after using this approach, the constraint of positive definiteness is equivalent to $I_{[u_{ij}, v_{ij}]}(C_{ij})$ where $u_{ij}$, $v_{ij}$

are functions of $\boldsymbol{C}_{-ij}$. Accordingly, the full conditional distribution is

$$C_{ij}|\boldsymbol{C}_{-ij}, \sigma^2, \tau_{ij} \propto |\boldsymbol{\Omega}|^{n/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega Y Y}^T\} - \frac{1}{\tau_{ij}}|C_{ij}|\} I_{[u_{ij}, v_{ij}]}(C_{ij}) I_{[-1,1]}(C_{ij}),$$

As this distribution is not in a closed form, we can employ the MH algorithm to sample from this distribution. However, $C_{ij}$ lies within an interval, so rather than using the MH algorithm, we discretize this interval in grids and then evaluate the conditional distribution at these grid values. The next step is to normalize the grid values and make a discrete draw of $C_{ij}$ from the grid values using those normalized values as the corresponding probabilities. This is similar to performing discrete bootstrap sampling from the conditional distribution. Furthermore, we used this discrete grid based method with resolution .001.

**Sampling** $\tau_{ij}$:

The full conditional distribution for $\tau_{ij}$ is

$$\tau_{ij}|C_{ij}, \boldsymbol{C}_{-ij} \propto K(\tau_{ij}) \frac{1}{\tau_{ij}} exp(\frac{-|C_{ij}|}{\tau_{ij}}) \times \tau_{ij}^{-g-1} exp(-\frac{h}{\tau_{ij}}) I(\boldsymbol{C} \in \mathbb{C}_p),$$

where $K$ is the normalizing constant constrained by the truncation and positive definiteness constraint on $\boldsymbol{C}$. First, based on $C_{-ij}$ we can identify the largest possible interval of $C_{ij}$, say $u_{ij}$ and $v_{ij}$, which will keep $\boldsymbol{C}$ positive-definite. Then, we evaluate $K(\tau_{ij})$ as

$$\begin{aligned} K^{-1}(\tau_{ij}) &= \int_{-1}^{1} \frac{1}{2\tau_{ij}} exp\{\frac{-|C_{ij}|}{\tau_{ij}}\} I_{[u_{ij}, v_{ij}]}(C_{ij}) dC_{ij} \\ &= \frac{1}{2}[sgn(v_{ij})\{1 - exp\{\frac{-|v_{ij}|}{\tau_{ij}}\}\} - sgn(u_{ij})\{1 - exp\{\frac{-|u_{ij}|}{\tau_{ij}}\}\}], \end{aligned}$$

where $sgn$ is the sign function

$$sgn(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

We draw $\tau_{ij}$'s from this distribution using the MH algorithm.

**Sampling** $\sigma^2$:

The full conditional distribution of $\sigma^2$ is in a closed form so we directly draw from the inverse gamma distribution as

$$k^* = k + np/2, \quad l^* = l + \frac{1}{2}tr(\boldsymbol{\Omega}\boldsymbol{Y}\boldsymbol{Y}^T)$$

$$\sigma^2|\boldsymbol{\Omega}, \boldsymbol{Y} \sim IG(k^*, l^*).$$

**Sampling S**:

The full conditional distribution of $S_i$ is

$$\begin{aligned} S_i|\boldsymbol{S}_{-i}, \boldsymbol{Y}, \sigma^2 &\propto |\boldsymbol{SCS}|^{n/2} exp\{-\frac{1}{2\sigma^2}tr\{\boldsymbol{SCSYY}^T\}\}S_i^{-g-1}exp(\frac{-h}{S_i}) \\ &\propto S_i^n exp\{-\frac{1}{2\sigma^2}tr\{\boldsymbol{SCSYY}^T\}\}S_i^{-g-1}exp(\frac{-h}{S_i}). \end{aligned}$$

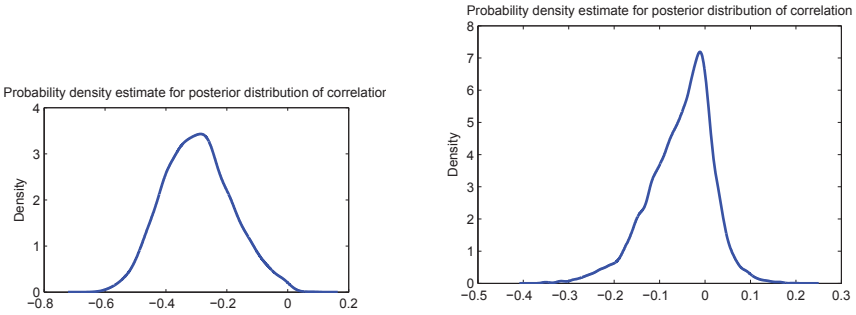
We use MH algorithm to sample $S_i$ from this distribution.

The conditionals for the model which are not in closed form are limited to an interval. So we can use griding to calculate the exact distribution and draw from it directly. We use a Metropolis Hastings step for drawing $S_i$ and $\tau_{ij}$, which converges quickly with a vague prior. All other conditionals are directly drawn from their distributions.

## 2. Posterior thresholding for sparse solutions in Bayesian Lasso Models

The Bayesian lasso model yields (adaptively) shrunk estimates of the precision matrix, whose entries are close to zero but not exactly zero i.e. the Laplace prior induces sparsity by shrinking the off-diagonal elements $C_{ij}$ close to zero depending on the shrinkage parameter $\tau_{ij}$, but they will not be exactly zero. . To explore the zero entries in the precision matrix, we introduce a thresholding rule based on the variability of the estimates. We show this for the cork boring dataset example. The posterior kernel density estimates of the MCMC chains

for coefficients that were determined to be nonzero and determined to be exactly zero are as shown in Figures 1(a) and 1(b), respectively. To achieve sparsity, we compute the 95% bootstrap confidence interval for the mode of $C_{ij}$ from the MCMC samples of $C_{ij}$. The mode for each data set of the bootstrap sample is computed by finding the kernel density for the sample and finding the mode of the estimated density. We use the method used in Botev et al. (2010) to automatically select the optimal bandwidth for density estimation. If zero is contained in the interval then the corresponding $C_{ij}$ is zero, and if zero is not




(a) Posterior distribution for nonzero correlation

(b) Posterior distribution for zero correlation

Fig. 1.: Shows the kernel density estimate of the empirical distributions of the MCMC samples of the correlations.

contained in the interval then the corresponding $C_{ij}$ is the estimate of the mode. Generally the empirical distributions of the MCMC samples are unimodal, but in rare cases when they are multi-modal, the mode of the sample set is defined as the highest point in the empirical p.d.f. By using the method described above we get a graphical model that corresponds to

the model averaging of the best models, containing zero entries.

## C. The Bayesian Lasso Selection Model for Sparse Graphical Models

In this section, we develop a selection model to identify the off-diagonal elements of the precision matrix that are exactly zero. We have a likelihood function for this model that is similar to the previous one as,

$$\begin{aligned} \boldsymbol{Y}|G &\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Omega^{-1}}, \sigma^2 \boldsymbol{I}_n) \\ &= (2\pi\sigma^2)^{-\frac{np}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} exp\{-\frac{1}{2\sigma^2} tr\{\boldsymbol{\Omega Y Y^T}\}\}, \end{aligned}$$

where $\boldsymbol{\Omega} = \boldsymbol{SCS}$ is similarly structured as in the Bayesian lasso model, but the correlation matrix $\boldsymbol{C}$ is now modeled as

$$\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{R}$$

where $\odot$ is the Hadamard operator that does the element-wise multiplication.

### 1. Modelling the shrinkage matrix $\boldsymbol{R}$

In order to achieve adaptive shrinkage of the partial correlations, we assign a Laplace prior to the off-diagonal elements of $\boldsymbol{R}$, $R_{ij}$'s for $i < j$, where the Laplace prior is defined as

$$f(R_{ij}|\tau_{ij}) \propto \frac{1}{2\tau_{ij}} exp(-\frac{|R_{ij}|}{\tau_{ij}}),$$

with each individual element having its own scale parameter, $\tau_{ij}$, that controls the level of sparsity. As discussed previously, Laplace priors have been widely used for shrinkage applications.

Since $\boldsymbol{R}$ is a correlation matrix with elements that lie between [-1, 1], we incorporate this fact as an additional constraint on the overall convolution matrix, $\boldsymbol{C} \in \mathbb{C}_p$, where $\mathbb{C}_p$ is the space of all correlation matrices of dimension $p$. Hence the prior for $R_{ij}$ can be written

as,

$$R_{ij}|\boldsymbol{A} \sim \text{Laplace}(0, \tau_{ij})I(\boldsymbol{C} \in \mathbb{C}_p),$$

where the indicator function ensures that the correlation matrix is positive definite. The full specification of the constraints on the $R_{ij}$'s to ensure the positive definiteness are discussed in Appendix A.

In this setting, the shrinkage parameter $\tau_{ij}$ controls the degree of sparsity, i.e., determines how much the $ij^{th}$ element of $\boldsymbol{R}$ will be shrunk towards zero. We assign an exchangeable inverse gamma prior as

$$\tau_{ij} \sim IG(e, f), i < j,$$

where $(e, f)$ are the shape and scale parameters, respectively. Note that if we set $\tau_{ij} = \tau \ \ \forall i, j$ along with $\boldsymbol{A} = \mathbf{1}_n$ (i.e., a matrix of all 1's), this gives rise to the special case of the Bayesian version of the graphical lasso of Friedman et al. (2008) and Yuan and Lin (2007), where the single penalty parameter ($\tau$) controls the sparsity of the graph and is estimated via cross-validation or by using a criterion similar to the Bayesian information criterion (BIC). By allowing the penalty parameter to vary locally for each node, we allow for additional flexibility, which has been shown to result in better properties than those of the lasso prior and which also satisfies the oracle property (consistent model selection), as shown by Griffin and Brown (2007) in the variable selection context. This fact is also illustrated in our data analysis and simulations studies.

## 2. Modelling the selection matrix $\boldsymbol{A}$

Since $\boldsymbol{A}$ is the selection matrix that performs the variable selection on the elements of the correlation matrix $\boldsymbol{R}$, it thus consists of only binary variables with the off-diagonal elements being either zeros or ones. The most general prior is an exchangeable Bernoulli

prior on the off-diagonal elements of $\boldsymbol{A}$, given as

$$A_{ij}|q_{ij} \sim \text{Bernoulli}(q_{ij}), i < j,$$

where $q_{ij}$ is the probability that the $ij^{th}$ element will be selected as 1; and $q_{ij}$ is assigned a beta prior as

$$q_{ij} \sim \text{Beta}(a, b), i < j.$$

In this construction the hyperparameters $q_{ij}$ control the probability that the $ij^{th}$ element will be selected as a non-zero element. To evaluate a highly sparse model the hyper-parameters should be specified such that the beta distribution is skewed towards zero, and for a dense model the hyper-parameters should be specified such that the beta distribution is skewed towards one. Furthermore, prior beliefs about the existence of edges can be incorporated at this stage of the hierarchy by giving greater weights to important edges while down-weighting redundant edges.

In conclusion, the joint specification of $\boldsymbol{A}$ and $\boldsymbol{R}$ above gives us the *graphical lasso selection* that performs simultaneous shrinkage and selection. To complete the hierarchical specification of the graphical lasso selection, we use an inverse gamma prior on the inverse of the partial standard deviations $S_i$:

$$S_i \sim IG(g, h), i = 1, 2, \ldots, p.$$

The complete hierarchical model can be succinctly summarized as

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\Omega}, \sigma^2 &\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}, \sigma^2 \boldsymbol{I}_n) \\
\boldsymbol{\Omega} &= \boldsymbol{S}(\boldsymbol{A} \odot \boldsymbol{R})\boldsymbol{S} \\
A_{ij}|q_{ij} &\sim \text{Bernoulli}(q_{ij}), i < j \\
\boldsymbol{R}|\boldsymbol{A} &\sim \prod_{i<j} \text{Laplace}(0, \tau_{ij}) I(\boldsymbol{C} \in \mathbb{C}_p) \\
\tau_{ij} &\sim IG(e, f), i < j \\
q_{ij} &\sim \text{Beta}(a, b), i < j \\
S_i &\sim IG(g, h) \\
\sigma^2 &\sim IG(k, l),
\end{aligned}$$

where $i = 1, \ldots, p$, $j = 1, \ldots, p$ and $\odot$ is the Hadamard product.

3. Conditional distributions and the posterior sampling for the selection model

We again use MCMC methods for posterior inference as the joint posterior is not of explicit form. All the full conditional distributions of the parameters are not in closed form, so we employ the MH algorithm to draw those parameters. For simplicity, let $\theta_{ij} = \{\boldsymbol{R}_{-ij}, \boldsymbol{A}_{-ij}, q_{ij}, \boldsymbol{Y}\}$ where $\boldsymbol{R_{-ij}}$ and $\boldsymbol{A}_{-ij}$ contain all other off-diagonal elements of $\boldsymbol{R}$ and $\boldsymbol{A}$, respectively, except the $ij^{th}$ one.

**Joint sampling of** $[A_{ij}, R_{ij}]$:

First, we consider the complete conditional distribution of $R_{ij}$ as

$$[R_{ij}|A_{ij}, \theta_{ij}] \propto |\boldsymbol{\Omega}|^{n/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}\boldsymbol{Y}\boldsymbol{Y}^T\} - \frac{1}{\tau_{ij}}|R_{ij}|\} I(\boldsymbol{C} \in \mathbb{C}_p)$$

We use this conditional distribution to draw $R_{ij}$. We use the discrete bootstrap method to draw $R_{ij}$ similarly to drawing $C_{ij}$ in the Bayesian lasso model. To sample $A_{ij}$, we need to

evaluate its complete conditional distribution

$$[A_{ij}|R_{ij},\theta_{ij}] \propto |\boldsymbol{\Omega}|^{n/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}\boldsymbol{Y}\boldsymbol{Y}^T\}\} q_{ij}^{A_{ij}}(1-q_{ij}^{1-A_{ij}}) I(\boldsymbol{C}\in\mathbb{C}_p)$$

and use it to draw the binary variable $A_{ij}$.

An alternative way to sample $A_{ij}$ is to marginalize $R_{ij}$ from the joint distribution of $A_{ij}$ and $R_{ij}$ and use the marginal distribution for sampling $A_{ij}$. As the marginalization is not explicitly available, we use a Riemann approximation of this integral. We take $M$ grid points within the interval $[u_{ij}, v_{ij}]$, which is the range of values $R_{ij}$ can take, and use the approximation

$$P(A_{ij}=0|\theta_{ij}) \propto (1-q_{ij}) \sum_{k=1}^{M} |\boldsymbol{\Omega}_{(R_{ij}(k),A_{ij}=0)}|^{\frac{n}{2}} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}_{(R_{ij}(k),A_{ij}=0)}\boldsymbol{Y}\boldsymbol{Y}^T\}\}$$

$$P(A_{ij}=1|\theta_{ij}) \propto q_{ij} \sum_{k=1}^{M} |\boldsymbol{\Omega}_{(R_{ij}(k),A_{ij}=1)}|^{\frac{n}{2}} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}_{(R_{ij}(k),A_{ij}=1)}\boldsymbol{Y}\boldsymbol{Y}^T\}\}$$

Consequently, we draw $A_{ij}$ as a discrete binary variable using these probabilities as weights.

**Sampling** $\tau_{ij}, q_{ij}$:

The full joint conditional distribution for $\tau_{ij}$ and $q_{ij}$ is

$$\tau_{ij}, q_{ij}|A_{ij}, R_{ij}, \theta_{ij} \propto K(\tau_{ij}, q_{ij}) \frac{1}{\tau_{ij}} exp(\frac{-|A_{ij}R_{ij}|}{\tau_{ij}}) \times \tau_{ij}^{-g-1} exp(-\frac{h}{\tau_{ij}})$$
$$\times q_{ij}^{A_{ij}}(1-q_{ij})^{(1-A_{ij})} I(\boldsymbol{C}\in\mathbb{C}_p),$$

where $K$ is the normalizing constant constrained by the truncation and positive definiteness constraint on $\boldsymbol{C}(=\boldsymbol{A}\odot\boldsymbol{R})$. First, based on $\boldsymbol{R}_{-ij}$ we can identify the largest possible interval of $R_{ij}$, say $u_{ij}$ and $v_{ij}$ (Barnard et al. (2000)), which will keep $\boldsymbol{C}$ positive-definite.

Then, we evaluate $K(\tau_{ij}, q_{ij})$ :

$$K^{-1}(\tau_{ij}, q_{ij}) = \sum_{A_{ij}=\{0,1\}} q_{ij}^{A_{ij}}(1-q_{ij})^{(1-A_{ij})} \int_{-1}^{1} \frac{1}{2\tau_{ij}} exp\{\frac{-|A_{ij}R_{ij}|}{\tau_{ij}}\} I_{[u_{ij},v_{ij}]}(A_{ij}R_{ij}) dR_{ij}$$
$$= \frac{(1-q_{ij})}{2} \frac{(v_{ij}-u_{ij})}{\tau_{ij}} I_{[u_{ij},v_{ij}]}(0) I_{A_{ij}}(0) + \frac{q_{ij}}{2} CLap(u_{ij}, v_{ij}) I_{[u_{ij},v_{ij}]}(R_{ij}) I_{A_{ij}}(1)$$

where $CLap(u_{ij}, v_{ij}) = [sgn(v_{ij})\{1 - exp\{\frac{-|v_{ij}|}{\tau_{ij}}\}\} - sgn(u_{ij})\{1 - exp\{\frac{-|u_{ij}|}{\tau_{ij}}\}\}]$and $sgn$ is the sign function

$$sgn(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Now we can draw $\tau_{ij}$ and $q_{ij}$ from their conditional distributions :

$$\tau_{ij}|q_{ij}, A_{ij}, R_{ij}, \boldsymbol{Y} \propto K(\tau_{ij}, q_{ij}) \frac{1}{\tau_{ij}} exp(\frac{-|A_{ij}R_{ij}|}{\tau_{ij}}) \times \tau_{ij}^{-g-1} exp(-\frac{h}{\tau_{ij}})$$
$$q_{ij}|\tau_{ij}, A_{ij}, R_{ij}, \boldsymbol{Y} \propto K(\tau_{ij}, q_{ij}) q_{ij}^{a_{ij}}(1-q_{ij})^{(1-a_{ij})} q_{ij}^{\alpha-1}(1-q_{ij})^{(\beta-1)}.$$

Both of these conditionals do not have an explicit form, so we need to use the Metropolis Hastings algorithm to draw $\tau_{ij}$ and $q_{ij}$ from their conditionals.

**Sampling** $\sigma^2$:

The full conditional distribution of $\sigma^2$ is in a closed form so we directly draw from the inverse gamma distribution as

$$k^* = k + np/2, \quad l^* = l + \frac{1}{2} tr(\boldsymbol{\Omega Y Y}^T)$$

$$\sigma^2|\boldsymbol{\Omega}, \boldsymbol{Y} \sim IG(k^*, l^*).$$

**Sampling** $S_i$ The full conditional distribution of $S_i$ is

$$S_i|\boldsymbol{S}_{-i}, \boldsymbol{Y}, \sigma^2 \propto |\boldsymbol{S}(\boldsymbol{A} \odot \boldsymbol{R})\boldsymbol{S}|^{n/2} exp\{-\frac{1}{2\sigma^2} tr\{\boldsymbol{S}(\boldsymbol{A} \odot \boldsymbol{R})\boldsymbol{S}\boldsymbol{Y}\boldsymbol{Y}^T\}\} S_i^{-g-1} exp(\frac{-h}{S_i})$$
$$\propto S_i^n exp\{-\frac{1}{2\sigma^2} tr\{\boldsymbol{S}(\boldsymbol{A} \odot \boldsymbol{R})\boldsymbol{S}\boldsymbol{Y}\boldsymbol{Y}^T\}\} S_i^{-g-1} exp(\frac{-h}{S_i}).$$

We use the Metropolis Hastings algorithm to sample $S_i$ from this distribution.

## 4. Model selection using marginal probabilities

In this subsection, we propose a metric using marginal probabilities to compare different graphs visited by the MCMC chains. The marginal posterior probability of a given graphical ($G$) structure can be expressed as,

$$p(G|Y) \propto \int p(Y|\theta, G)p(\theta|G)p(G)d\theta, \tag{1.1}$$

where $Y$ denotes the data and $G$ encodes the variables that define the graphical structure and $\theta$ represents all the other parameters in the model. In standard graphical models $p(\theta|G)$ is usually assigned a conjugate prior such as hyper Inverse-Wishart (Jones et al. (2004); Carvalho et al. (2007)) and hence the integral in (1.1) can be obtained explicitly. Although, making computations tractable, the conjugate priors restricts the search to to small classes of graphical models like decomposable graphical models (Giudici and Green (1999); Scott and Carvalho (2008)). In our framework, we explore a larger class of graphical models in addition to inducing sparsity which comes with an added computational complexity – the marginal density (1.1) is not available in explicit form.

However, one method to approximate the marginal posterior probability using our MCMC samples is as below.

1. We rank the top graphs based on some model selection criteria. For our examples we choose Bayes Information Criteria(BIC) which penalizes the complex models in

favor of balanced models and is defined as,

$$-2\log p(Y|G) + const \approx -2L(Y,\hat{\theta}) + m_{\mathcal{M}} log(n) \equiv BIC$$

where $p(Y|G)$ is the (integrated) likelihood of the data for the graph $G$, $L(Y,\hat{\theta})$ is the maximized mixture log likelihood for the model, and $m_G$ is the number of independent parameters to be estimated in the model. The number of parameters to be estimated in the model is considered as the number of nonzero edges and all the other parameters in the model.

2. Select top $K$ (say 200) graphs in accordance with the BIC values.

3. Re-run the MCMC (for M iterations) to get sufficient samples to approximate the marginal probabilities using the Harmonic mean estimate (Newton and Raftery (1994); Gelfand and Dey (1994)).

4. Use the Harmonic mean estimate $P(G|Y) \approx (M^{-1}\sum_{i=1}^{M} p(Y|\theta_i)^{-1})^{-1}$ and normalize it to calculate the posterior probabilities of the models.

The resulting marginal posterior probabilities now come with appropriate uncertainty bounds and can be used for inference.

This approach has a major drawback which is the volatility of the harmonica mean estimators. This has been criticized widely in literature and we chose to use an alternative method to approximate posterior probabilities based on the frequency of appearance of models in the MCMC. We obtain the Monte-Carlo estimates of these posterior probabilities by counting the proportion of MCMC samples to have the specific graphical structure. Hence, if $I(\boldsymbol{A} = \boldsymbol{A}^*)$ denote the indicator function for the graphical model $\boldsymbol{A} = \boldsymbol{A}^*$ , then

the ergodic average or the Monte Carlo frequency estimator of this model $\boldsymbol{A}^*$ is given by

$$\pi(\boldsymbol{A}^*|\boldsymbol{Y}) = \frac{1}{K}\sum_{b=1}^{K} I(\boldsymbol{A}_b = \boldsymbol{A}^*),$$

where $\boldsymbol{A_b}$ is graphical model visited on the $b^{th}$ MCMC draw and $K$ is the total number of draws from the Markov chain.

## D. A Naive Bayesian Model

We also develop a naive Bayesian model expressing $\boldsymbol{C}$ as

$$\boldsymbol{C} = \boldsymbol{A} \odot \hat{\boldsymbol{R}},$$

where $\odot$ is the Hadamard operator that does element wise multiplication. Here $\hat{\boldsymbol{R}}$ is a plug-in estimate of the correlation matrix obtained from factorizing the estimate of the precision matrix $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{S}}\hat{\boldsymbol{R}}\hat{\boldsymbol{S}}$ where $\hat{\boldsymbol{R}}$ is a correlation matrix and $\hat{\boldsymbol{S}}$ is a diagonal matrix. The relation of $\hat{\boldsymbol{R}}$ to partial correlation is described in Subsection C. For a relatively large sample size, the inverse of the sample correlation matrix is an obvious choice for this estimate. $\boldsymbol{A}$ is the shrinkage matrix such that the elements of $\boldsymbol{A}$ will shrink the elements in $\hat{\boldsymbol{R}}$. In this way some of the elements of $\hat{\boldsymbol{R}}$ will be shrunk towards $0$. This approach is similar in spirit to the nonnegative garrote estimator proposed by Breiman (1995) and Yuan and Lin (2007).

We assign a Laplace prior on the off-diagonal elements of $\boldsymbol{A}$

$$A_{ij} \sim \text{Laplace}(0, \tau), i < j.$$

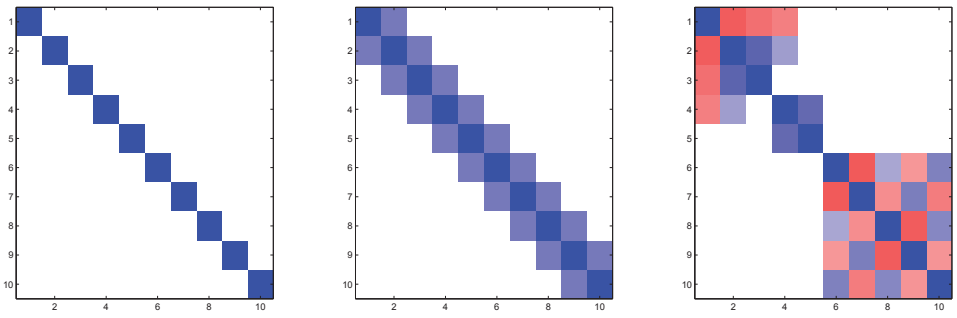
The posterior inference is similar to previous analyses, hence we skip the details.

## E. Simulations

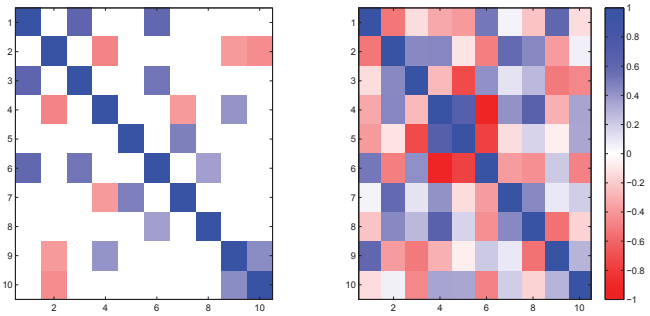
In this subection we compare different methods to assess the performance of the Bayesian lasso models. We simulate five types of concentration matrices, in order of increasing structural complexity:

1. Identity matrix
2. Banded diagonal matrix.
3. Block diagonal matrix
4. Sparse unstructured matrix.
5. Dense unstructured matrix.

An identity matrix is a simple matrix with ones in its diagonal and zeros in its off diagonal. Banded diagonal matrix is a tridiagonal matrix with ones in its diagonal and all the elements in the diagonals adjacent to the main diagonal set to 0.5. Before explaining simulations of more complex matrix structures, we describe the process used for generating a random positive definite correlation matrix. A random lower triangular matrix $\boldsymbol{L}$ was generated with ones in its diagonal and normal random numbers in its lower triangle. Then $\boldsymbol{L}\boldsymbol{L}^T$ gave us a positive definite matrix. The matrix was then factored as $\boldsymbol{Q}\boldsymbol{\Omega}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is a diagonal matrix and $\boldsymbol{\Omega}$ is a correlation matrix with ones in its diagonal which is the desired positive definite correlation matrix. A block diagonal matrix was generated as follows. Two positive definite matrix correlation matrices of sizes $p-k$ and $k$ were generated, where k is a random number between $1$ and $p$, and were concatenated in the diagonals to create a matrix of size $p \times p$ as shown in Figure 2(c). the sparse unstructured matrix was simulated as follows: Let $\boldsymbol{\Sigma} = \boldsymbol{B} + \delta \boldsymbol{I}_p$ where each off-diagonal entry in $\boldsymbol{B}$ is generated independently and equals a random number between $[-1, -.5]$ and $[.5, 1]$ with probability $\pi$ or 0 with probability



(a) Identity Matrix (b) Banded Diagonal Matrix (c) Block Diagonal Matrix



(d) Sparse Matrix (e) Dense Matrix

Fig. 2.: This figure shows the simulated matrices for different types of structures for precision matrix. The colorbar is same for all the matrices. White indicates a zero in the precision matrix whereas colored cells indicate non-zero elements.

$1 - \pi$, all diagonal entries of $\boldsymbol{B}$ are zero and $\delta$ is chosen such that the resulting matrix is positive definite. In the end we do the factorization of $\boldsymbol{\Sigma}$ as $\boldsymbol{Q\Omega Q}$, where $\boldsymbol{Q}$ is a diagonal matrix and $\boldsymbol{\Omega}$ is a correlation matrix with ones in its diagonal, which is the desired sparse positive definite correlation matrix. We can vary the sparsity of the matrices generated by changing the value of $\pi$. We chose $\pi = 0.1$ for the sparse unstructured matrix. The dense unstructured matrix is the full matrix that is a random positive definite correlation matrix of size $p$ generates using the method described above. The simulated matrices for size $p = 10$

are shown in Figure 2. In Figure 2 the white blocks in the off diagonal are the zeros in the matrices, the colors correspond to the magnitude of nonzero off-diagonal elements in the matrices as represented by the colorbar at the end of the figure.

We compare our methods with the "glasso" approach of Friedman et al. (2008) and the method ("MB") proposed by Meinshausen and Búhlmann (2006) as both these methods use the L1- regularization and are closest to our approach using Laplace priors. We try to assess the performance of these methods in terms of the Kullback-Leibler loss (KL), the number of false positives (FP; incorrectly identified edges) and the number of false negatives (FN; incorrectly missed edges). Both the methods were implemented using the glasso package in R. We implemented them using Matlab-R link to call the the functions in Matlab.

It should be noted that both these methods are frequentist methods and they give a point estimate for the precision matrix, whereas the Bayesian methods can also provide the uncertainty estimates for the covariance matrix, so we are comparing the performance regarding the final estimate of the precision matrix. For the Bayesian lasso model and the Bayesian lasso selection model we use the estimate of the precision matrix as the matrix that has the highest joint log posterior of all of the unique models visited in the MCMC simulation. The joint log posterior is computed at every iteration of the MCMC simulation, and the sample with the highest joint log posterior is the most likely map estimate, which can be compared with the estimates of the above two frequentist methods.

The Kullback -Leibler Loss is defined as $\Delta_{KL}(\hat{\mathbf{\Omega}}, \mathbf{\Omega}) = trace(\mathbf{\Omega}\hat{\mathbf{\Omega}}^{-1}) - log|\mathbf{\Omega}\hat{\mathbf{\Omega}}^{-1}| - p$ , whose ideal value should be zero when $\hat{\mathbf{\Omega}} = \mathbf{\Omega}$. Figures 3, 4 and 5 show the means and standard errors for the KL, FP and FN for sample size $n = \{25\}$ and number of covariates $p = \{5, 10, 15, 25\}$ averaged over 10 data sets.

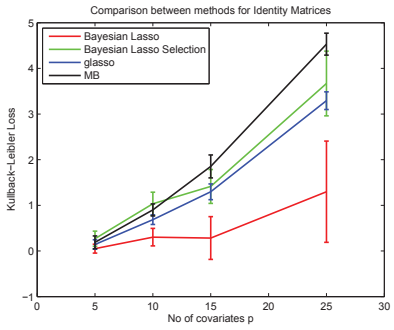
The "glasso" method and the "MB" method were performed using $\rho = 0.1$ which is the tuning parameter for the lasso penalty in both the methods because this setting gave good results for all the scenarios. As shown in Figure 3, the proposed Bayesian methods

perform better than the other methods in some of the cases while in others all the methods are competitive with each other. The Bayesian Lasso model does better than the Bayesian lasso selection model in simpler correlation structures as the Bayesian lasso is a shrinkage model from which the zeros were selected post-MCMC. As it is a continuous model it has a better probability to get to good estimate of the precision matrix in simpler models such as the identity matrix structures, where as the Bayesian lasso selection model is more of a model searching method which searches over all the models of the precision matrix to find which are the probable models. The Bayesian lasso has a higher probability of getting stuck in a local mode than the Bayesian lasso selection model. As the Bayesian lasso selection model makes discrete jumps in the model space, it is more likely to explore the whole space.

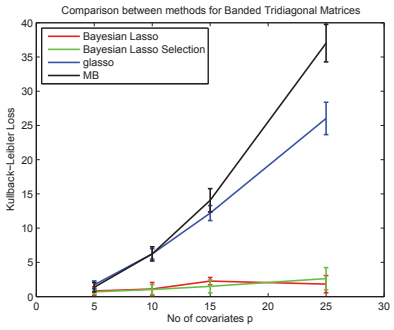
We can see that all the methods perform more or less the same in Identity and Dense Matrix structures. In sparse unstructured matrices and banded diagonal matrices the Bayesian models outperform the "glasso" and "MB" methods. This is because of the adaptive regularization on the partial correlations in Bayesian models. If "glasso" and "MB" did adaptive regularization the methods would have been competitive with each other in these scenarios.

To compute the false positive and false negative rate for the Bayesian lasso model we need to use the bootstrap confidence intervals to find the zeros in the model. This is not necessary for the Bayesian lasso selection model as the zeros are directly incorporated in the model. We also computed the false negative and false positive rates for the methods and compared them in Figures 5 and 4 respectively. This is mostly dependent on the parameter for tuning the sparsity. If you want more sparser models you are more likely to get false negatives and less likely to get false positives. All the methods have similar false negative rates except for dense and block diagonal matrices. Both these scenarios are dense matrices so there are a lot of elements in the matrix which have small partial correlations but not exactly zero, so all the models are likely to make them zero as they are small enough. So there is a higher chance of getting a false negative in these scenarios than others. For the scenario of Identity matrices there is no chance of getting a false negative as all elements are zeros.

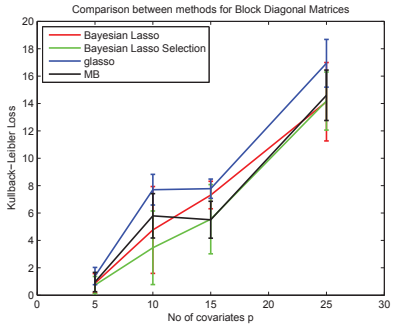
The false positive rates tell us how likely you are to make an error by changing an element which was actually zero to a nonzero one. We can see that the Bayesian models have smaller false positive rates compare to the “glasso” and “MB” methods.



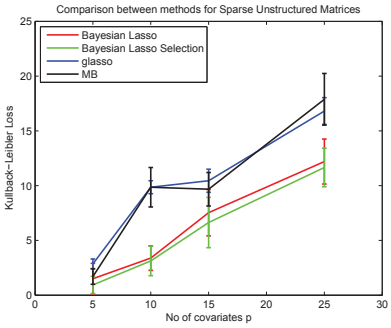

(a) Identity Matrix



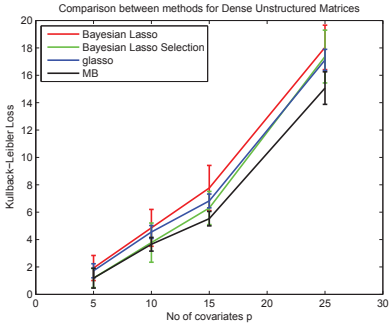

(b) Banded Diagonal Matrix




(c) Block Diagonal Matrix

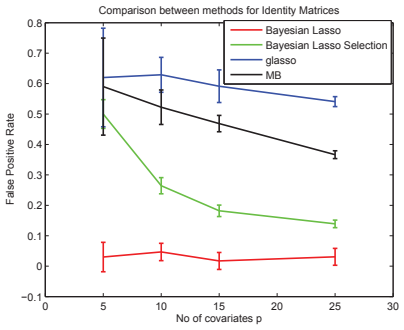



(d) Sparse Matrix

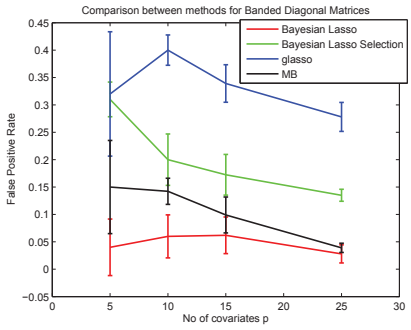



(e) Dense Matrix

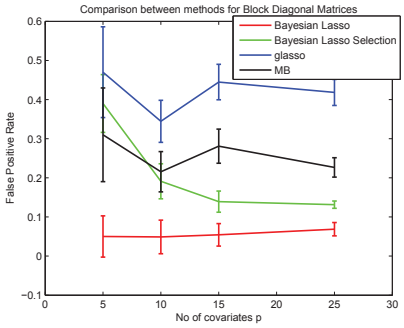
Fig. 3.: This figure shows the comparison between 4 methods "glasso" -Friedman et al. (2008), "MB"- Meinshausen and Búhlmann (2006), "Bayesian lasso" model and "Bayesian lasso selection" model in terms of Kullback-Leibler loss (K-L) for the simulated simulated matrices for different types of structures for precision matrix for $p = 25$. Lower is better.



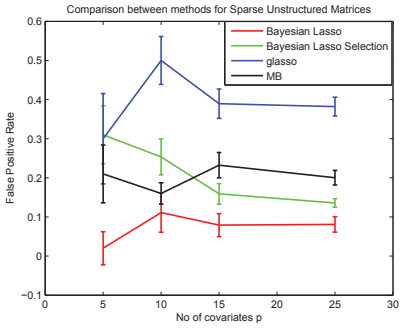

(a) Identity Matrix



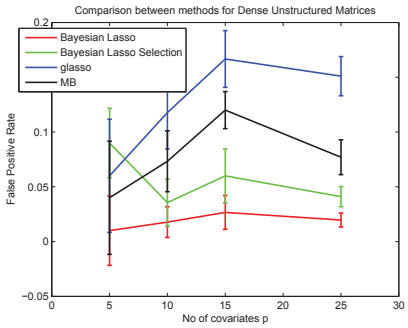

(b) Banded Diagonal Matrix




(c) Block Diagonal Matrix

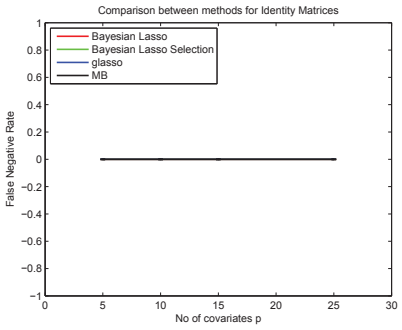



(d) Sparse Matrix

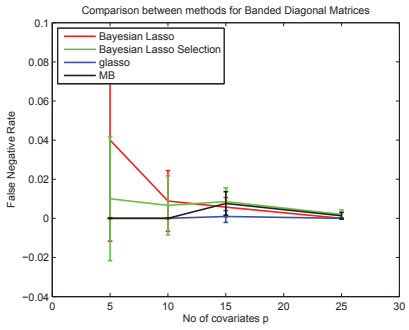



(e) Dense Matrix

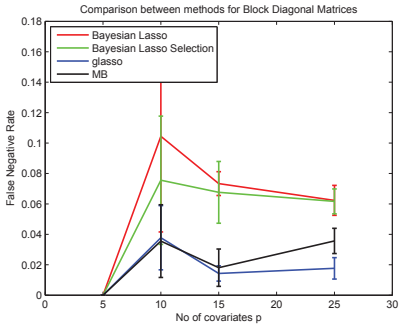
Fig. 4.: This figure shows the comparison between 4 methods "glasso" -Friedman et al. (2008), "MB"- Meinshausen and Búhlmann (2006), "Bayesian lasso" model and "Bayesian lasso selection" model in terms of false positive rates for the simulated simulated matrices for different types of structures for precision matrix for $p = 25$. Lower is better.



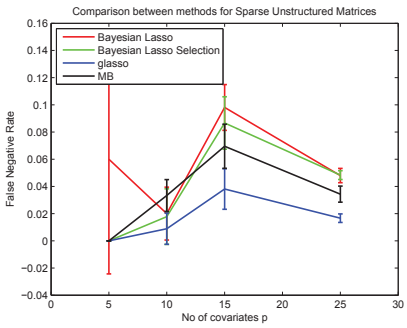

(a) Identity Matrix



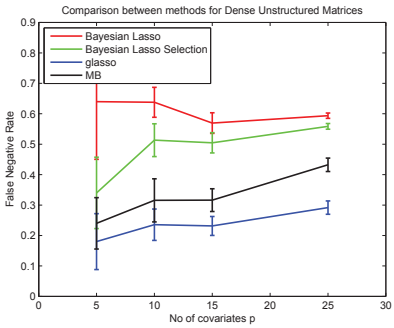

(b) Banded Diagonal Matrix




(c) Block Diagonal Matrix




(d) Sparse Matrix




(e) Dense Matrix

Fig. 5.: This figure shows the comparison between 4 methods "glasso" -Friedman et al. (2008), "MB"- Meinshausen and Búhlmann (2006), "Bayesian lasso" model and "Bayesian lasso selection" model in terms of false negative rates for the simulated simulated matrices for different types of structures for precision matrix for $p = 25$. Lower is better.

## F. Model Comparison with Benchmark Data

We chose to compare our methods with three existing methods that were earlier used in different papers. The Lasso and non negative type garrotte estimator are used in Yuan and Lin (2007) and Mixed Interaction Modeling (MIM) is one of the leading softwares for graphical modeling. For determining the best models for Bayesian lasso model we use the model obtained with the bootstrap confidence intervals. For the Bayesian lasso selection model we compute the joint log posterior for all of the unique models visited in the MCMC simulation and we select the model with the highest joint log posterior as the best model.

**Lasso Model**: The Lasso model is a penalized-likelihood method that does model selection and parameter estimation simultaneously in the Gaussian concentration graph model and uses an $L-1$ penalty on the off-diagonal elements of the concentration matrix that encourages encourages sparsity and simultaneously shrinks the estimates.

**Non-Negative Garrote Model**: This model is similar to the Lasso model but the fact that we have a relatively reliable estimate of the concentration matrix changes the penalty function by incorporating the estimate into it (Yuan and Lin (2007)). This approach is similar to the non-negative garrote estimator proposed by Breiman (1995) for linear regression.

**MIM**: MIM is the only available software supporting graphical modeling with both discrete and continuous variables. MIM is designed for graphical modeling using undirected graphs, directed acyclic graphs and chain graphs. It is based on a comprehensive class of statistical models for discrete and continuous data. The dependence properties of the models can be displayed in the form of a graph. The backward stepwise selection method in Edward‘s MIM package with the option of unrestricted selection, wherein both decomposable and non-decomposable models are considered, is used. Implementation of the stepwise model selection procedure in MIM is based on removing only one edge, the

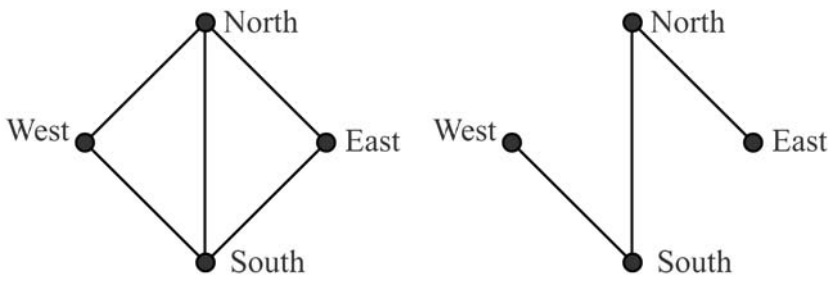
least significant one, at a time.

## 1. Examples

We consider two benchmark real datasets and a stock market dataset to compare our methods

### a. Example 1: Cork borings data set

Cork borings data are presented in Whittaker (1990)(Exercise 8.6.5) and were originally used by Rao (1948). The $p = 4$ measurements are the weights of cork borings on $n = 28$ trees in four directions: north, east, south and west.




(a) (b)

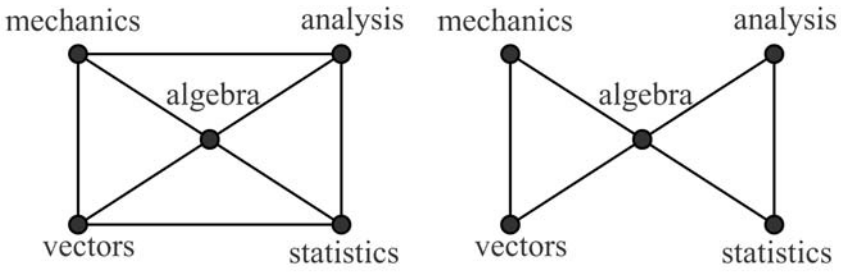
Fig. 6.: (A) was selected by Lasso, Garrote and Naive Bayes Models and (B) was selected by Bayesian lasso, Bayesian lasso selection and MIM Models.

Figure 6 depicts the best graphs for the cork borings data set. We can see that the Bayesian lasso, Bayesian lasso selection and MIM models select the same graph, Figure 6(b) as the best graph. This graph had the highest joint posterior value for both the Bayesian lasso and Bayesian lasso selection models. Whereas the graph in Figure 6(a) is selected as the best graph by Lasso, Garrote and Naive Bayes models. As these are benchmark datasets

with small number of covariates, the results for all the models are very similar because the models that best describe the data are the same. We confirm that we get the same models that best describe the data as in Yuan and Lin (2007).

b. Example 2: The mathematics marks data set

The Mathematics marks dataset (Mardia et al. (1979)) contains the marks of $n = 88$ students in the $p = 5$ examinations in mechanics, vectors, algebra, analysis and statistics,




(a) (b)

Fig. 7.: (A) was selected by the Lasso model and (B) was selected by Bayesian lasso, Bayesian lasso selection, MIM, garrote and Naive Bayes Models.

Figure 7 depicts the best graphs for the mathematics marks data set. Here Bayesian lasso, Bayesian lasso selection, Garrote, Naive Bayes and MIM models select the same graph, Figure 7(b) as the best graph. This graph had the highest joint posterior value for both the Bayesian lasso and Bayesian lasso selection models. The graph in Figure 7(a) is selected as the best model by the Lasso model. As these are benchmark datasets with small number of covariates, the results for all the models are very similar because the models that best describe the data are the same. We confirm that we get the same models that best describe the data as in Yuan and Lin (2007).

c. Example 3: Enron stock market data example

We take a motivating example from a stock market data set Liechty et al. (2004), which may be used by the finance community to group and analyze companies according to their areas of operation. This grouping requires knowledge of the companies and is determined by people who are experts in the field. Grouping companies according to the services or products they offer may be complicated by companies redirecting their efforts, e.g., in response to changing economic situations or consumer demands.

Enron was a company that provided a good illustration of this type of change. Enron began as an energy company, but changed its business focus and transformed itself into a finance company. It was not known whether Enron provided more service to energy clients or to finance clients; therefore, the category into which Enron fit was uncertain. One approach to resolving this uncertainty is to examine the behavior of a companys stock to determine its primary service. We undertook such an analysis using the same data set that was used by Liechty et al. (2004), which consists of data on nine companies. Four of the companies were known to provide energy services, four were known to provide financial services, and the ninth was Enron. The energy companies were Reliant, Chevron, British Petroleum and Exxon. The finance companies were Citi-Bank, Lehman Brothers, Merrill Lynch and Bank of America. The data included monthly stock data for each company over a period of 73 months. This example is also motivated by the need for accurate estimates of pairwise correlations of assets in dynamic portfolio-selection problems. Graphical models offer a potent tool for regularization and stabilization of these estimates, leading to portfolios with the potential to uniformly dominate their traditional counterparts in terms of risk, transaction costs, and overall profitability.

We report the best graphs supported by the data by computing the posterior probabilities for the graphs using the following scheme. The MCMC samples obtained from

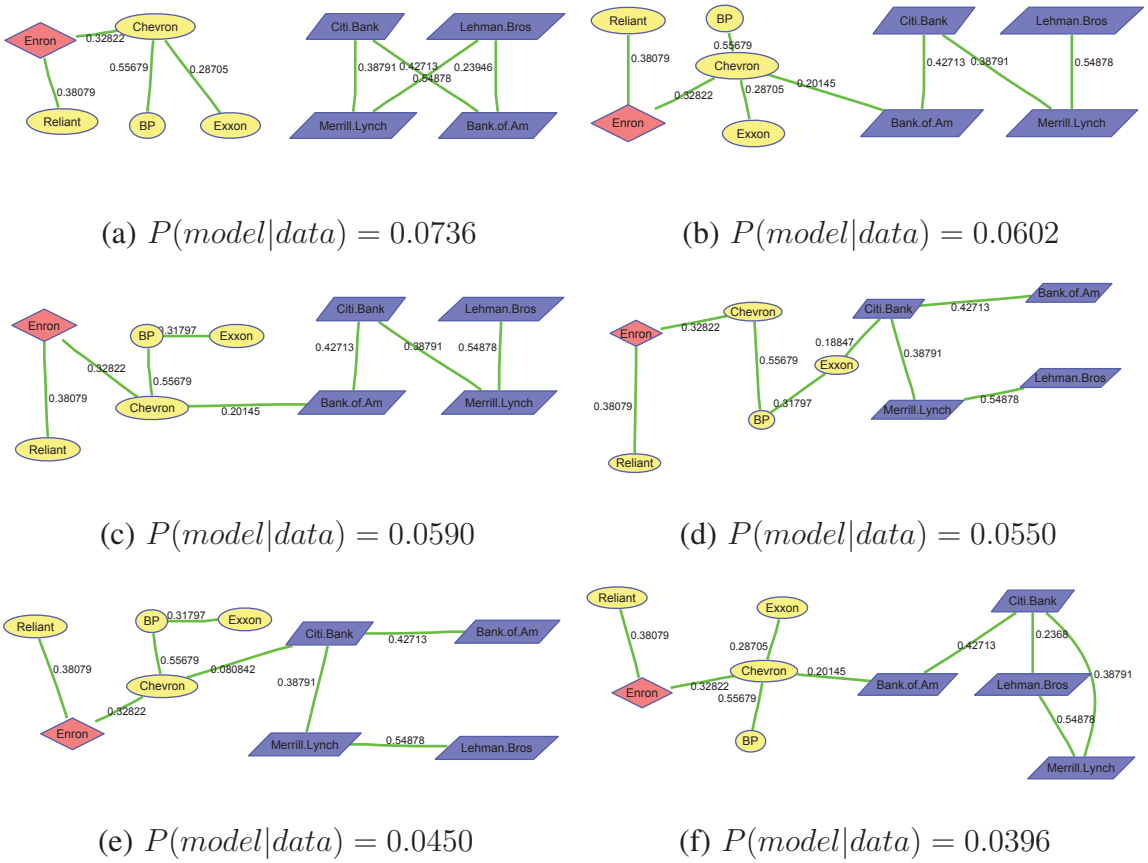
the analysis explore the distribution of possible graphical configurations suggested by the data, with each configuration represented by the selection matrix $\boldsymbol{A}$ encoding the indicators of the possible edges. To explore the space of valid graphs, we follow the strategy of selecting the model with the highest marginal posterior probability over the space of all possible graphs. We obtain the Monte-Carlo estimates of these posterior probabilities by counting the proportion of MCMC samples to have the specific graphical structure. Hence, if $I(\boldsymbol{A} = \boldsymbol{A}^*)$ denote the indicator function for the graphical model $\boldsymbol{A} = \boldsymbol{A}^*$ , then the ergodic average or the Monte Carlo frequency estimator of this model $\boldsymbol{A}^*$ is given by

$$\pi(\boldsymbol{A}^*|\boldsymbol{Y}) = \frac{1}{K}\sum_{b=1}^{K} I(\boldsymbol{A}_b = \boldsymbol{A}^*),$$

where $\boldsymbol{A_b}$ is graphical model visited on the $b^{th}$ MCMC draw and $K$ is the total number of draws from the Markov chain.

The top six graphs identified using our lasso selection model are shown in Figure 8 sorted by the posterior probabilities. It is clear from the illustrated network (e.g Figure 8(a)) that Enron is grouped with the energy companies and was not successful, in terms of stock performance, in transitioning from an energy company to a finance company. Liechty et al. (2004) also found Enron to be more closely related to the energy companies than the finance companies.

For comparison with our proposed method, we selected two methods that use L1-regularization and are similar to our approach using Laplace priors: the "glasso" approach of Friedman et al. (2008) and the method ("MB") proposed by Meinshausen and Búhlmann (2006). As both approaches are frequentist, hence they incorporate no notion of marginal likelihoods and posterior probabilities, we used prediction performance to compare the methods. We split the 73-month data sample into a 60-month training set and a 13-month prediction set. Using the training set to find the top 10 graphs (where top graphs are ranked




(a) $P(model|data) = 0.0736$ (b) $P(model|data) = 0.0602$

(c) $P(model|data) = 0.0590$ (d) $P(model|data) = 0.0550$

(e) $P(model|data) = 0.0450$ (f) $P(model|data) = 0.0396$

Fig. 8.: This figure shows the top 6 graphical models for the stock market data, sorted by the marginal posterior probabilities of the models.

by marginal posterior probabilities), we applied the Bayesian lasso selection model and found the estimates of the precision matrix for each graph. We then predicted the stock value of each sample of the test set given all other stocks for each of the test samples and averaged them over the 10 graphs – thus employing Bayesian model mixing. For the glasso and MB methods, we used the estimate for the precision matrix derived by these methods to predict the test samples using $\rho = 0.1$, where $\rho$ is the tuning parameter for the lasso penalty in both methods. For the sake of a fair comparison of the frequentist methods, we also included a Bayesian model with a single penalty parameter, making $\tau_{ij} = \tau$ and

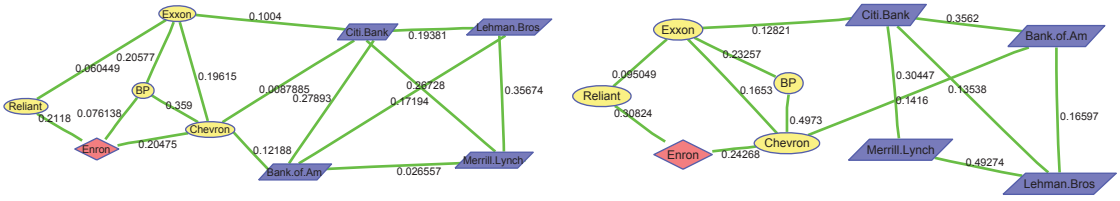
Table I.: Predictive squared error comparison for Enron stock data

| Bayesian lasso selection | Bayesian lasso (single penalty) | glasso | MB |
|---|---|---|---|
| 30.6764 | 31.9765 | 32.1968 | 32.7445 |

$q_{ij} = q$ to make it equivalent to the frequentist models with a single penalty parameter. The results are shown in Table I.

We can see that the Bayesian lasso selection model has the lowest (better) predictive squared error compared to the frequentist methods, thus showing how Bayesian model mixing can help improve prediction accuracy. Of interest is that the performance of the Bayesian lasso model with the single penalty parameter was worse than that of the lasso selection model with a locally varying penalty, and its prediction performance was close to those of the glasso and MB methods. We show the graphs derived from the glasso and MB methods in Figure 9. The inferences are similar using these approaches in the sense that Enron is linked more with oil companies than finance companies. However, these approaches show more connections than are shown in our selection models. Thus the methods seem to differ in imparting sparse solutions, with the Bayesian lasso selection models giving sparser outputs, which is reflected in the prediction performance.

In addition, we compared our graphical method to a simple cluster analysis to see how the companies cluster together in terms of their stock performance. We clustered the data using the model-based clustering software MCLUST (Fraley and Raftery, 2002). We used the "VVV" parameterization to estimate the unconstrained covariance matrix for the




(a) Glasso (b) MB

Fig. 9.: The graphical models for the stock market data obtained using (a) the glasso method and (b) the MB method.

data and used BIC to find the optimal number of clusters. The optimal number of clusters found by BIC was one cluster, which grouped all nine companies together. In contrast, the graph with the highest posterior probability as determined by our method, Figure 8(a), detected two distinct subgraphs, those of energy companies and finance companies, with Enron being connected to the energy companies. This clustering also appeared in the other graphs in Figure 8. In essence, cluster analysis missed this relationship and was unable to distinctly answer the scientific question that was posed.

# CHAPTER II

# BAYESIAN SPARSE GRAPHICAL MODELS FOR CLASSIFICATION WITH APPLICATION TO PROTEIN EXPRESSION DATA

## A. Introduction

### 1. Protein signaling pathways in cancer

The treatment of cancer is rapidly evolving due to an improved understanding of the signaling pathways that are activated in tumors. Global profiling of DNA mutations, chromosomal copy number changes, DNA methylations, and expression of mRNA and miRNA have greatly improved our appreciation of the heterogeneity of cancer [Nishizuka et al. (2003); Blower et al. (2007); Gaur et al. (2007); Shankavaram et al. (2007); Ehrich et al. (2008)]. However, the characterization of protein signaling networks has proven to be much more challenging. Several reasons underscore the critical importance of overcoming this challenge: First, changes in cellular DNA and RNA both ultimately result in changes in protein expression and/or function; thus, protein networks represent the summation of changes that happen at the DNA and RNA levels. Second, research has demonstrated that many of the most common oncogenic genetic changes activate proteins in kinase signaling pathways. Examples include activating mutations of PIK3CA, EGFR, and RAS family members; amplification of HER2/neu; and a loss of the PTEN function.

Numerous studies of protein networks and expression analysis have shown promising results. Due to the hyper-activation of kinase signaling pathways, numerous kinase inhibitors have been used in clinical trials, frequently with dramatic clinical activity. Inhibitors that target protein signaling pathways are now FDA-approved in a variety of cancers, including chronic myelogenous leukemia, breast cancer, colon cancer, renal cell carcinoma, and gastrointestinal stromal tumors [reviewed in Davies et al. (2006)]. While

most of these treatments directly target prevalent genetic changes, or the downstream effectors of the mutated proteins, there is emerging evidence that carcinogenesis frequently involves the concurrent activation of multiple pathways. This is clinically important, as these events may cause resistance to targeted therapies. For example, EGFR inhibitors are FDA-approved for the treatment of metastatic colon cancer. However, research has demonstrated that this treatment is ineffective in colon cancer patients with an RAS mutation in their tumor [Linardou et al. (2008); Siena et al. (2009)]. There is also evidence that concurrent activation of the PI3K-AKT signaling pathway reduces the efficacy of trastuzumab in breast cancer patients who have an amplified level of the HER2/neu gene [Nagata et al. (2004)].

Protein networks need to be assessed directly, as DNA or RNA analyses often do not accurately reflect or predict the activation status of protein networks. Many proteins are regulated by post-translational modifications, such as phosphorylation or cleavage events that are not detected by the analysis of DNA or RNA. Several studies have also demonstrated marked discordance between mRNA and protein expression levels, particularly for genes in kinase signaling and cell cycle regulation pathways [Varambally et al. (2005); Shankavaram et al. (2007)]. Recently, it has been demonstrated, in both cancer cell lines and tumors that different genetic mutations in the same signaling pathway can result in significant differences in the quantitative activation levels of downstream pathway effectors [Stemke-Hale et al. (2008); Davies et al. (2009); Vasudevan et al. (2009); Park et al. (2010)]. While these observations support that direct measurements are essential to measure protein network activation, a number of studies have demonstrated that signaling pathways are frequently regulated by complex feed-forward and feedback regulatory loops, as well as cross-talk between different pathways [Mirzoeva et al. (2009); Zhang et al. (2009); Halaban et al. (2010)]. Thus, developing an accurate understanding of the regulation of protein signaling networks will be optimized by approaches that (1) assess multiple path-

ways simultaneously for different tumor types and/or conditions, and (2) allow for the use of rigorous statistical approaches to identify differential functional networks.

## 2. Graphical models for network analysis

A convenient and coherent statistical representation of protein networks is accorded by graphical models [Lauritzen (1996)]. By "protein network" we mean any graph with proteins as nodes, where the edges between proteins may code for various biological information. For example, an edge between two proteins may represent the fact that their products interact physically (protein-protein interaction network), the presence of an interaction such as a synthetic-lethal or suppressor interaction [Kelley and Ideker (2005)], or the fact that these proteins code for enzymes that catalyze successive chemical reactions in a pathway [Vert and Kanehisa (2003)]. An example plot of the PI3K-AKT signaling pathway, a protein interaction network that is a focus of our study, is shown in Figure 12.

Our focus is on undirected graphical models and on Gaussian graphical models (GGM) in particular [Whittaker (1990)]. These models provide representations of the conditional independence structure of the multivariate distribution – to develop and infer protein networks. In such models, the nodes represent the variables (proteins) and edges represent pairwise dependencies, with the edge set defining the global conditional independence structure of the distribution. We develop an adaptive modeling approach for the covariance structure of high-dimensional distributions with a focus on sparse structures, which are particularly relevant in our setting in which the number of variables ($p$) can exceed the number of observations ($n$).

GGMs have been under intense methodological development over the past few years in both frequentist [Meinshausen and Búhlmann (2006); Chaudhuri et al. (2007); Yuan and Lin (2007); Friedman et al. (2008); Bickel and Levina (2008) ] and Bayesian settings [Giudici and Green (1999); Roverato (2002); Carvalho and Scott (2009)] . In high-dimensional

settings, Dobra et al. (2004) used regression analysis to find directed acyclic graphs and converted them to undirected (sparse) graphs to explore the underlying network structure. However, most of the approaches we cited focused on inferring the conditional independence structure of the graph and did not consider classification, which is one of the foci of our article. Rapaport et al. (2007) used spectral decomposition to detect the underlying network structure and classify genetic data using support vector machines (SVM). More recently Monni and Li (2010) proposed a graph-based regression approach incorporating pathway information as a prior for classification procedures, but their method does not detect differential networks based on available data. In this article, we propose a constructive method for sparse graphical models using selection priors on the conditional relationships in the presence of class information. Our method has several advantages over classical approaches. First, we incorporate (integrate) the uncertainty of the parameters in deriving the optimal rule via Bayesian model mixing. Second, our network model provides an adaptively regularized estimate of the covariance matrix and hence is capable of handling $n < p$ situations. More importantly, our model uses this information in deriving the optimal classification boundary.

With available online databases containing tens of thousands of reactions and interactions, there is a pressing need for methods integrating *a priori* pathway knowledge in the proteomic data analysis models. This challenge has been addressed in several studies. Vert and Kanehisa (2003) developed a method for correlating interaction graphs and different types of quantitative data. For gene expression data, Rahnenfhrer et al. (2004) showed that explicitly taking into account the pathway distance between pairs of genes enhances the statistical scores when identifying activated pathways. Hanisch et al. (2002) proposed co-clustering of gene expression and gene networks, and Galbraith et al. (2006) proposed constructing linear models of gene regulation based on *a priori* known network information. Sivachenko et al. (2002) proposed a method to find significantly affected pathway

regulators when *a priori* network topology information was used jointly with microarray data. In our approach, we integrate prior network information directly in the model in an intuitive way such that the presence of an edge can be specified by providing the probability of an edge to be present in the correlation matrix. Our method is fully Bayesian and allows for posterior inference on the network topologies both within and between groups. After fitting the Bayesian model, we obtain the posterior probabilities of the edge inclusion, which leads to false discovery rate (FDR)-based calls on significant edges.

## B. Model

Our data construct for modeling is as follows. We observe a tuple: $(Z_i, \boldsymbol{Y}_i)$, where $Z_i$ is a categorical outcome denoting the type or subtype of cancer (binary or multicategory) and $\boldsymbol{Y}_i$ is a vector of $p$ proteins for the $i$th sample/patient/array. We proceed by modeling the tuple using the following conditional representation: $P(\boldsymbol{Y}_i|Z_i)P(Z_i)$, where the first term defines a sampling model on the network via a Bayesian GGM. In combination with the second term, this provides the classification scheme.

### 1. Bayesian Sparse Gaussian Graphical Model with selection priors

Let $\boldsymbol{Y}_{p\times n} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a $p \times n$ matrix with $n$ samples and $p$ covariates (proteins). Each sample $\boldsymbol{Y}_i = (Y_i^{(1)}, \ldots, Y_i^{(p)})$ is a $p$ dimensional vector. We associate each vertex of the graph $\boldsymbol{G}$ with a covariate in $\boldsymbol{Y}$ and assume that the graphical model is a family of probability distributions that is Markov in G [Lauritzen (1996)]. $\boldsymbol{Y}$ follows a matrix normal distribution, $\boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2\boldsymbol{I}_n)$, where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is a nonsingular covariance matrix between the covariates, and $\sigma^2$ is a scaling factor for the covariance. For ease of exposition, we set $\boldsymbol{\mu} = 0$ in the ensuing discussion, assuming that the mean effects have been accounted for either via centering or integration.

Given a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ , we wish to estimate the concentration matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, which encodes the conditional dependencies between the proteins. The likelihood of the Gaussian graphical model can then be written as

$$\begin{aligned} \mathbf{Y}|G &\sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1}, \sigma^2 \boldsymbol{I}_n) \\ &\propto (2\pi\sigma^2)^{-\frac{np}{2}} |\boldsymbol{\Omega}|^{n/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}\}\} \end{aligned} \tag{2.1}$$

The key idea behind GGMs is the modeling of the concentration (precision) matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, which will dictate the network structure. In this framework, of particular interest is the identification of zero entries in the concentration matrix $\Omega$. A zero entry at the $ij^{th}$ element of $\Omega$ indicates conditional independence between the two random variables $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$, given all other variables. This is the covariance selection problem in the Gaussian graphical models [Dempster (1972); Cox and Wermuth (2002)]. Typical estimation is carried out either via shrinkage estimation [Yuan and Lin (2007)] or using continuous priors such as hyper-inverse Wishart priors [Carvalho and West (2007)], which yields estimates that are close to zero (but not exactly zero) entries, and thus results in many non-zero entries. We propose a different kind of selection prior over $\boldsymbol{\Omega}$ to explore these zero entries in the next subsection.

a. Parameterization of the concentration matrix

Due to the complicated structure of the covariance matrix, it is helpful to start by breaking it down into components. For some applications (e.g., shrinkage modeling), it is desirable to work directly with standard deviations and a correlation matrix [Barnard et al. (2000)] that do not correspond to any type of parameterization (e.g., Cholesky, etc.). This separation has a strong practical motivation, as most practitioners are trained to think in terms of standard deviations and correlations, thus easing prior elicitation. In this model we would

like to use partial correlations and the inverse of the partial standard deviations to model the concentration matrix instead of modeling the covariance matrix [Wong et al. (2003)].

To this end, we parameterize the concentration matrix as $\boldsymbol{\Omega} = \boldsymbol{S} \times \boldsymbol{C} \times \boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix and $\boldsymbol{C}$ is a correlation matrix. The partial correlation coefficients are related to $C_{ij}$ as

$$\rho_{ij} = \frac{-\Omega_{ij}}{(\Omega_{ii}\Omega_{jj})^{\frac{1}{2}}} = -C_{ij}$$

To aid a more intuitive interpretation for the model, we model the correlation matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{R},$$

where $\odot$ is the Hadamaard operator indicating element-wise multiplication between the two matrices. Here $\boldsymbol{A}$ can be defined as a selection matrix that consists of only binary (0/1) variables as its elements. The off-diagonal elements of $\boldsymbol{A}$ are zeros or ones only. By a selection matrix we mean that the elements in $\boldsymbol{A}$ select which of the elements in $\boldsymbol{R}$ are zeros or not. In other words, $\boldsymbol{A}$ performs variable selection on the elements of the correlation matrix $\boldsymbol{R}$. This parameterization is intuitive in the sense that we work with individual elements of the correlation matrix to determine if each is a zero or not.

We assign a Bernoulli prior on the off-diagonal elements of $A$ as they are binary variables as

$$A_{ij}|q_{ij} \sim \text{Bernoulli}(q_{ij}), i \neq j$$

where $q_{ij}$ is the probability of the $ij$th element being selected as 1.

Since $\boldsymbol{R}$ is a correlation matrix, all of its off-diagonal elements are in the range $[-1, 1]$. Hence, we can assign an independent uniform prior over $[-1, 1]$ for all $R_{ij}$s for $i < j$. Note that all the values of $\boldsymbol{R}$ in this range do not guarantee that $\boldsymbol{C}(= \boldsymbol{A} \odot \boldsymbol{R})$ will be positive-definite. For that we need the additional constraint that $\boldsymbol{C} \in \mathbb{C}_p$ where $\mathbb{C}_p$ is the space of

all correlation matrices of dimension $p$. Hence a coherent prior for $\boldsymbol{R}$ is

$$\boldsymbol{R}|\boldsymbol{A} \sim \prod_{i<j} \text{Uniform}(-1,1) I(\boldsymbol{C} \in \mathbb{C}_p),$$

where $I(\bullet)$, the indicator function, ensures that the correlation matrix is positive-definite and introduces dependence among the $R_{ij}$'s.

Instead of defining a joint prior on the space of the correlation matrices, it is simpler to work with the individual elements $R_{ij}$. Following the method of Barnard et al. (2000), we find the range $[u_{ij}, v_{ij}]$ on the individual elements of $R$ that will guarantee the positive definiteness of $\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{R}$. The resulting prior on the off-diagonal elements $R_{ij}$ can be written as

$$R_{ij}|a_{ij}, A_{-ij}, R_{-ij} \sim Uniform(u_{ij}, v_{ij}) I(-1 < R_{ij} < 1), i \neq j,\ i < j,$$

where $R_{-ij}$ contains all other off-diagonal elements of $\boldsymbol{R}$ except the $ij$th element and $A_{-ij}$ contains all elements of $\boldsymbol{A}$ except the $ij$th element. In the calculations, $u_{ij}$ and $v_{ij}$ have to be chosen such that $\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{R}$ remains positive-definite and (conditionally) $u_{ij}$ and $v_{ij}$ are functions of $R_{-ij}$ and $A_{-ij}$.

The parameter $q_{ij}$ is the probability that the $ij^{th}$ element will be selected as a non-zero element; it controls the degree of sparsity in an adaptive manner by element-wise selection of the entries of the correlation matrix. We assign a beta hyper-prior for the probabilities $q_{ij}$ as

$$q_{ij} \sim \text{Beta}(a_{ij}, b_{ij}), i \neq j,$$

where the hyper-parameters $a_{ij}, b_{ij}$ can be set to induce prior information on the graph structure . To complete the hierarchical specification, we choose an (exchangeable) inverse-gamma prior on the inverse of the partial standard deviations $S$, which is a diagonal matrix containing entries $S_i = \Omega_{ii}^{\frac{1}{2}}$ as $S_i \sim IG(g,h), i = 1, 2, \ldots, p.$, and on the error variance,

$\sigma^2 \sim IG(k, l)$.

All the above parameters described are different for each of the groups. So there will be two $\boldsymbol{A}$'s (i.e.$\boldsymbol{A}^1$ and $\boldsymbol{A}^2$ one for each group) as is the case with all the above parameters. But the main advantage of Bayesian methodology lies in borrowing strength between the groups. This can be accomplished by having a variable which connects the groups. We introduce a latent variable $\lambda$ which is defined as

$$\lambda_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{A}^1_{ij} \neq \boldsymbol{A}^2_{ij} \\ 0 & \text{if } \boldsymbol{A}^1_{ij} = \boldsymbol{A}^2_{ij} \end{cases}$$

The parameter $\lambda_{ij}$ signifies the presence or absence of the same edge in the graphical model of both the groups. In other words $\lambda_{ij} = 1$ signifies a differential edge (i.e. the relation between the covariates $i, j$ is significant in only one group but not the other) whereas $\lambda_{ij} = 0$ signifies a common edge( i.e. the relation between the covariates $i, j$ is significant in both the groups). This information is vital for understanding the biological processes and inferring conclusions from the analysis.

As $\lambda_{ij}$ are binary random variables we propose a Bernoulli prior on $\lambda_{ij}$ as

$$\lambda_{ij} \sim \text{Bernoulli}(\pi_{ij}), i < j$$

The parameter $\pi_{ij}$ is the probability that the relation between $i^{th}$ and $j^{th}$ covariate is differential. We assign a beta hyper-prior for the probabilities $\pi_{ij}$ as

$$\pi_{ij} \sim \text{Beta}(e_{ij}, f_{ij}), i \neq j,$$

The complete hierarchical formulation of the network component of the model can be

succinctly summarized as follows for $k = 1, 2$:

$$\begin{aligned}
\boldsymbol{Y}^{(k)}|\boldsymbol{\Omega}^{(k)}, \sigma^{2(k)} &\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Omega}^{-1(k)}, \sigma^{2(k)}\boldsymbol{I}_n) \\
\boldsymbol{\Omega}^{(k)} &= \boldsymbol{S}^{(k)}(\boldsymbol{A}^{(k)} \odot \boldsymbol{R}^{(k)})\boldsymbol{S}^{(k)} \\
A_{ij}^{(1)}, \lambda_{ij}|q_{ij}^{(1)}, \pi_{ij} &\sim \text{Bernoulli}(q_{ij}^{(1)})\text{Bernoulli}(\pi_{ij}), i < j \\
\boldsymbol{R}^{(k)}|\boldsymbol{A}^{(k)} &\sim \prod_{i<j} \text{Uniform}(-1, 1) I(\boldsymbol{C}^{(k)} \in \mathbb{C}_p) \\
q_{ij}^{(1)} &\sim \text{Beta}(\alpha_{ij}^{(1)}, \beta_{ij}^{(1)}) \\
\pi_{ij} &\sim \text{Beta}(e_{ij}, f_{ij}), i \neq j, \\
S_i^{(k)} &\sim IG(g, h) \\
\sigma^{2(k)} &\sim IG(m, l)
\end{aligned}$$

where $i, j = 1, \ldots, p$.

An important thing to note is that by the introduction of the latent variable $\lambda$ we are actually reparameterizing the model by making one of the $\boldsymbol{A}$ matrices fixed, i.e. given $\boldsymbol{A}^{(1)}$ and $\lambda$, $\boldsymbol{A}^{(2)}$ is fixed. So we only need 2 priors one on $\lambda$ and $\boldsymbol{A}^{(1)}$ as $\boldsymbol{A}^{(2)}$ is no longer a random variable. Because of the same reason we also dont need to draw $q_{ij}^{(2)}$.

b. Incorporating prior pathway information

As we mentioned before, there exists a huge amount of literature (prior knowledge) on pathways and other functional behaviors of proteins such as metabolic, signaling or other regulation pathways. We formally incorporate this *a priori* knowledge in our model through the prior specification on $q_{ij}$, the probability that the edge between protein $(i, j)$ will be selected as shown in Figure 10. In particular, we impose an informative prior on $(q_{ij}) \sim$ Beta$(a_{ij}, b_{ij})$, and set the hyper-parameters $a_{ij}$and $b_{ij}$ such that the distribution has a higher mean to reflect our prior knowledge of the presence of an edge. For example we set the,

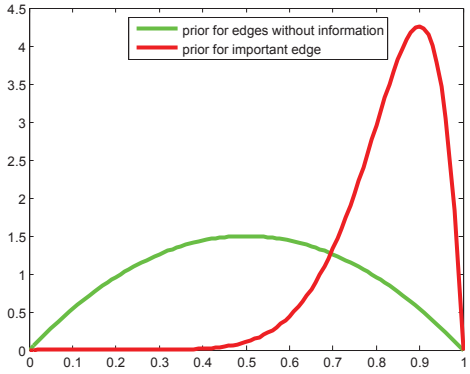



Fig. 10.: This figure shows the how the prior information is incorporated in the model. $q_{ij}$ is the model parameter which is the probability of there being an edge between protein i and protein j . If no information is available , prior on $q_{ij}$ is Beta(2,2) with mean 0.5, reflecting no prior information about the edge and the prior on $q_{ij}$ is Beta(10,2) with mean 0.83, if there is biological evidence that the edge plays an important role in the pathway.

- prior on $q_{ij}$ as Beta(2,2) with mean 0.5, in absence of prior information and
- prior on $q_{ij}$ is Beta(10,2) with mean 0.83, if there is biological evidence that the edge plays an important role in the pathway.

The prior information incorporated in $q_{ij}$ is the pathway information which is when the biological process is normal, whose information is available in on-line databases. The information on which relations between proteins are affected when there is a mutation is not readily available and is one of the goals of our methodology. We can get this information using expert opinion from the biologists who can tell us which relations are perturbed due to mutation and we can incorporate that information to draw $\lambda$. As $\pi_{ij}$ is the probability of a differential edge, we can incorporate the information about perturbed relations during mutations in a similar way as above.

we set the,

- prior on $\pi_{ij}$ as Beta(2,2) with mean 0.5, if the relationship between $i, j$ proteins is not perturbed by a mutation.

- prior on $\pi_{ij}$ is Beta(10,2) with mean 0.83, if there is biological evidence that the relationship between $i, j$ proteins is perturbed by a mutation .

## 2. Conditionals

**Sampling of $R_{ij}^{(k)}$, $k = 1, 2$:**

First, we consider the complete conditional distribution of $R_{ij}^{(k)}$ as

$$[R_{ij}^{(k)}|A_{ij}^{(k)}, R_{-ij}^{(k)}, A_{-ij}^{(k)}, \boldsymbol{Y}^{(k)}] \propto |\boldsymbol{\Omega}^{(k)}|^{n^{(k)}/2} exp\{\frac{-1}{2\sigma^2} tr\{\boldsymbol{\Omega}^{(k)}\boldsymbol{Y}^{(k)}\boldsymbol{Y}^{(k)^T}\}\}$$
$$I_{\{u_{ij}^{(k)}, v_{ij}^{(k)}\}}(A_{ij}^{(k)} R_{ij}^{(k)}).$$

We use this conditional distribution to draw $R_{ij}^{(k)}$. We use the discrete bootstrap method to draw $R_{ij}^{(k)}$

**Joint Sampling of $A_{ij}^{(1)}$ and $\lambda_{ij}$:**

To sample $A_{ij}^{(1)}$ and $\lambda_{ij}$, we need to evaluate its complete conditional distribution which is

$$\begin{aligned}[A_{ij}^{(1)}, \lambda_{ij}|others] \propto\ & |\boldsymbol{\Omega}^{(1)}|^{n^{(1)}/2} exp\{\frac{-1}{2\sigma^{2(1)}} tr\{\boldsymbol{\Omega}^{(1)}\boldsymbol{Y}^{(1)}\boldsymbol{Y}^{(1)^T}\}\} I_{\{u_{ij}^{(1)}, v_{ij}^{(1)}\}}(A_{ij}^{(1)} R_{ij}^{(1)}) \\ & |\boldsymbol{\Omega}^{(2)}|^{n^{(1)}/2} exp\{\frac{-1}{2\sigma^{2(2)}} tr\{\boldsymbol{\Omega}^{(2)}\boldsymbol{Y}^{(2)}\boldsymbol{Y}^{(2)^T}\}\} I_{\{u_{ij}^{(2)}, v_{ij}^{(2)}\}}(A_{ij}^{(2)} R_{ij}^{(2)}) \\ & {q_{ij}^{(1)}}^{A_{ij}^{(1)}} (1 - q_{ij}^{(1)})^{1 - A_{ij}^{(1)}} \pi_{ij}^{\lambda_{ij}} (1 - \pi_{ij})^{(1-\lambda_{ij})} \pi_{ij}^{e_{ij}-1} (1 - \pi_{ij})^{(f_{ij}-1)}\end{aligned}$$

and use it to jointly draw the binary variable $A_{ij}^{(1)}$ and $\lambda_{ij}$. Lets label this equation as $F_{A,\lambda}(.)$.

Note here that there are only 4 cases we need to draw for $A_{ij}^{(1)}$ and $\lambda_{ij}$ ( i.e. $[\{0,0\}, \{0,1\}, \{1,0\}, \{1,1\}]$). So we directly find the probabilities for each of the states .

$P(A_{ij}^{(1)} = 0, \lambda_{ij} = 0|others) \propto F_{A,\lambda}(A_{ij}^{(1)} = 0, \lambda = 0, A_{ij}^{(2)} = 0)$

$P(A_{ij}^{(1)} = 0, \lambda_{ij} = 1|others) \propto F_{A,\lambda}(A_{ij}^{(1)} = 0, \lambda = 1, A_{ij}^{(2)} = 1)$

$P(A_{ij}^{(1)} = 1, \lambda_{ij} = 0|others) \propto F_{A,\lambda}(A_{ij}^{(1)} = 1, \lambda = 0, A_{ij}^{(2)} = 1)$

$P(A_{ij}^{(1)} = 1, \lambda_{ij} = 1|others) \propto F_{A,\lambda}(A_{ij}^{(1)} = 1, \lambda = 1, A_{ij}^{(2)} = 0)$ Consequently, we sample one of the configurations $A_{ij}^{(1)}, \lambda_{ij}$ as a discrete binary variable using these probabilities as weights.

**Complete conditional for $Q^{(1)}$,$\pi_{ij}$:**

$$q_{ij}^{(1)}|A_{ij}^{(1)} \propto {q_{ij}^{(1)}}^{A_{ij}^{(1)}}(1-q_{ij}^{(1)})^{(1-A_{ij}^{(1)})}{q_{ij}^{(1)}}^{\alpha_{ij}-1}(1-q_{ij}^{(1)})^{(\beta_{ij}-1)}$$

$$q_{ij}^{(1)}|A_{ij}^{(1)} \sim \text{Beta}(A_{ij}^{(1)} + \alpha_{ij}, \beta_{ij} + 1 - A_{ij}^{(1)}).$$

Similarly

$$\pi_{ij}|\lambda_{ij} \sim \text{Beta}(\lambda_{ij} + e_{ij}, f_{ij} + 1 - \lambda_{ij}).$$

**Complete conditional for $\sigma^{2(k)}$**

$$m^* = m + n^{(k)}p/2, \quad l^* = l + \frac{1}{2}tr\{\boldsymbol{\Omega}^{(k)}\boldsymbol{Y}^{(k)}\boldsymbol{Y}^{(k)^T}\}$$

$$\sigma^{2(k)}|\Omega^{(k)}, \boldsymbol{Y}^{(k)} \sim IG(m^*, l^*).$$

**Complete conditional for $S^{(k)}$**

$$\begin{aligned} S_i^{(k)}|S_{-i}^{(k)}, \boldsymbol{Y}^{(k)}, \sigma^{2(k)} &\propto |S^{(k)}(A^{(k)} \odot R^{(k)})S^{(k)}|^{n^{(k)}/2} \\ &\quad exp\{\frac{-1}{2\sigma^{2(k)}}tr\{S^{(k)}(A^{(k)} \odot R^{(k)})S^{(k)}\boldsymbol{Y}^{(k)}\boldsymbol{Y}^{(k)^T}\}\} \\ &\quad {S_i^{(k)}}^{-g-1}exp(\frac{-h}{S_i^{(k)}}) \\ &\propto {S_i^{(k)}}^n exp\{\frac{-1}{2\sigma^{2(k)}}tr\{S^{(k)}(A^{(k)} \odot R^{(k)})S^{(k)}\boldsymbol{Y}^{(k)}\boldsymbol{Y}^{(k)^T}\}\} \\ &\quad {S_i^{(k)}}^{-g-1}exp(\frac{-h}{S_i^{(k)}}). \end{aligned}$$

### 3. Bayesian classification based on posterior predictive probabilities

We develop a Bayesian classification procedure based on posterior predictive probabilities using the network information obtained from the previous subsection. Our task here is to build a predictor or classifier for $K$ tumor classes/subtypes that partitions the space into $K$ disjoint subsets, $(B_1, \ldots, B_K)$, such that if a sample with protein expression profile $\boldsymbol{Y}_i = (y_1, \ldots, y_p) \in B_k$, the predicted class is $k$. This utilizes the fact that the proteins in the $k$th class share a common network profile.

We propose a model-based Bayesian classification procedure for this problem. In particular, given a training data set, $\{(Z_i^T, \boldsymbol{Y}_i^T), i = 1, \ldots, N\}$, we wish to build a discrimination rule that we can use to classify future samples based on their protein expression $\boldsymbol{Y}^{new}$, i.e., predict $Z^{new}$ based on the posterior predictive probabilities. Suppose $p_k \equiv P(Z^{new} = k|\boldsymbol{Y}^{new}, \boldsymbol{Y}^T, \boldsymbol{Z}^T)$ is the posterior predictive probability of the new sample belonging to the $k$th class, which is defined as

$$p_k \propto \left[\int_\theta P(\boldsymbol{Y}^{new}|Z^{new} = k, \mathcal{M}) P(\mathcal{M}|\boldsymbol{Y}^T, \boldsymbol{Z}^T) d\mathcal{M}\right] P(Z^{new} = k) \tag{2.2}$$

and is known up to a proportionality constant. Here $\mathcal{M}$ encodes all the unknown model parameters (from the previous subsection), $P(\mathcal{M}|\bullet)$ denotes the posterior distribution of $\mathcal{M}$ based on the current model, and $P(Z^{new} = k)$ is the prior probability of the new samples belonging to the $k$th class.

In our framework, due to the construction of our network model, this posterior predictive density is not available in closed form. We numerically evaluate this integral based on Markov chain Monte Carlo (MCMC) techniques. Suppose $\widehat{\mathcal{M}}_m$ is the $m$th random sample from our MCMC chain, then we can approximate (2.2) as

$$p_k \propto \left[\frac{1}{M}\sum_{m=1}^{M} P(\boldsymbol{Y}^{new}|Z^{new} = k, \widehat{\mathcal{M}}_m)\right] P(Z^{new} = k), \tag{2.3}$$

where $M$ is the total number of MCMC samples, and the approximation (2.3) converges to the true value in (2.2) as $M \longrightarrow \infty$.

A Bayesian classification scheme assigns the new sample to the $k$th class as $k^* = \arg\max_k(p_k)$. This is akin to the usual Bayes discriminant rule under a 0/1 loss function. Under a Bayesian GGM framework, $P(\boldsymbol{Y}^{new}|Z^{new} = k, \theta)$ is a Gaussian distribution, hence our classification rule is similar to a (quadratic) discriminant analysis. Discriminant analysis is a well-studied problem in classical multivariate statistics, in which the data are projected onto a low-dimensional space providing the maximum class separability [Duda et al. (2000)] and includes linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) as special cases. The LDA and QDA differ in the form of the optimal decision boundaries, which is linear in the former and nonlinear in the latter. Our Bayesian discriminant rule has three key advantages over the classical approach. First, we incorporate (integrate) the uncertainty of the parameters in deriving the optimal rule via Bayesian model mixing. Second, our network model provides an adaptively regularized estimate of the covariances and hence is capable of handling $n < p$ situations. Third, our network model uses this information in deriving the optimal classification boundary.

To illustrate the underpinnings of our network-based classifier, we illustrate a simple case using $k = 2$ groups. Assume

$$\begin{aligned} \boldsymbol{Y} \sim f_1 &= N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \text{with probability } \pi_1 \\ \boldsymbol{Y} \sim f_2 &= N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \quad \text{with probability } \pi_2, \end{aligned}$$

where $(\pi_1, \pi_2)$ are the prior odds of belonging to the classes and $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, which in our context codes for the network/pathway information. In this framework the classical decision boundary for discrimination is given by $\lambda(\boldsymbol{Y}) = log\frac{f_1(\boldsymbol{Y})}{f_2(\boldsymbol{Y})}$, which also happens to be the *a posteriori* log-odds ratio for popu-

lation 1 versus 2 (assuming equal prior odds for both populations, i.e., $\pi_1 = \pi_2$). Using simple algebraic manipulations of the multivariate normal densities given previously, one can show that

$$\lambda(\boldsymbol{Y}) = \beta_0 + \beta_L \boldsymbol{Y} + \boldsymbol{Y}^T \beta_Q \boldsymbol{Y}$$

where $\beta_0$ is the intercept component, $\beta_L$ is the linear component (in$\boldsymbol{Y}$), and $\beta_Q$ is the quadratic component of the discriminant function.

$$\begin{aligned}
\beta_0 &= \frac{1}{2}\{-log|\boldsymbol{\Sigma}_1| + log|\boldsymbol{\Sigma}_2| - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2\} \\
\beta_L &= -\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}_2^{-1} \\
\beta_Q &= \frac{1}{2}\{\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\}
\end{aligned}$$

Therefore, we can see that $\beta_0$ and $\beta_L$ are functions of $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ and $\beta_Q$ is a function of$(\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1})$ only. Note that the nonlinear classification decision boundary is only because of the presence of the term $\boldsymbol{Y}^T \beta_Q \boldsymbol{Y}$, which is a sole function of the covariances. Hence, network information from each of the classes is used to decide the boundary. We exploit this advantage of using the precision matrix information, which has been estimated using prior pathways of the proteins, to determine the optimal boundaries .

## C. Estimation Via MCMC

This subsection sets up the framework to carry out the Markov chain Monte Carlo (MCMC) calculations for our model parameters. The parameters and random variables to be estimated in the model are $\mathcal{M} \equiv \{\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{q}, \boldsymbol{S}, \sigma^2\}$. The conditional distribution of all the parameters except $\boldsymbol{q}$ and $\sigma^2$ are not available in closed form, hence we resort to a hybrid of the Gibbs and Metropolis-Hastings algorithms to explore the posterior distribution. We train the model using the training data set and use the classification scheme to predict the class of a new observation. In addition to this classification scheme, we can perform poste-

rior inference on the network parameters for significant graphs using MCMC samples, as outlined hereafter.

## D. FDR-based Determination of Significant Networks

Once we apply the MCMC methods, we are left with posterior samples of the model parameters that we can use to perform Bayesian inference. Our objective is twofold: to detect the "best" network/pathway based on the significance of the edges and also to detect differential networks between treatment groups/classes. Given $p$ proteins, our network consists of $p(p+1)/2$ unique edges, which could be large even for a moderate number of proteins. Therefore we need a mechanism that will control for these large scale comparisons, discover edges that are significant, and also detect differential edges between groups. We accomplish this in a statistically coherent manner using false discovery rate (FDR)-based thresholding to find significant networks and also to differentiate networks across samples.

The MCMC samples explore the distribution of possible network configurations suggested by the data, with each configuration leading to a different topology of the network based on the model parameters. Some edges that are strongly supported by the data may appear in most of the MCMC samples, whereas others with less evidence may appear less often. There are different ways to summarize this information in the samples. One could choose the most likely (posterior mode) network configuration and conduct conditional inference on this particular network topology. The benefit of this approach would be the yielding of a single set of defined edges, but the drawback is that the most likely configuration may still appear only in a very small proportion of MCMC samples. Alternatively, one could use all of the MCMC samples and, applying Bayesian model averaging (BMA) [Hoeting et al. (1999)], mix the inference over the various configurations visited by the sampler. This approach better accounts for the uncertainty in the data, leads to estimators

of the precision matrix with the smallest mean squared error, and should lead to better predictive performance in class predictions [Raftery et al. (1997)]. We will use this Bayesian model averaging approach.

From our MCMC method, suppose we have $M$ posterior samples of the corresponding parameter set $\{A_{ij}^{(m)}, m = 1, \ldots, M\}$, for which the selection indicator of the $ij$th edge is in the model. Suppose further that the model averaged set of posterior probabilities is set $\mathcal{P}$, the $ij$th element of which $\mathcal{P}_{ij} = M^{-1} \sum_m A_{ij}^{(m)}$ and is a $p \times p$ dimensional matrix. Note that $1 - \mathcal{P}_{ij}$ can be considered Bayesian q-values, or estimates of the local false discovery rate [Storey and Tibshirani (2003); Newton et al. (2004)] as they measure the probability of a false positive if the $ij$th edge is called a "discovery" or is significant. Given a desired global FDR bound $\alpha \in (0, 1)$, we can determine a threshold $\phi_\alpha$ to flag a set of edges $\mathcal{X}_\phi = \{(i, j) : \mathcal{P}_{ij} \geq \phi_\alpha\}$ as significant edges.

The significance threshold $\phi_\alpha$ can be determined based on classical Bayesian utility considerations such as those described in Muller et al. (2004) and based on the elicited relative costs of false-positive and false-negative errors or can be set to control the average Bayesian FDR, as in Morris et al. (2008). The latter is the process we follow here. For example, suppose we are interested in finding the value $\phi_\alpha$ that controls the overall average FDR at some level $\alpha$, meaning that we expect that only $100\alpha\%$ of the edges that are declared significant are in fact false positives. Let $\mathrm{vec}(\mathcal{P}) = [\mathcal{P}_\mathrm{t}; \mathrm{t} = 1, \ldots, \mathrm{p}^2]$ be the vectorized probability of the set $\mathcal{P}$, stacked columnwise. We first sort $\mathcal{P}_t$ in descending order to yield $\mathcal{P}_{(t)}, t = 1, \ldots, p^2$. Then $\phi_\alpha = \mathcal{P}_{(\xi)}$, where $\xi = \max\{j^* : j^{*-1} \sum_{j=1}^{j^*} \mathcal{P}_{(t)} \leq \alpha\}$. The set of regions $\mathcal{X}_{\phi_\alpha}$ then can be claimed to be significant edges based on an average Bayesian FDR of $\alpha$.

This FDR-based thresholding procedure can also be extended to find differential networks between different populations (tumor classes/subtypes), for example, to identify edges that are significantly different between tumor types. To this end, we use the cor-

responding parameter set $\{\lambda_{ij}^{(m)}, m = 1, \ldots, M\}$, for which the selection indicator of the differential edge between the $ij$th covariates in the model. The model averaged set of posterior probabilities is set $\mathcal{P}^d$, the $ij$th element of which $\mathcal{P}_{ij}^d = M^{-1} \sum_m \lambda_{ij}^{(m)}$. We this same procedure to arrive at a set of differential edges $\mathcal{X}_\phi = \{(i,j) : \mathcal{P}_{ij}^d \geq \phi_\alpha\}$ with $\phi_\alpha$ chosen to control the Bayesian FDR at level $\alpha$. We use a similar procedure on the parameter set $\{1 - \lambda_{ij}^{(m)}, m = 1, \ldots, M\}$, to arrive at a set of common edges $\mathcal{X}_\phi = \{(i,j) : \mathcal{P}_{ij}^c \geq \phi_\alpha\}$ with $\phi_\alpha$ chosen to control the Bayesian FDR at level $\alpha$.

### 1. Application of the methodology to reverse-phase protein lysate arrays

As explained, there is a strong rationale for methods that will directly assess the activation status of protein signaling networks in cancer. Traditional protein assays include immunohistochemistry (IHC), Western blotting, enzyme-linked immunosorbent assay (ELISA), and mass spectroscopy. Although IHC is a very powerful technique for the detection of protein expression and location, it is critically limited in network analyses by its non- to semi-quantitative nature. Western blotting can also provide important information, but due to its requirement for relatively large amounts of protein, it is difficult to use when comprehensively assessing protein networks, and also is semi-quantitative in nature. The ELISA method provides quantitative analysis, but is also limited by requirements of relatively high amounts of specimen and by the high cost of analyzing large pools of specimens. Mass spectroscopy is a powerful, quantitative approach, but its utility is mainly limited by the cost and time required to analyze individual samples, which limits the ability to run large sets that are needed to appropriately assess characteristics of disease heterogeneity and protein networks. Reverse-phase protein array (RPPA) analysis is a relatively new technology that allows for quantitative, high-throughput, time- and cost-efficient analysis of protein networks using small amounts of material [Paweletz et al. (2001); Tibes et al. (2006)]. In order to perform RPPA, proteins are isolated from cell lines, tumors, or serum using standard

methods. The protein concentrations are determined for samples. The concentrations of the samples are normalized and then the samples are denatured. Serial dilutions prepared from each specimen are then arrayed on a nitrocellulose-coated slide. Due to the small amount of protein spotted on each slide, 30 $\mu g$ of protein (comparable to the amount used on a single Western blot) can be printed on $\geq$ 150 slides. Each slide is probed with an antibody that recognizes a specific protein epitope, including phosphorylated residues that reflect the activation state of the protein. A visible signal is then generated through the use of HRP-conjugated secondary antibodies, a signal amplification system, and staining. The signal reflects the relative amount of that epitope in each spot on the slide, as shown in Figure 11. The arrays are then scanned and the resulting images are analyzed with MicroVigene, imaging software specifically designed for the quantification of RPPA analysis (VigeneTech Inc., Carlisle, MA). The relative signal intensities are used to determine background correction, to quantitate the relative concentration of each sample, and then to normalize loading differences [Hu et al. (2007); Neeley et al. (2009); Zhang et al. (2009)]. Background correction is used to separate the signal from the noise by subtracting the extracted background intensity from the foreground intensity. The quantification step determines the amount of protein present in a dilution series relative to other samples in the array. There are various ways to quantify the proteins in the sample depending on the underlying statistical model. For example, MicroVigene fits a four-parameter logistic model to each dilution series, whereas the method of Mircean et al. (2005) models the log intensity of the spots as a linear function of the dilution series. Both of these methods work on one sample at a time. Tabus et al. (2006) discussed a joint estimation method that used a logistic model in which a sigmoid shape is consistent with the observed intensity of a spot and the true protein concentration. This is due to quenching at high levels and background noise at low levels. An R package, SuperCurve, developed to use with this joint estimation method is available at `http://bioinformatics.mdanderson.org/Software/OOMPA`. As with

most high-throughput technologies, the normalization of the resulting intensities is conducted before any downstream analysis in order to adjust for sources of systematic variation not attributable to biological variation. We refer the reader to Paweletz et al. (2001) for more biological and technical details concerning RPPAs. The efficient, sensitive and

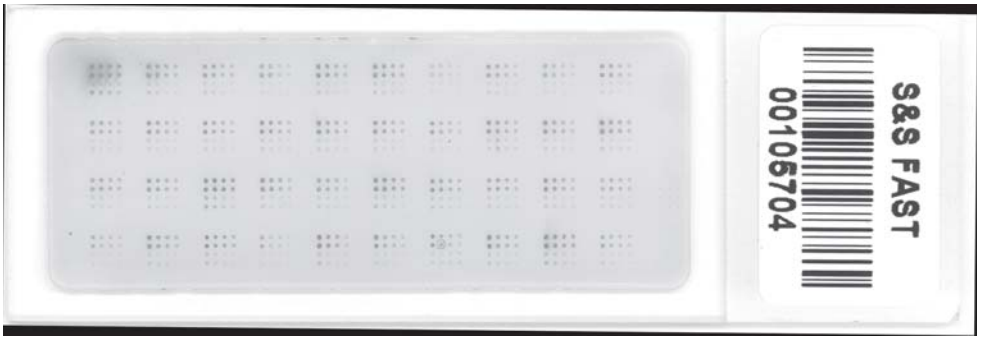



Fig. 11.: An example of a reverse-phase protein array (RPPA) slide with 40 samples shown as the 40 batches on the slide. Each batch represents one individual sample with 16 spots, which are the results of duplicates of 8-step dilutions.

quantitative nature of the RPPA technology allows for detailed and integrated analyses of protein signaling networks. We and other investigators have used RPPA to study kinase signaling pathways in multiple tumor types, including breast, ovarian, lung, and skin cancer [Sheehan et al. (2005); Stemke-Hale et al. (2008); Agarwal et al. (2009); Davies et al. (2009); O'Reilly et al. (2009); Park et al. (2010)]. We have used RPPA to characterize time-dependent changes in signaling networks in response to growth factor stimulation [Amit et al. (2007)]. We have also used RPPA to characterize signaling events that correlate with sensitivity and resistance to therapeutic agents [Hennessy et al. (2007); Mirzoeva et al. (2009)]. While these exploratory analyses have provided valuable information, the large amounts of novel data generated by the production of RPPAs provide us with the opportunity to develop and test rigorous statistical approaches to identify functional protein networks.

The scientific aims we address using RPPA data in this paper are three-fold: to utilize *a priori* information in inferring protein network topology within tumor classes/subtypes; to infer differential networks between tumor classes/subtypes; and finally to utilize network information in designing optimal classifiers for tumor classification. We believe this will improve our understanding of the regulation of protein signaling networks in cancer. Understanding the differences in protein networks between various cancer types and subtypes may allow for improved therapeutic strategies for each specific type of tumor. Such information may also be relevant when determining the origin of a tumor, which is clinically important in cases with indeterminate histologic analysis, particularly for patients who have more than one type of cancer.

E. Data Analysis

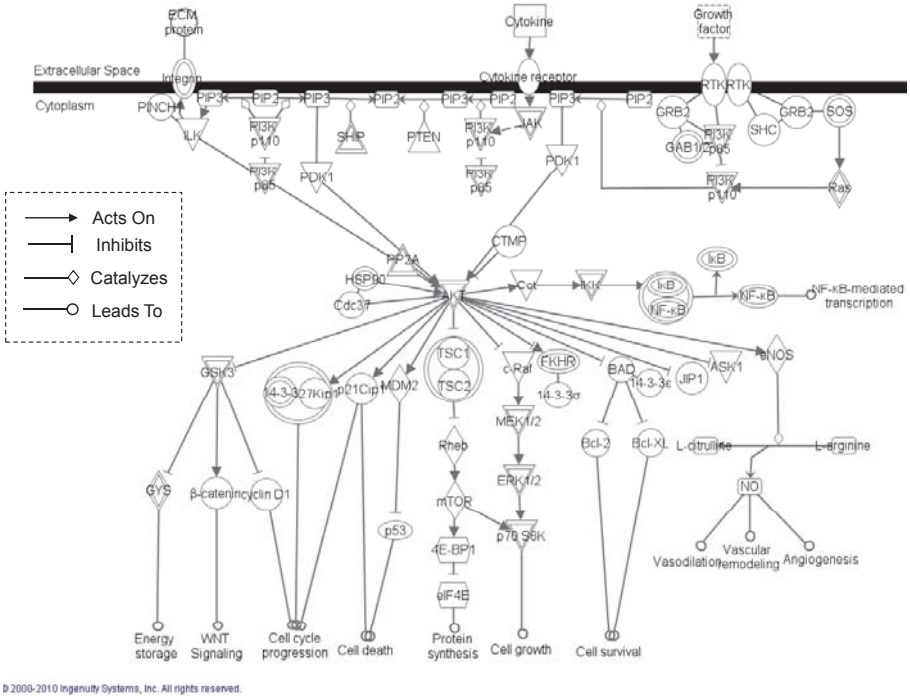
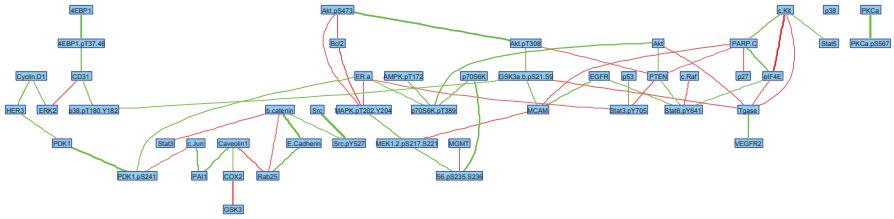
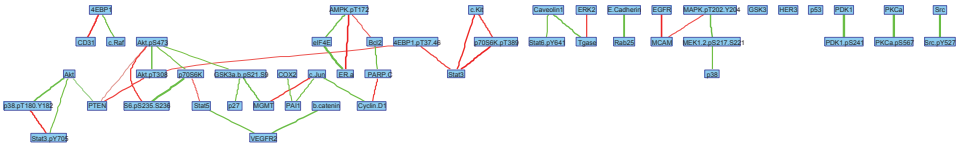



Fig. 12.: The PI3K-AKT Signaling Pathway. The pathway was generated through the use of Ingenuity Pathways Analysis (www.ingenuity.com).




(a) Breast Network

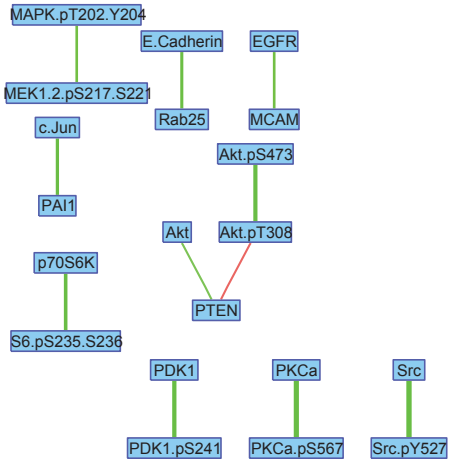



(b) Ovary Network

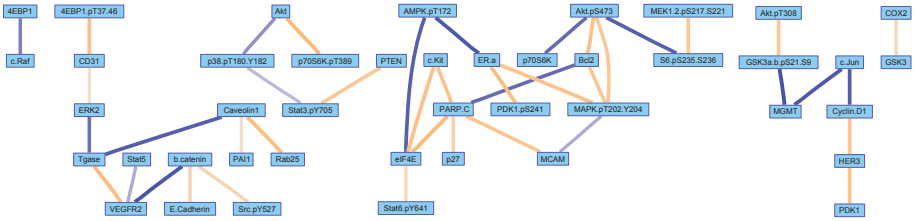
Fig. 13.: Significant edges for the proteins in the PI3K-AKT kinase pathway for breast (left panel) and ovarian cancer cell lines (right panel) computed using Bayesian FDR of 0.10. The red (green) lines between the proteins indicate a negative (positive) correlation between the proteins. The thickness of the edges corresponds to the strength of the associations, with stronger associations having greater thickness.

### 1. Classification of breast and ovarian cancer cell lines

Breast and ovarian cancer are two of the leading causes of cancer-related deaths in women [Jemal et al. (2008)]. Both of these diseases are frequently affected by mutations in kinase signaling cascades, particularly those involving components of the PI3K-AKT pathway [Mills et al. (2003); Hennessy et al. (2008); Yuan and Cantley (2008); Bast et al. (2009)]. The PI3K-AKT pathway is one of the most important signaling networks in carcinogenesis [Vivanco and Sawyers (2002)]. Our previous data have demonstrated that different mutations in the PI3K-AKT pathway may result in the activation of and functional dependence upon different effectors in this pathway [Vasudevan et al. (2009)]. PI3K is lipid kinase, which is activated by a number of different signals in carcinogenesis, including the stimulation of growth factors and other proteins that are frequently mutated in cancer tis-




(a) Conserved network between ovarian and breast cancer cell lines.




(b) Differential network between ovarian and breast cancer cell lines.

Fig. 14.: Conserved and differential networks for the proteins in the PI3K-AKT kinase pathway between breast and ovarian cancer cell lines computed using Bayesian FDR set to 0.10. In the conserved network (top panel), the red (green) lines between the proteins indicate a negative (positive) correlation between the proteins. In the differential network (bottom panel) the blue lines between the proteins indicate a relationship significant in ovarian cell lines that was not significant in the breast cell lines; the orange lines between the proteins indicate a significant relationship in the breast cell lines that was not significant in the ovarian cell lines. The thickness of the edges corresponds to the strength of the associations, with stronger associations having greater thickness.

sue. The importance of the PI3K-AKT pathway in carcinogenesis is supported by findings that this pathway is affected by activating mutations in cancer tissues more than any other signaling pathway[Yuan and Cantley (2008)]. This pathway may also be activated by the loss of function of the PTEN gene, which has been detected in many cancers (i.e., glioblastoma multiforme, and breast and ovarian cancer) and results in constitutive activation of the pathway [Davies et al. (1998, 1999)]. Broad genomic characterization of various cancers has demonstrated that while the prevalence of the individual mutations varies significantly among different cancer types, it is very common for tumors to have at least one genetic event that will activate the PI3K-AKT pathway. For example, the Cancer Genome Atlas analysis of mutations and copy number changes in glioblastoma multiforme identified at least one activating genetic event in genes in or immediately upstream of the PI3K-AKT pathway in 86% of the tumors [Cancer Genome Atlas Research Network (2008)]. Due to the body of evidence that the PI3K-AKT pathway plays a critical role in many cancers, this pathway has also been the subject of aggressive drug development efforts. Inhibitors of multiple different components of this pathway have been developed and are in various stages of preclinical and clinical testing [Hennessy et al. (2005); Courtney et al. (2010)].

We applied our methodology to identify differences in the regulation of the PI3K-AKT signaling network in breast and ovarian cancers. For this analysis, we used data for the expression of $p = 50$ protein markers in signaling pathways from an RPPA analysis of human breast ($n_1 = 51$) and ovarian ($n_2 = 31$) cancer cell lines grown under normal tissue culture conditions [Stemke-Hale et al. (2008)]. We used the known connections in the PI3K-AKT pathway (Figure 12) as *a priori* information in our model, by replacing the directed edges with undirected edges.

The significant networks based on a Bayesian FDR cutoff of $\alpha$ =0.1 for breast and ovarian cancer samples are shown in Figures 13(a) and 13(b), respectively. The red edges indicate a negative association (regulation) and the green edges indicate a positive

interaction between the proteins. The edges are represented by lines of varying degrees of thickness based on the strength of the association (correlation), with higher weights having thicker edges and lower weights having thinner edges. In order to identify biological similarities and differences between breast and ovarian cancers, we compared the results of our network analyses of the two cancer types. Plotted in Figure 14(a) are the conserved (common) edges between the two cancer types. The differential network between the two cancer types, controlling for a Bayesian FDR cutoff of $\alpha = 0.1$, is shown in Figure 14(b).

A number of protein-protein relationships demonstrated significant similarity between the two cancer types. For example, both breast cancer and ovarian cancer cell lines exhibited a marked negative association between the levels of PTEN and phosphorylated AKT (Akt.pT308). This relationship was expected due to the critical regulation of 3-phopshatidylinositols by the lipid phosphatase activity of PTEN, and has previously been demonstrated as a significant interaction in multiple tumor types [Davies et al. (1998, 1999); Stemke-Hale et al. (2008); Vasudevan et al. (2009); Davies et al. (2009); Park et al. (2010)]. Although this concordance was expected, our analysis also identified a large network of differential protein interactions in breast and ovarian cancers (Figure 14(b)). In this figure, the edges in blue indicate relationships between proteins that were present in the ovarian cancer cell lines but not in the breast cancer cell lines using our FDR cutoff, and the orange edges indicate relationships present in the breast cancer cell lines but not in the ovarian cancer cell lines. In addition, the thickness of the edges corresponds to the strength of the association. Notable differential connections in this analysis include the association of phosphorylated AKT (Akt.pS473) with BCL-2 (Bcl2) and phosphorylated MAPK (MAPK.pT202.Y204) in breast cancer. Both of these, BCL-2(Bcl2) and phosphorylated (activated) MAPK (MAPK.pT202.Y204), may contribute to tumor proliferation and survival, and are therapeutic targets with available inhibitors. The association of different proteins with the expression of the estrogen receptor, phosphorylated

PDK1 (PDK1.pS241) and MAPK (MAPK.pT202.Y204) in breast cancer and phosphorylated AMPK (AMPK.pT172) in ovarian cancer, may also have therapeutic implications, as the estrogen-receptor blockade is a treatment used in both advanced breast and ovarian cancers.

We used this network information to build a classifier to distinguish between breast cancer and ovarian cancer samples using the predictive probabilities approach, as explained. We assessed the performance of the classifiers using cross-validation techniques. In particular, we generated 100 random selections of test and training data sets with 66% and 33% splits of training and test data, respectively. We fit our Bayesian graph-based classifier (BGBC) and compared our method to four other methods: the K-nearest neighbor (KNN), linear discriminant analysis (LDA), diagonal linear discriminant analysis (DLDA) and diagonal quadratic discriminant analysis (DQDA) methods. The average misclassification errors (along with standard errors) across all splits for all the methods on the test set are shown in Table II. The BGBC method had much lower misclassification rates compared to the other methods (the other methods ignore the underlying network structure of the proteins). We believe that this improved precision is due to the fact that the mean expression profiles of the breast and ovarian cancers are very similar so there is not enough information in the mean to classify the two cases. So means-based classifiers, especially KNN and LDA (both of which use identity and diagonal covariances), underperform as compared to our method. The results of the DQDA method could be a bit closer to that of the BGBC method, but the former method ignores the cross-connections, i.e., network information, and hence results in a higher misclassification rate. The QDA could not be performed because the estimation of different covariance matrices for different classes is an ill-posed problem for $n < p$.

Nonlinear (quadratic) boundaries are obtained by using network information whereas linear boundaries are obtained by ignoring the network information. Figure 15 exemplifies

Table II.: Misclassification error rates for different classifiers for ovarian and breast cancer data sets. The methods compared here are LDA (linear discriminant analysis), KNN (K-nearest neighbor), DQDA (diagonal quadratic discriminant analysis), DLDA (diagonal linear discriminant analysis) and BGBC (Bayesian graph-based classifier), which is the method studied in this paper. The mean and the standard deviation are values of the percentage misclassification over 100 random splits of the data.

| Ovary Vs Breast | LDA | KNN | DQDA | DLDA | BGBC |
|---|---|---|---|---|---|
| Mean | 23.74 | 14.89 | 12.67 | 9.89 | 5.89 |
| Standard deviation | 11.64 | 5.81 | 5.70 | 5.40 | 4.41 |

our intuition and approach. We have a $p(=50)$-dimensional quadratic classification boundary. In order to visualize this we projected the boundary and the data onto two randomly selected dimensions/covariates. Two of those projections are shown in Figure 15. We can see how a nonlinear boundary is more effective than a linear boundary in classifying the data.

### 2. Effects of tissue culture conditions on network topology

Cell lines derived from tumors are a powerful research tool, as they allow for detailed characterization and functional testing. Genetic studies support the concept that cell lines generally mirror the changes that are detected in tumors, particularly at the DNA and RNA levels [Neve et al. (2006)]. However, the activation status of proteins can be impacted by the use of different environmental conditions in the culturing of cells. A key scientific question in the analysis of protein networks in cancer cell lines is the variability of network topologies due to differing tissue culture conditions. In order to test if different network connectivity is observed under varying culture conditions, we used three different tissue

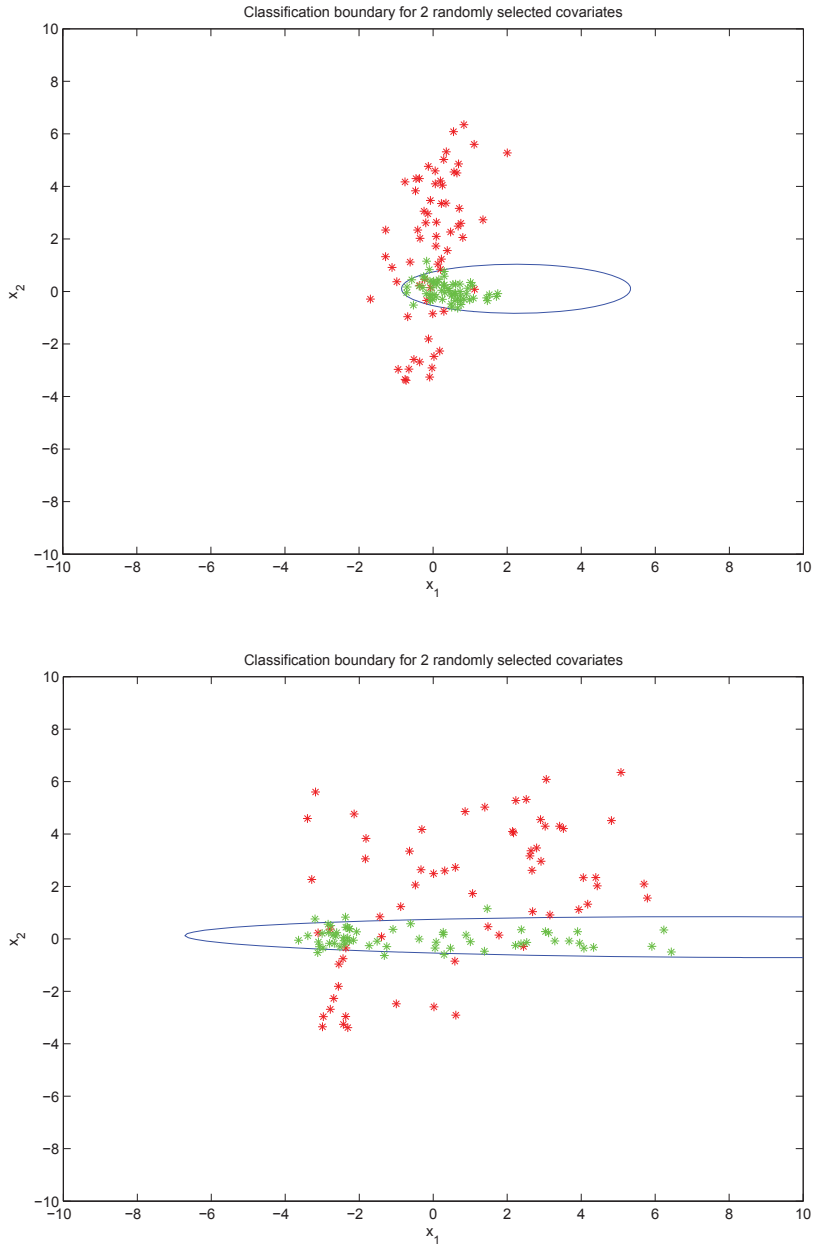
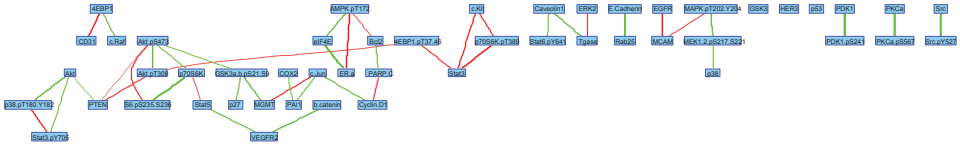
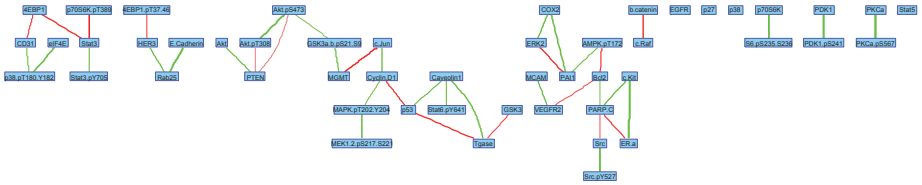



Fig. 15.: Nonlinear classification boundaries for two randomly selected covariates. Green points represent breast data and red points represent ovarian data. The blue line is the classification boundary determined by the model, which tries to differentiate between breast and ovarian data.

culture conditions to grow the 31 ovarian cancer cell lines used in the previous analysis. For condition “A”, the cells were grown in tissue culture media that was supplemented with growth factors in the form of fetal calf serum (5% of the total volume), which is a standard condition for the culturing of cancer cells. For condition “B”, the cells were harvested after being cultured in the absence of growth factors (serum) for 24 hours. For condition “C”, cells were grown in the absence of growth factors for 24 hours, then they were stimulated acutely (20 minutes) with growth factors (5% fetal calf serum). Proteins were harvested from each cell line for each tissue culture condition. The samples were then analyzed by RPPA. The RPPA data for each condition were then analyzed for protein-protein interactions using the GGM method. The topology maps for the ovarian cancer cells for the A, B, and C tissue culture conditions are shown in Figures 16(a), 16(b), and 16(c), respectively. We then performed comparisons of the results based on each of the three conditions in order to identify protein topology networks that were similar and different between each of the tissue culture conditions.



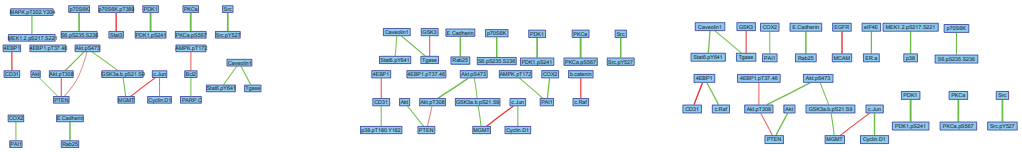
(a) Ovary-A Network



(b) Ovary-B Network

(c) Ovary-C Network

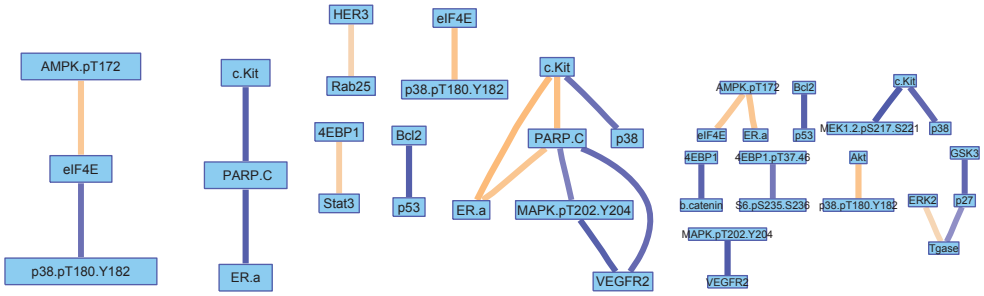
Fig. 16.: Significant edges for the proteins in the PI3K-AKT kinase pathway for ovarian cell lines grown in three different tissue culture conditions: A, B and C (see main text) computed using Bayesian FDR set to 0.10. The red (green) lines between the proteins indicate a negative (positive) correlation between the proteins. The thickness of the edges corresponds to the strength of the associations, with stronger associations having greater thickness.



(a) Conserved Network of OvaryA and OvaryB

(b) Conserved Network of OvaryB and OvaryC

(c) Conserved Network of OvaryA and OvaryC




(d) Differential Network of OvaryA and OvaryB

(e) Differential Network of OvaryB and OvaryC

(f) Differential Network of OvaryA and OvaryC

Fig. 17.: Conserved and differential networks for the proteins in the PI3K-AKT kinase pathway between ovarian cell lines grown in three different tissue culture conditions: A, B and C computed using Bayesian FDR set to 0.10. In the conserved network , the red (green) lines between the proteins indicate a negative (positive) correlation between the proteins. In the differential network, the blue lines between the proteins indicate a relationship significant in ovarian cell lines that was not significant in the breast cell lines; the orange lines between the proteins indicate a significant relationship in the breast cell lines that was not in the ovarian cell lines. The thickness of the edges corresponds to the strength of the associations, with stronger associations having greater thickness.

As conditions A (growth-factor replete media) and B (growth-factor starved media) both represented steady-state tissue culture conditions, we initially compared these protein networks using a Bayesian FDR of 10%. The networks that are shared between the two conditions are shown in Figure 17(a); the differential associations are presented in Figure 17(d). We detected 21 significant protein interactions that were common for conditions A and B, and 4 interactions that were different. Thus, the overwhelming majority of protein-protein associations that were observed were maintained regardless of the presence or absence of growth factors (serum) in the tissue culture media. We then compared the significant relationships identified for condition B (growth-factor starved media) versus condition C (starvation followed by acute stimulation). This comparison showed increased discordance of results, as we detected 20 associations that were common for conditions B and C [Figure 17(b)], but 11 associations that differed significantly [Figure 17(e)]. Similarly, the comparison of networks between the A and C conditions identified 22 shared protein interactions [Figure 17(c)] and 12 differential interactions [Figure 17(f)]. Of the differential interactions noted for the comparisons of conditions B versus C and A versus C, only 2 were observed in both comparisons (c-KIT and P38; VEGFR2 and MAPK.pT202.Y204). Neither of these 2 relationships was among the differential protein interactions in the analysis of condition A versus condition B. Of the 4 relationships that differed in the comparison of condition A versus condition B, 3 of the relationships were also identified as differing significantly when comparing condition B versus condition C (eIF4E and P38.pT180.Y182; c-Kit and PARP.cleaved; PARP.cleaved and ER.alpha), and the fourth differed significantly for the comparison of condition A versus condition C (AMPK.pT172 and eIF4E). This analysis suggests that protein-protein relationships are largely maintained under steady-state tissue culture conditions. However, these interactions may differ significantly in the setting of acute growth factor stimulation.

## F. Discussion and Conclusions

We present methodology to model sparse graphical models in the presence of class variables in high-dimensional settings, with a particular focus on protein signaling networks. Our method allows for the effective use of prior information about signaling pathways that is already available to us from various sources to help in decoding the complex protein networks. We also emphasize the differential and common networks between the classes of cancers/tumor conditions. Improved understanding of the differential networks can be crucial for biologists when designing their experiments, by allowing them to concentrate on the most important factors that distinguish tumor types. Such information may also help to narrow the drug targets for specific cancers. Knowledge of the common networks can be used to develop a drug for two different cancers that targets proteins that are active in both cancers. Data on the differential edges may be used as a good screening analysis, allowing researchers to eliminate unimportant proteins and concentrate on effective proteins when designing advanced patient-based translational experiments. In this article we focused on undirected graphical models and not on directed (casual) networks. Directed graphical models, such as Bayesian networks and directed acyclic graphs (DAGs), have explicit causal modeling goals that require further modeling assumptions. In our formulation, we provide a natural and useful technical step in the identification of high posterior probability undirected graphical models, assuming a random sampling paradigm. In addition, our models infer network topologies that assume a steady-state network. Some of the protein networks may be dependent on causal relations between the nodes, which would require us to model data over time to infer the complete dynamics of the network. We leave this task for future consideration.

With regard to computation time, our MCMC chains are fairly fast for a high-dimensional data set like those we considered, with a 5000-iteration run taking about 15 minutes. The

source code, in MATLAB (The Mathworks, Inc., Natick, MA), takes advantage of several matrix optimizations available in that language environment. The computationally-involved step is the imposition of a positive definiteness on the correlation matrix. Optimizations to the code have been made by porting some functions into C. The software is available by emailing the first author.

# CHAPTER III

# MIXTURES OF GAUSSIAN GRAPHICAL MODELS

## A. Finite Mixtures of Gaussian Graphical Models

### 1. Introduction

One of the strengths of the proposed methods is it can be employed in a more complex modeling framework in a hierarchical manner. We use it to develop finite mixture graphical models, where in each each mixture component is assumed to follow a Gaussian graphical model with a adaptive covariance structure. Thus we model the dependencies of variables within the mixture components in a flexible manner in addition as opposed to traditional mixture models (Mclachlan and Peel, 2000), which typically assume independence.

Our motivation for this model arises from a high-throughput genomics example. Suppose we have a gene expression data set with $n$ samples and $g$ genes. We are interested in detecting $k$ sub-types of cancer among the $n$ samples. Furthermore, we assume a different network structure of these $g$ genes for each cancer sub-type and it is our primary goal to use this information efficiently to cluster the samples into the correct sub-type of cancer. Additionally, we wish to learn about these networks for different sub-types of cancer to identify biologically significant differences among them that explain the variation between the sub-types.

### 2. The hierarchical model

Let $\boldsymbol{Y}_{p\times n} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a $p \times n$ matrix with $n$ samples and $p$ covariates. Here each of the $n$ samples belongs to one of the $K$ hidden groups or strata. Each sample $\boldsymbol{Y}_i$ follows a multivariate normal distribution $\boldsymbol{N}(\theta_j, \boldsymbol{\Sigma_j})$ if it belongs to the $j$th group. Given a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ , we wish to estimate the number of mixtures $k$ as well as the precision

matrices $\boldsymbol{\Omega_j} = \boldsymbol{\Sigma_j^{-1}}$, $j = 1, \cdots, k$. Conditional on the number of mixtures ($K$) we fit a finite mixture model, then vary the number of mixtures and select the optimal number of mixtures using BIC, as explained in Appendix.

We introduce the latent indicator variable $L_i \in 1, 2, \ldots, K$, which corresponds to every observation $\boldsymbol{Y_i}$ that indicates which component of the mixture is associated with $\boldsymbol{Y}_i$, i.e., $L_i = j$ if $\boldsymbol{Y}_i$ belongs to the $j^{th}$ group. *A priori*, we assume $P(L_i = j) = p_j$ such that $p_1 + p_2 + \ldots + p_K = 1$. We can then write the likelihood of the data conditional on the latent variables as

$$\boldsymbol{Y}_i | L_i = j, \boldsymbol{\theta}, \boldsymbol{\Omega} \sim N(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1}).$$

The latent indicator variables are allowed *a priori* to follow a multinomial distribution with probabilities $p_1, \ldots, p_K$ as

$$L_i \sim \text{Multinomial}(1, [p_1, p_2, \ldots, p_K]),$$

and the associated class probabilities follow a Dirichlet distribution as

$$p_1, p_2, \ldots, p_K | \alpha \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K).$$

We allow the individual means of each group to follow a Normal distribution as

$$\boldsymbol{\theta}_j | \boldsymbol{B} \sim N(\boldsymbol{0}, \boldsymbol{B}),$$

We assign a common inverse Wishart prior for covariance matrix $\boldsymbol{B}$ across groups as $\boldsymbol{B} \sim IW(\boldsymbol{\nu_0}, \boldsymbol{B_0})$, where $\boldsymbol{\nu_0}$ is the shape parameter and $\boldsymbol{B_0}$ is the scale matrix.

The hierarchical specification of the GGM structure for each group $\boldsymbol{\Omega}_j$ parallels the development of the previous subsection, with each GGM indexed by its own mixture-specific parameters to allow the sparsity to vary within each cluster component. The hierarchical

model for finite mixture GGMs can be summarized as follows:

$$
\begin{aligned}
\boldsymbol{Y}_i|L_i = j, \boldsymbol{\theta}, \boldsymbol{\Omega} &\sim N(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1}) \\
L_i &\sim \text{Multinomial}(1, [p_1, p_2, \ldots, p_K]) \\
p_1, p_2, \ldots, p_K|\alpha &\sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K) \\
\boldsymbol{\theta}_j|\boldsymbol{B} &\sim N(\boldsymbol{0}, \boldsymbol{B}) \\
\boldsymbol{B} &\sim IW(\boldsymbol{\nu_0}, \boldsymbol{B_0}) \\
\boldsymbol{\Omega}_j &= \boldsymbol{S}_j(\boldsymbol{A_j} \odot \boldsymbol{R_j})\boldsymbol{S}_j \\
A_{j_{(lm)}}|q_{j_{(lm)}} &\sim \text{Bernoulli}(q_{j_{(lm)}}), i < j \\
\boldsymbol{R_j}|\boldsymbol{A_j} &\sim \prod_{l<m} \text{Laplace}(0, \tau_{j_{(lm)}}) I(\boldsymbol{C}_j \in \mathbb{C}_p) \\
\tau_{j_{(lm)}} &\sim IG(e, f) \\
S_{j_{(l)}} &\sim IG(g, h),
\end{aligned}
$$

where $i$ denotes the sample, $j$ denotes the mixture component, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, K$. In addition, $A_{j_{(lm)}}$ and $\tau_{j_{(lm)}}$ denote the $lm^{th}$ component of the $\boldsymbol{A}_j$ and the $\boldsymbol{\tau}_j$, $l = 1, 2, \ldots, p$ , $m = l, \ldots, p$.

### 3. Posterior inference and the conditional distributions

We perform the posterior inference using MCMC methods; hence we derive the full conditionals for all the parameters. Not all the full conditionals are in a closed form; and in those situations we employ the MH algorithm to simulate those parameters.

**Sampling probabilities** $p_j$.

We draw the probabilities from a Dirichlet distribution, which can be done by drawing each probability from a gamma distribution with the corresponding Dirichlet parameter

and normalizing them so that their sum is equal to 1.

$$\begin{aligned} p_1, p_2, \ldots, p_K \qquad |\text{Others} \propto \qquad & \prod_{j=1}^{K} p_j^{\alpha_j - 1} \prod_{j=1}^{K} p_j^{n_j} \\ \sim \qquad & Dirichlet(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_K + \alpha_K). \end{aligned}$$

**Sampling Class Indicators** $L_i$.

The full conditional of $L_i$ is

$$P(L_i = j|\text{Others}) \qquad = \qquad \frac{p_j \phi_{Y_i}(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1})}{\sum_{j=1}^{K} p_j \phi_{Y_i}(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1})}.$$

Each of the class indicators $L_i$ can be drawn from a multinomial distribution with the above probability.

**Sampling class means** $\boldsymbol{\theta}_j$.

The conditionals for the means of the corresponding mixtures are from a multivariate normal distribution, so we can directly sample them:

$$\begin{aligned} \boldsymbol{\theta}_j|\text{Others} \qquad \propto \qquad & N_{\boldsymbol{\theta}_j}(\mathbf{0}, \boldsymbol{B}) \times \prod_{i=1}^{n_j} N_{\boldsymbol{Y}_i}(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1}) \\ \propto \qquad & exp(-\frac{1}{2}\boldsymbol{\theta}_j^T \boldsymbol{B}^{-1} \boldsymbol{\theta}_j) \times exp(-\frac{1}{2}\sum_{i=1}^{n_j} (\boldsymbol{\theta}_j - \boldsymbol{Y}_i)^T \boldsymbol{\Omega}_j (\boldsymbol{\theta}_j - \boldsymbol{Y}_i)) \\ \propto \qquad & exp(-\frac{1}{2}\boldsymbol{\theta}_j^T [n_j \boldsymbol{\Omega}_j + \boldsymbol{B}^{-1}] \boldsymbol{\theta}_j - 2\boldsymbol{\theta}_j \boldsymbol{\Omega}_j \sum_{i=1}^{n_j} \boldsymbol{Y}_i \\ & \qquad + (\sum_{i=1}^{n_j} \boldsymbol{Y}_i)^T \boldsymbol{\Omega}_j [n_j \boldsymbol{\Omega}_j + \boldsymbol{B}^{-1}]^{-1} \boldsymbol{\Omega}_j \sum_{i=1}^{n_j} \boldsymbol{Y}_i) \\ \sim \qquad & N_{\boldsymbol{\theta}_j}([n_j \boldsymbol{\Omega}_j + \boldsymbol{B}^{-1}]^{-1} \boldsymbol{\Omega}_j \sum_{i=1}^{n_j} \boldsymbol{Y}_i, [n_j \boldsymbol{\Omega}_j + \boldsymbol{B}^{-1}]^{-1}). \end{aligned}$$

**Sampling Correlation and Other Parameters Related to the Precision Matrix:**

The sampling of all these conditionals is similar to sampling from the previous selection

model with slightly different expressions as we have to sample from each cluster

$$[R_{j_{(lm)}}|A_{j_{(lm)}}, others] \propto |\boldsymbol{\Omega}_j|^{\frac{n_j}{2}} exp\{-\frac{1}{2}\sum_{i=1}^{n_j}\{(\boldsymbol{Y}_i-\boldsymbol{\theta}_j)^T\boldsymbol{\Omega}_j(\boldsymbol{Y}_i-\boldsymbol{\theta}_j) - \frac{1}{\tau_{j_{(lm)}}}|R_{j_{(lm)}}|\}\}$$
$$I(\boldsymbol{C}\in\mathbb{C}_p)$$

$$[A_{j_{(lm)}}|R_{j_{(lm)}}, others] \propto |\boldsymbol{\Omega}_j|^{\frac{n_j}{2}} exp\{-\frac{1}{2}\sum_{i=1}^{n_j}\{(\boldsymbol{Y}_i-\boldsymbol{\theta}_j)^T\boldsymbol{\Omega}_j(\boldsymbol{Y}_i-\boldsymbol{\theta}_j)\}\}$$
$$q_{j_{(lm)}}^{A_{j_{(lm)}}}(1-q_{j_{(lm)}}^{1-A_{j_{(lm)}}})I(\boldsymbol{C}\in\mathbb{C}_p)$$

Here we use the similar approaches as used in the selection model by griding the conditional distribution between $\{u_{j_{(lm)}}, v_{j_{(lm)}}\}$ and drawing directly from the conditional.

We draw $\tau_{j_{(lm)}}$'s and $q_{j_{(lm)}}$'s using the MH algorithm. The expression for the normalizing constant $K(\tau_{j_{(lm)}}, q_{j_{(lm)}})$ is similar to the expression given before

$$\tau_{j_{(lm)}}|q_{j_{(lm)}}, A_{j_{(lm)}}, R_{j_{(lm)}}, \boldsymbol{Y} \propto K(\tau_{j_{(lm)}}, q_{j_{(lm)}})\frac{1}{\tau_{j_{(lm)}}}exp(\frac{-|A_{j_{(lm)}}R_{j_{(lm)}}|}{\tau_{j_{(lm)}}})$$
$$\times \tau_{j_{(lm)}}^{-g-1}exp(-\frac{h}{\tau_{j_{(lm)}}})$$
$$q_{j_{(lm)}}|\tau_{j_{(lm)}}, A_{j_{(lm)}}, R_{j_{(lm)}}, \boldsymbol{Y} \propto K(\tau_{j_{(lm)}}, q_{j_{(lm)}})q_{j_{(lm)}}^{A_{j_{(lm)}}}(1-q_{j_{(lm)}})^{(1-A_{j_{(lm)}})}$$
$$q_{j_{(lm)}}^{\alpha-1}(1-q_{j_{(lm)}})^{(\beta-1)}.$$

Similarly we draw $S_{j_{(l)}}$ using the MH algorithm from the conditional distribution:

$$S_{j_{(l)}}|\boldsymbol{S}_{j_{(-l)}}, Y \propto \prod_{i=1}^{n_j}|\boldsymbol{S}_j(\boldsymbol{C}_j)\boldsymbol{S}_j|^{1/2}exp\{\frac{-1}{2}\{(\boldsymbol{Y}_i-\boldsymbol{\theta}_j)^T(\boldsymbol{S}_j(\boldsymbol{C}_j)\boldsymbol{S}_j)(\boldsymbol{Y}_i-\boldsymbol{\theta}_j)\}\}$$
$$\times S_{j_{(l)}}^{-g-1}exp(\frac{-h}{S_{j_{(l)}}}).$$

# 4. Real data example

We used the leukaemia data from Golub et al. (1999) as a case study to illustrate our graphical mixture model. In this study, the authors measured the human gene expression signatures of acute leukaemia. They used supervised learning to predict the type of leukaemia and used unsupervised learning to discover new classes of leukaemia. The motivation for this work was to improve cancer treatment by distinguishing between subclasses of cancers or tumors. The data are available from `http://www.genome.wi.mit.edu/MPR`. The data set includes 6817 genes and 72 patient samples. We selected the 50 most relevant genes, identified using a Bayesian gene selection algorithm (Lee et al., 2003). The heat map of the top 50 genes in the data set is shown in Figure 18. In the heat map, which shows the expression profiles of the genes, we can observe distinct groups of genes that behave concordantly. We wanted to explicitly explore the dependence patterns that vary by group.

We fit our Bayesian mixture of graphical models to this data set using Bayesian lasso selection models and used the methods detailed in chapter 1 to find the top graphs for the data. We ran the MCMC simulation for 100000 samples and removed the first 20000 samples as burn-in. We selected the number of mixtures using BIC, as described in Appendix B. Using this criterion and without *a priori* knowledge, we determined two clusters as corresponding best to two subtypes of leukaemia: (1) acute lymphoblastic leukaemia (ALL) and (2) acute myelogenous leukaemia (AML). The respective networks corresponding to the two clusters are shown in Figure 19 and Figure 20. As shown in the figures, the networks for these two clusters are quite different, which suggests possible interactions between genes that differ depending on the subtype of leukaemia.

We further explored the biological ramifications of our findings using the gene annotations also used by Golub et al. (1999). Most of the genes active in the ALL network are inactive in the AML network and vice versa. It is known that ITGAX and CD33 encode

cell surface proteins for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells. We can see in the networks of the clusters that CD33 and ITGAX are active in the AML network but inactive in the ALL network. The zyxin gene has been shown to encode a LIM domain protein important in cell adhesion in fibroblasts, but a role for zyxin in haematopoiesis has not been reported. Zyxin is also active in the AML network but not in the ALL network. In general, the genes most useful in distinguishing AML vs. ALL class prediction are markers of haematopoietic lineage, which are not necessarily related to cancer pathogenesis. However, many of these genes encode proteins critical for S-phase cell cycle progression (CCND3, STMN1, and MCM3), chromatin remodelling (RBBP4 and SMARC4), transcription (GTF2E2), and cell adhesion (zyxin and ITGAX), or are known oncogenes (MYB, TCF3 and HOXA9). The genes encoding proteins for S-phase cell cycle progression (CCND3, STMN1, and MCM3) were all found to be active in the ALL network but inactive in the AML network. This suggests a connection of ALL with the S-phase cell cycle. Genes responsible for chromatin remodelling and transcriptional factors were present in both networks, indicating they are common to both types of cancer. This information can be used to discover a common drug for both types of leukaemia. Among the oncogenes, MYB was related to the ALL network, whereas TCF3 and HOXA9 were related to the AML network. HOXA9 is rearranged by a t(7;11)(p15;p15) chromosomal translocation in a rare subset of individuals with AML who tend to have poor outcomes. Furthermore, HOXA9 overexpression has been shown to transform myeloid cells *in vitro* and to cause leukaemia in animal models. A general role for the HOXA9 expression in predicting AML outcomes has been suggested by Golub et al. (1999). We also confirmed that HOXA9 is active in the AML network, but not in the ALL network.

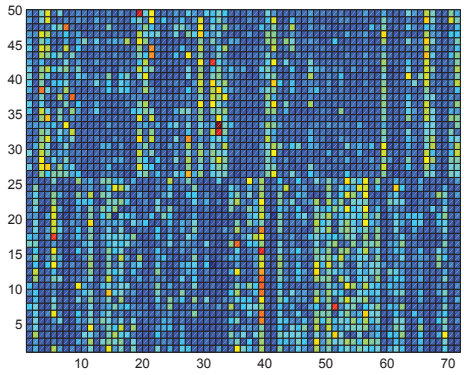



Fig. 18.: Heat map of top 50 genes in leukaemia data set.

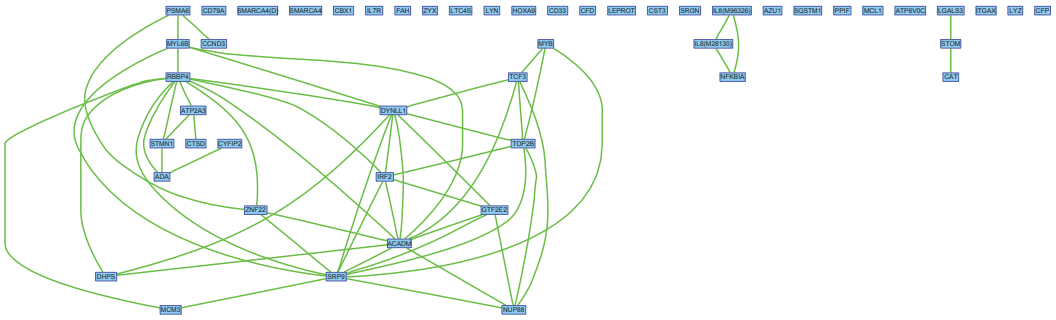



Fig. 19.: Significant edges for the genes in the ALL cluster. The red (green) lines between the proteins indicate a negative (positive) correlation between the proteins.

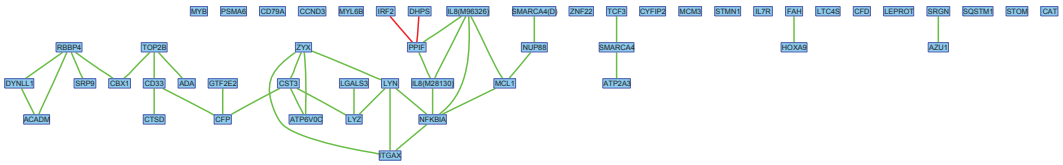



Fig. 20.: Significant edges for the genes in the AML cluster. The red (green) lines between the proteins indicate a negative (positive) correlation between the proteins.

## 5. Simulations

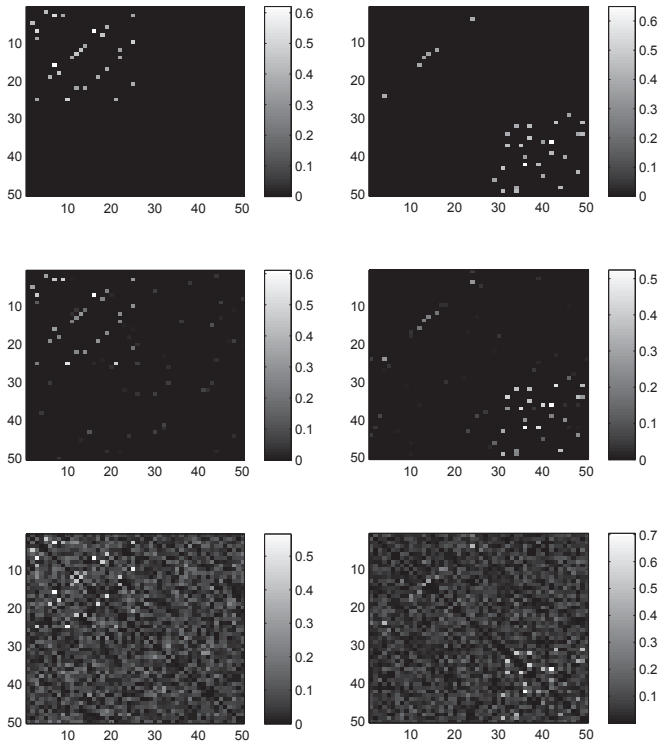
We performed a posterior predictive simulation study to evaluate the operating characteristics of our methodology for mixtures of graphical models. We simulated data from our fitted model of the leukaemia data set using the estimated precision matrices for the two groups, ALL and AML. The simulation was conducted as follows. Let $(\hat{\boldsymbol{\mu}}_{\boldsymbol{j}}, \hat{\boldsymbol{\Omega}}_{\boldsymbol{j}}^{-\mathbf{1}})$ denote the estimates of the mean and precision matrices corresponding to the ALL ($j = 1$) and AML ($j = 2$) groups, respectively, as obtained in the previous subsection. We generated data under the convolution of the following multivariate normal likelihood,

$$\boldsymbol{Y}_j \sim N(\hat{\boldsymbol{\mu}}_{\boldsymbol{j}}, \hat{\boldsymbol{\Omega}}_{\boldsymbol{j}}^{-\mathbf{1}}),$$

with 100 samples and 50 covariates.

We (re-)fitted our models to the simulated data and compared the estimates of the covariance matrices obtained from a non-adaptive finite mixture model (MCLUST) of Fraley and Raftery (2007). We used the "VVV" setting, which implies the use of an unconstrained covariance estimation method in their procedure. We completed 100000 runs of the MCMC simulation and removed the first 10000 runs as burn-in. The true and corresponding estimates of the precision matrices using the two methods are shown in Figure 21, where the absolute values of the precision matrix excluding the diagonal are plotted.

As shown in the figure, fitting our adaptive model to the data (middle row of images) yields estimates that are closer (sparser) to the true data generating precision matrices, whereas fitting the non-adaptive model to the data (bottom row of images) yields noisier estimates, with less local shrinkage of the off-diagonal elements. In addition to a visual inspection, we compared the performance of both methods using the K-L distance. The corresponding estimates of the K-L distances were 3.5592 and 7.2210 for the adaptive and non-adaptive model fits, respectively. For the AML cluster we obtained respective K-L

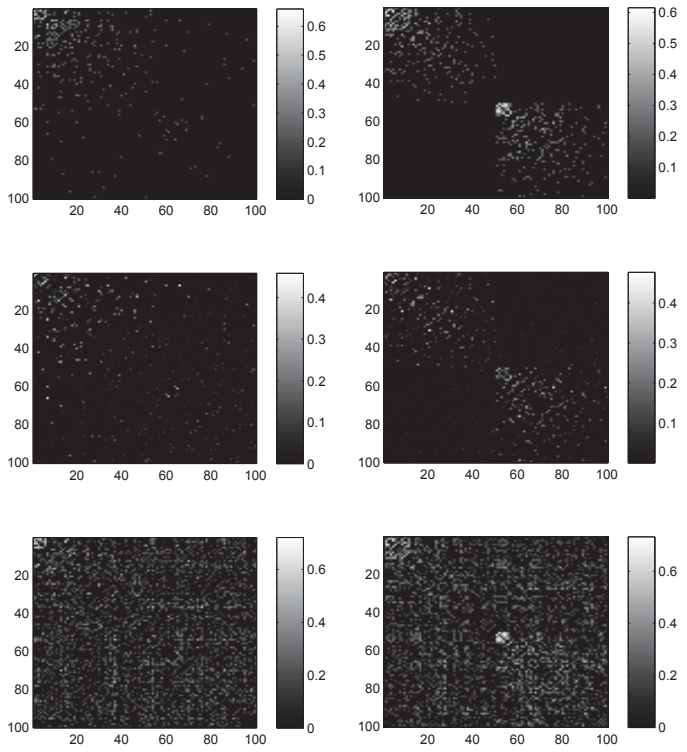


(a) ALL (b) AML

Fig. 21.: Simulation study (p=50). The true and estimated precision matrices for two subtypes of leukaemia: (a) ALL and (b) AML. The top row of images shows the true data generating precision matrix; the middle row shows the estimated precision matrix using our adaptive Bayesian model; and the bottom row shows the estimated precision matrix using a non-adaptive fit. Note that the absolute values of the partial correlations are plotted in the above figures without the diagonal. The colorbars are shown to the right of each image.

distances of 4.0836 and 7.7881 for the two methods. In addition, we also compared the false positive and false negative rates for finding true edges using each method. It should be noted that the purpose of the MCLUST approach is not covariance selection, hence we imposed selection on the elements of the estimated precision matrix by thresholding the coefficients to zero if they were less than a defined constant. We chose a fairly generous thresholding constant so that the false negatives and false positives were minimized. We applied the thresholding constant of 0.15 to the coefficients of the precision matrices that were estimated for the two clusters. For the AML cluster, we found false positive rates of (0.0049, 0.0645) and false negative rates of (0.0106, 0) for our adaptive model and the MCLUST approach, respectively. For the ALL cluster, we found false positive rates of (0.0041, 0.0661) and false negative rates of (0.0131, 0) for the adaptive and non-adaptive model fits, respectively. In summary, our adaptive method performs substantially better in recovering the true sparse precision matrix compared to the simple (non-adaptive) clustering approaches.

To explore how the method scales with the number of covariates, we ran another simulation with 100 covariates and 200 samples. The results are plotted in Figure 22. We find a similar pattern of performance from fitting our adaptive model to the data (middle row of images), which yields estimates that are closer to the true data generating precision matrices. By contrast, fitting the non-adaptive model to the data (bottom row of images) yields noisier estimates, with less local shrinkage of the off-diagonal elements. Again we compared the performance of both methods using the K-L distance and determined that the corresponding estimates were 10.1241 and 25.3378 for the adaptive and non-adaptive model fits, respectively. For the AML cluster, we obtained K-L distances of 12.1244 and 27.4851 for the respective methods. We chose a thresholding constant of 0.15 and applied that to the coefficients of the precision matrices that were estimated for the two clusters. For the AML cluster we found false positive rates of (0.0063, 0.2822) and false negative rates of (0.0222, 0.0081) for our adaptive model and the MCLUST approach, respectively. For the ALL cluster, we found false positive rates of (0.0044, 0.2497) and false negative rates of (0.101, 0.0372) for the adaptive and non-adaptive fits, respectively. Thus, compared to the non-adaptive approaches, our adaptive method performed substantially better in recovering the true sparse precision matrix. We found that our methods scale reasonably until we reach around 500 covariates but above that level the high computational complexity did not allow for a reasonable computation time. Parallel computation in cluster machines can be used to speed up the process when the number of covariates extremely high. Alternatively, we plan to explore faster deployments of our algorithm through variational approach or other approximations.




(a) ALL (b) AML

Fig. 22.: Simulation study(p=100). True and estimated precision matrices for two subtypes of leukaemia: (a) ALL and (b) AML. The top row of images shows the true data generating precision matrix; the middle row shows the estimated precision matrix using our adaptive Bayesian model; and the bottom row shows the estimated precision matrix using a non-adaptive fit. Note that the absolute values of the partial correlations are plotted in the above figures without the diagonal. The colorbars are shown to the right of each image.

## B. Infinite Mixtures of Graphical Models

When the number of mixtures is unknown, the parameters $\gamma_i$ of the model are specified by Dirichlet process priors for clustering. The Dirichlet Process (DP) is a non-parametric two-parameter conjugate family in the sense that there is a positive probability that a sample distribution will approximate arbitrarily well any distribution that is dominated by the base distribution $H_\phi$. DPs are also a.s. discrete and comprise a certain partitioning of the parameter space. These properties allow us to model clustering configurations of a set of variables by DP priors without fixing the number of clusters beforehand.

In a sequence of draws $\gamma_1, \gamma_2, \ldots$ from the Polya urn representation of the Dirichlet process (Blackwell and MacQueen, 1973), the $n$th sample is either distinct with a small probability $\alpha/(\alpha + n - 1)$ or is tied to previous sample with positive probability to form a cluster. Let $\gamma_{-n} = \{\gamma_1, \ldots, \gamma_n\} - \{\gamma_n\}$ and $d_{n-1}$= number of preexisting clusters of tied samples in $\gamma_{-n}$ at the $n$th draw, then we have

$$f(\gamma_n|\gamma_{-n}, \alpha, \phi) = \frac{\alpha}{\alpha + n - 1} H_\phi + \sum_{j=1}^{d_{n-1}} \frac{n_j}{\alpha + n - 1} \delta_{\bar{\gamma}_j}, \tag{3.1}$$

where $H_\phi$ is the base prior, and the $j$th cluster has $n_j$ tied samples that are commonly expressed by $\bar{\gamma}_j$ subject to $\sum_{j=1}^{d_{n-1}} n_j = n - 1$. After $n$ sequential draws from the Polya urn, there are several ties in the sampled values and we denote the set of distinct samples by $\{\bar{\gamma}_1, \ldots, \bar{\gamma}_{d_n}\}$, where $d_n$ is essentially the number of clusters.

Let $\boldsymbol{Y}_{p\times n} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a $p \times n$ matrix with $n$ samples and $p$ covariates. Each sample $\boldsymbol{Y}_i$ follows a multivariate normal distribution $\boldsymbol{N}(\boldsymbol{\theta}_j, \boldsymbol{\Sigma_j})$. Given a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ , we wish to estimate the number of mixtures $k$ as well as the precision matrices for each cluster $\boldsymbol{\Omega_j} = \boldsymbol{\Sigma_j^{-1}},\ j = 1, \cdots, k$. We can write the likelihood of the data as

$$\boldsymbol{Y}_i|\boldsymbol{\theta}_i, \boldsymbol{\Omega}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Omega}_i^{-1}),$$

Let $\gamma_i = (\boldsymbol{\theta}_i, \boldsymbol{\Omega}_i^{-1})$. We propose a Dirichlet process prior on $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ which can be written as

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n) \sim DP(\alpha, H_\phi),$$

Here $H_\phi$ is the base distribution of the DP. We induce sparsity into the model using the base distribution which defines the cluster configuration. We allow the individual means of each group to follow a Normal distribution as

$$\boldsymbol{\theta}_j | \boldsymbol{\Omega} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{-1}),$$

The hierarchical specification of the GGM structure for each group $\boldsymbol{\Omega}_j$ parallels the development of the previous subsection, with each GGM indexed by its own mixture-specific parameters to allow the sparsity to vary within each cluster component. The hierarchical model for the baseline prior can be summarized as follows:

$$\begin{aligned}
H_\phi &\propto N_{\boldsymbol{\theta}}(.) F_{\boldsymbol{\Omega}} \\
\boldsymbol{\Omega} &= \boldsymbol{S}(\boldsymbol{A} \odot \boldsymbol{R})\boldsymbol{S} \\
F_\Omega &\propto F_A(.) F_R(.) F_S(.) \\
\boldsymbol{R} | \boldsymbol{A} &\sim Uniform(0,1) I(\boldsymbol{C}_j \in \mathbb{C}_p) \\
\boldsymbol{A} | \boldsymbol{Q} &\sim Bernoulli(\boldsymbol{Q}) \\
\boldsymbol{Q} &\sim Beta(\nu_c, \nu_d) \\
\boldsymbol{S} &\sim IG(\nu_\alpha, \nu_\beta),
\end{aligned}$$

The parameters are similar to the models detailed in the previous chapters. The base prior is not in conjugate form so the base prior is not integrable with the likelihood to draw from the posterior using Gibbs sampling framework ( Escobar and West(1995)). We need to use Metropolis Hastings framework to handle the non-conjugate priors(Neal(2000)).

Let us introduce a latent variable, the class indicator for the $i^{th}$ sample $c_i$ for ease of notation. We need to update all the $c_i$'s for each MCMC draw. The MCMC state then consists of $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$. Let $F_Y(.)$ be the data likelihood and $n_c$ be the number if samples in cluster $c$. We use the following algorithm to update the clustering configuration.

- **New Cluster Creation:** For $i = 1, \ldots, n$. If $c_i$ is not a singleton (i.e. $c_i = c_j$ for some $j \neq i$) , let $c_i^*$ be the new cluster indicator. Draw $\phi_{c_i^*} = [A, R, S, \theta, Q]$ from the base prior $H_\phi$. Probability that a new cluster is created is

$$p(c_i = c_i^*) = min\{1, \frac{\alpha}{n-1}\frac{F_Y(Y_i, \phi_{c_i^*})}{F_Y(Y_i, \phi_{c_i})}\},$$

  otherwise, if $c_i$ is a singleton, draw $c_i^*$ from $c_{-i}$ with probability $Pr(c_i^* = c) = n_c/(n-1)$. The the probability of the sample belonging to the cluster $c$ is

$$p(c_i = c_i^*) = min\{1, \frac{n-1}{\alpha}\frac{F_Y(Y_i, \phi_{c_i^*})}{F_Y(Y_i, \phi_{c_i})}\},$$

- **Existing Clusters:** For $i = 1, \ldots, n$. If $c_i$ is not a singleton, choose a new value for $c_i$ using the following probabilities,

$$Pr(c_i = c) \propto \frac{n_c}{n-1} F_Y(Y_i, \phi_c),$$

- **Cluster Parameters:** Update $\phi_c$ for each cluster $c = 1, \ldots, d_n$ using $\phi_c | Y_c$ where $Y_c$ are the samples in the cluster $c$. The sampling procedure for updating the cluster parameters is similar to the posterior inference of the Bayesian lasso selection model.

### 1. Sampling from $H_\phi$

When a new cluster is formed we need to draw the new cluster parameters $\gamma_i = (\boldsymbol{\theta}_i, \boldsymbol{\Omega}_i^{-1})$ from the base prior $H_\phi$ which can accomplished as follows:

- The mean of the cluster $\boldsymbol{\theta}$ can be drawn from the distribution $p(\boldsymbol{\theta}|\boldsymbol{\Omega})$ which is a multivariate normal distribution.
- Drawing $\boldsymbol{\Omega}$ is complicated as we do not have a closed form distribution to draw $\boldsymbol{\Omega}$ from. As $\boldsymbol{\Omega} = \{\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S}, \boldsymbol{Q}\}$, we need to draw each of these to get a draw for $\boldsymbol{\Omega}$.
- The probabilities $\boldsymbol{Q}$ can be sampled directly from the beta prior.
- $\boldsymbol{S}$ can be sampled directly from the Inverse Gamma prior specified.
- Sampling $\boldsymbol{A}, \boldsymbol{R}$ is complicated due to the condition of positive definiteness . We use a metropolis hastings algorithm to sample these variables because they do not have a closed form distribution.

## 2. Real data example

We use the leukemia data from Golub et al. (1999) as an case study to illustrate our Dirichlet process mixture model. In this study, the authors measured the gene expression signatures of human acute leukemia and included prediction of the type of leukemia using supervised learning and the discovery of new classes of leukemia using unsupervised learning. The motivation for this work was to improve cancer treatment by distinguishing between subclasses of cancers or tumors. The data is available from http://www.genome.wi.mit.edu/MPR. The data was first classified into two groups: (1) data from lymphoid precursors and (2) data from myeloid precursors. The first one is known as acute lymphoblastic leukemia (ALL) and the second one is known as acute myelogenous leukemia (AML). The data has 6817 genes and 72 patient samples. We selected the top 10 genes to do the analysis to cluster the data and found different graphs of relations between genes for different cancers. The top genes were selected using the Bayesian gene selection algorithm (Lee et al. (2003)).

We fit our Dirichlet processes model to this data. We ran the MCMC simulation

for 100000 samples and remove the first 20000 samples as burn-in. The Dirichlet process found two clusters with one cluster corresponding to ALL and the other one corresponded to AML. The networks corresponding to the two clusters ALL and AML are shown in Figure 23 and Figure 24, respectively. As can be seen the networks for these two clusters were quite different, which suggests possible interactions between genes is different depending on the sub-type of cancer. The biological conclusions based on the data are similar to the ones described above in the finite mixture model where the genes PSMA,TCF3,CCND3,CD79A and MYL6B play a major role in pathways related to ALL whereas CD33,LYN,ATP6V0C,SRGN and ZYX play a major role in AML pathways.

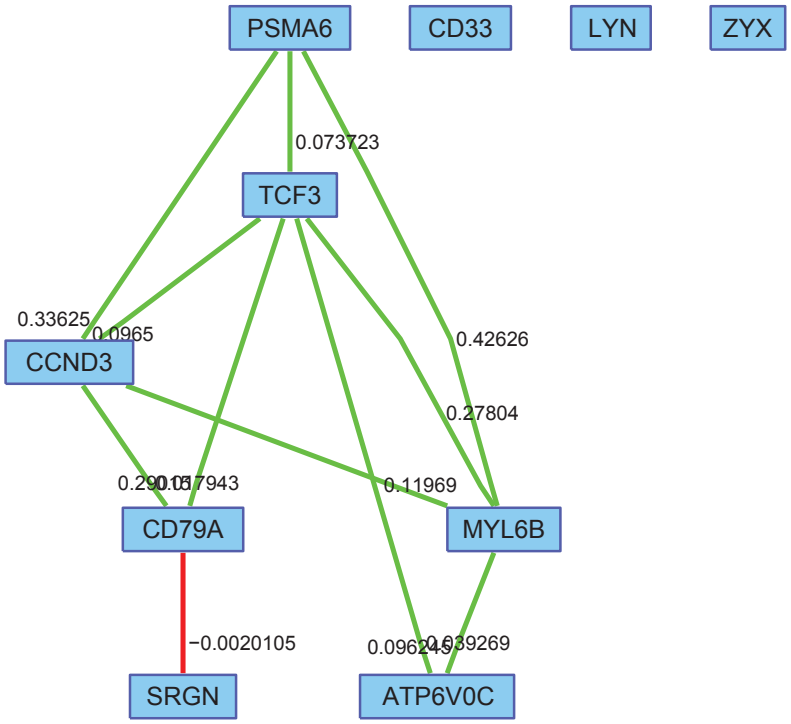



Fig. 23.: Graph for ALL Group.

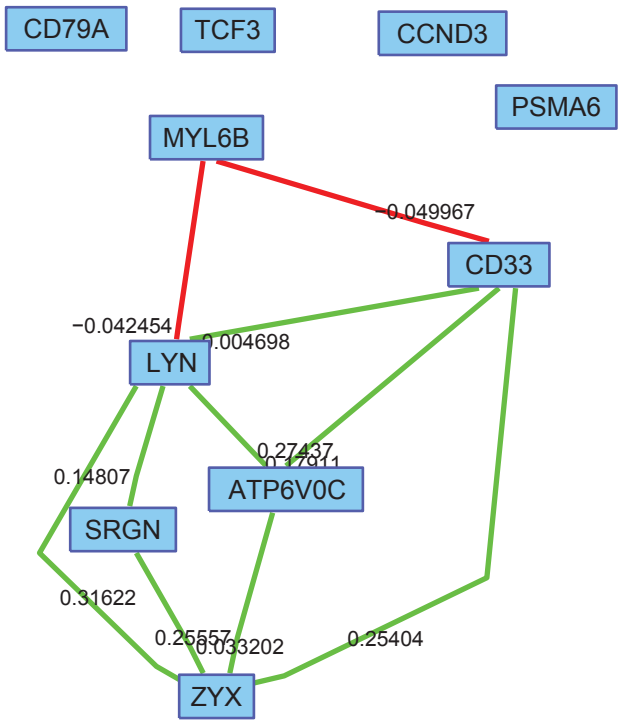



Fig. 24.: Graph for AML Group.

## 3. Simulations

We performed a posterior predictive simulation study to evaluate the operating characteristics of our methodology for Dirichlet process mixture of graphical models. We simulated data from our fitted model of the leukaemia data set using the estimated precision matrices for the two groups, ALL and AML. The simulation was conducted as follows. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Omega}}_j^{-1})$ denote the estimates of the mean and precision matrices corresponding to the ALL ($j = 1$) and AML ($j = 2$) groups, respectively, as obtained above. We generated data under the convolution of the following multivariate normal likelihood,

$$\boldsymbol{Y}_j \sim N(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Omega}}_j^{-1}),$$

with 100 samples and 10 covariates.

We compared the accuracy of the method using the K-L distance. The estimates of the K-L distances was 0.9584 for the ALL and 1.2689 for the AML cluster. In addition, we also compared the false positive and false negative rates for finding true edges For the AML cluster, we found false positive rates of 0.0222 and false negative rates of 0.0000 and for the ALL cluster, we found false positive rates of 0.0444 and false negative rates of 0 for the Dirichlet process model.

## C. Discussion and Conclusions

In this dissertation a Bayesian framework for adaptive estimation of precision matrices in Gaussian graphical models has been developed. We propose sparse estimators using L1-regularization and use lasso-based selection priors to obtain sparse and adaptively shrunk estimators of the precision matrix that conduct simultaneous model selection and estimation. We extend these methods to mixtures of Gaussian graphical models for clustered data, with each mixture component assumed to be Gaussian with an adaptive covariance structure. We discuss appropriate posterior simulation schemes for implementing posterior inference in the proposed models, including the evaluation of normalizing constants that are functions of the parameters of interest which result from constraints on the correlation matrix. We compare our methods with several existing methods from the literature using both real and simulated examples. We found our methods to be very competitive and in some cases to substantially outperform the existing methods.

Our simulations and analysis suggest that it is feasible to implement adaptive GGMs and mixtures of GGMs using MCMC for a reasonable number of variables. Applications to more high-dimensional settings may require more refined sampling algorithms and/or parallelized computations for our method to run in a reasonable time.

One nice feature of our modelling framework is that it can be generalized to other contexts in a straightforward manner. As opposed to the unsupervised setting we considered, another context would be that of supervised learning or classification using GGMs and showed that using GGM's improves the misclassification rate. Another interesting setting would be to extend our methods for situations in which the variables are observed over time and our models are used to develop time-dependent sparse dynamic graphs. We leave these tasks for future consideration.

# REFERENCES

Agarwal, R., Gonzalez-Angulo, A.-M., Myhre, S., Carey, M., Lee, J.-S., Overgaard, J., Alsner, J., Stemke-Hale, K., Lluch, A., Neve, R. M., Kuo, W. L., Sorlie, T., Sahin, A., Valero, V., Keyomarsi, K., Gray, J. W., Borresen-Dale, A.-L., Mills, G. B., and Hennessy, B. T. (2009), "Integrative Analysis of Cyclin Protein Levels Identifies Cyclin B1 as A Classifier and Predictor of Outcomes in Breast Cancer," *Clinical Cancer Research*, 15, 3654–3662.

Amit, I., Citri, A., Shay, T., Lu, Y., Katz, M., Zhang, F., Tarcic, G., Siwak, D., Lahad, J., Jacob-Hirsch, J., et al. (2007), "A Module of Negative Feedback Regulators Defines Growth Factor Signaling," *Nature Genetics*, 39, 503–512.

Bae, K., and Mallick, B. (2004), "Gene Selection Using A Two-level Hierarchical Bayesian model," *Bioinformatics*, 20, 3423–3430.

Barnard, J., McCulloch, R., and Meng, X. (2000), "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage," *Statistica Sinica*, 10, 1281–1312.

Bast, R., Hennessy, B., and Mills, G. (2009), "The Biology of Ovarian Cancer: New Opportunities for Translation," *Nature Reviews Cancer*, 9, 415–428.

Bickel, P., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

Blower, P., Verducci, J., Lin, S., Zhou, J., Chung, J., Dai, Z., Liu, C., Reinhold, W., Lorenzi, P., Kaldjian, E., et al. (2007), "MicroRNA Expression Profiles for the NCI-60 Cancer Cell Panel," *Molecular Cancer Therapeutics*, 6, 1483–1491.

Botev, Z., Grotowski, J., and Kroese, D. (2010), "Kernel Density Estimation Via Diffusion," *The Annals of Statistics*, 38, 2916–2957.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

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 (Statistical Methodology)*, 65, 3–39.

Cancer Genome Atlas Research Network (2008), "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways," *Nature*, 455, 1061–1068.

Carvalho, C., and West, M. (2007), "Dynamic Matrix-variate Graphical Models," *Bayesian Analysis*, 2, 69–98.

Carvalho, C. M., Massam, H., and West, M. (2007), "Simulation of Hyper-inverse Wishart Distributions in Graphical Models," *Biometrika*, 94, 647–659.

Carvalho, C. M., and Scott, J. G. (2009), "Objective Bayesian Model Selection in Gaussian Graphical Models," *Biometrika*, 96, 497–512.

Chaudhuri, S., Drton, M., and Richardson, T. (2007), "Estimation of a Covariance Matrix with Zeros," *Biometrika*, 94, 199–216.

Courtney, K., Corcoran, R., and Engelman, J. (2010), "The PI3K Pathway as Drug Target in Human Cancer," *Journal of Clinical Oncology*, 28, 1075–1083.

Cox, D. R., and Wermuth, N. (2002), "On Some Models for Multivariate Binary Variables Parallel in Complexity with the Multivariate Gaussian Distribution," *Biometrika*, 89, 462–469.

Davies, M., Hennessy, B., and Mills, G. B. (2006), "Point Mutations of Protein Kinases and Individualised Cancer Therapy," *Expert Opinion on Pharmacotherapy*, 7, 2243–2261.

Davies, M., Koul, D., Dhesi, H., Berman, R., McDonnell, T., McConkey, D., Yung, W., and Steck, P. (1999), "Regulation of Akt/PKB Activity, Cellular Growth, and Apoptosis in Prostate Carcinoma Cells by MMAC/PTEN," *Cancer Research*, 59, 2551–2556.

Davies, M., Lu, Y., Sano, T., Fang, X., Tang, P., LaPushin, R., Koul, D., Bookstein, R., Stokoe, D., Yung, W., Mills, G., and Steck, P. (1998), "Adenoviral Transgene Expression of MMAC/PTEN in Human Glioma Cells Inhibits Akt Activation and Induces Anoikis," *Cancer Research*, 58, 5285–5290.

Davies, M. A., Stemke-Hale, K., Lin, E., Tellez, C., Deng, W., Gopal, Y. N., Woodman, S. E., Calderone, T. C., Ju, Z., Lazar, A. J., Prieto, V. G., Aldape, K., Mills, G. B., and Gershenwald, J. E. (2009), "Integrated Molecular and Clinical Analysis of AKT Activation in Metastatic Melanoma," *Clinical Cancer Research*, 15, 7538–7546.

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 21, 1272–1317.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000), *Bayesian Variable Selection Using the Gibbs Sampler. Generalized Linear Models: A Bayesian Perspective*, eds: D.Dey, S.Ghosh and B.Mallick, New York: Marcel Dekker.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing*, 12, 27–36.

Dempster, A. (1969), *Elements of Continuous Multivariate Analysis*, New York: Addison Wesley.

Dempster, A. P. (1972), "Covariance Selection," *Biometrics*, 28, 157–175.

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.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000), *Pattern Classification (2nd Edition)*, New York: Wiley-Interscience.

Ehrich, M., Turner, J., Gibbs, P., Lipton, L., Giovanneti, M., Cantor, C., and van den Boom, D. (2008), "Cytosine Methylation Profiling of Cancer Cell Lines," *Proceedings of the National Academy of Sciences*, 105, 4844–4849.

Figueiredo, M. A. T. (2003), "Adaptive Sparseness for Supervised Learning," *IEEE Transactions on Pattern Analysis Machine Intelligence*, 25, 1150–1159.

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Fraley, C., and Raftery, A. E. (2007), "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering," *Journal of Classification*, 24, 155–181.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441.

Galbraith, S. J., Tran, L. M., and Liao, J. C. (2006), "Transcriptome Network Component Analysis with Limited Microarray Data," *Bioinformatics*, 22, 1886–1894.

Gaur, A., Jewell, D., Liang, Y., Ridzon, D., Moore, J., Chen, C., Ambros, V., and Israel, M. (2007), "Characterization of MicroRNA Expression Levels and Their Biological Correlates in Human Cancer Cell Lines," *Cancer Research*, 67, 2456.

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 501–514.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

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

Giudici, P., (1996), *Learning in Graphical Gaussian Models*, London: Oxford University Press.

Giudici, P. and Green, P. J. (1999), "Decomposable Graphical Gaussian Model Determination," *Biometrika*, 86, 785–801.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Griffin, J., and Brown, P. (2007), "Bayesian Adaptive Lassos with Non-convex Penalization," *Centre for Research in Statistical Methodology, University of Warwick, Coventry, UK, Technical Report*, 07–2.

Halaban, R., Zhang, W., Bacchiocchi, A., Cheng, E., Parisi, F., Ariyan, S., Krauthammer, M., McCusker, J., Kluger, Y., and Sznol, M. (2010), " PLX4032, a Selective BRAF V600E Kinase Inhibitor, Activates the ERK Pathway and Enhances Cell Migration and Proliferation of BRAF WT Melanoma Cells," *Pigment Cell & Melanoma Research*, 23, 190–200.

Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002), "Co-clustering of Biological Networks and Gene Expression Data," *Bioinformatics*, 18, S145–154.

Hennessy, B., Lu, Y., Poradosu, E., Yu, Q., Yu, S., Hall, H., Carey, M., Ravoori, M., Gonzalez-Angulo, A., Birch, R., et al. (2007), "Pharmacodynamic markers of perifosine efficacy," *Clinical Cancer Research*, 13, 7421–7431.

Hennessy, B., Murph, M., Nanjundan, M., Carey, M., Auersperg, N., Almeida, J., Coombes, K., Liu, J., Lu, Y., Gray, J., and Mills, G. (2008), "Ovarian Cancer: Linking Genomics to New Target Discovery and Molecular Markers - The Way Ahead," *Advances in Experimental Medicine and Biology*, 617, 23–40.

Hennessy, B., Smith, D., Ram, P., Lu, Y., and Mills, G. (2005), "Exploiting the PI3K/AKT Pathway for Cancer Drug Discovery," *Nature Reviews Drug Discovery*, 4, 988–1004.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

Hu, J., He, X., Baggerly, K. A., Coombes, K. R., Hennessy, B. T., and Mills, G. B. (2007), "Non-parametric Quantification of Protein Lysate Arrays," *Bioinformatics*, 23, 1986–1994.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., et al. (2008), "Cancer statistics, 2008." *CA: A Cancer Journal for Clinicians*, 58, 71–96.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2004), "Experiments in Stochastic Computation for High-Dimensional Graphical Models," *Statistical Science*, 20, 388–400.

Kelley, R., and Ideker, T. (2005), "Systematic Interpretation of Genetic Interactions Using Protein Networks," *Nat Biotech*, 23, 561–566.

Kuo, L., and Mallick, B. (1998), "Variable Selection for Regression Models," *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 65–81.

Lauritzen, S. (1996), *Graphical Models*, vol. 17, New York: Oxford University Press.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003), "Gene selection: A Bayesian Variable Selection Approach," *Bioinformatics*, 19, 90–97.

Liechty, J. C., Liechty, M. W., and Muller, P. (2004), "Bayesian Correlation Estimation," *Biometrika*, 91, 1–14.

Linardou, H., Dahabreh, I. J., Kanaloupiti, D., Siannis, F., Bafaloukos, D., Kosmidis, P., Papadimitriou, C. A., and Murray, S. (2008), "Assessment of Somatic k-RAS Mutations as a Mechanism Associated with Resistance to EGFR-targeted Agents: A Systematic Review and Meta-analysis of Studies in Advanced Non-small-cell Lung Cancer and Metastatic Colorectal Cancer," *The Lancet Oncology*, 9, 962 – 972.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Meinshausen, N. and Búhlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462.

Mills, G., Kohn, E., Lu, Y., Eder, A., Fang, X., Wang, H., Bast, R., Gray, J., Jaffe, R., and Hortobagyi, G. (2003), "Linking Molecular Diagnostics to Molecular Therapeutics: Targeting the PI3K Pathway in Breast Cancer," *Seminars in Oncology*, 30, 93–104.

Mircean, C., Shmulevich, I., Cogdell, D., Choi, W., Jia, Y., Tabus, I., Hamilton, S. R., and Zhang, W. (2005), "Robust Estimation of Protein Expression Ratios with Lysate Microarray Technology," *Bioinformatics*, 21, 1935–1942.

Mirzoeva, O. K., Das, D., Heiser, L. M., Bhattacharya, S., Siwak, D., Gendelman, R., Bayani, N., Wang, N. J., Neve, R. M., Guan, Y., Hu, Z., Knight, Z., Feiler, H. S., Gascard, P., Parvin, B., Spellman, P. T., Shokat, K. M., Wyrobek, A. J., Bissell, M. J., McCormick, F., Kuo, W.-L., Mills, G. B., Gray, J. W., and Korn, W. M. (2009), "Basal Subtype and MAPK/ERK Kinase (MEK)-Phosphoinositide 3-Kinase Feedback Signaling Determine Susceptibility of Breast Cancer Cells to MEK Inhibition," *Cancer Research*, 69, 565–572.

Monni, S., and Li, H. (2010), "Bayesian Methods for Network-Structured Genomics Data," *UPenn Biostatistics Working Papers*, 34.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008), "Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-based Functional Mixed Models." *Biometrics*, 64, 479 – 489.

Muller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal Sample Size for Multiple Testing," *Journal of the American Statistical Association*, 99, 990–1001.

Nagata, Y., Lan, K.-H., Zhou, X., Tan, M., Esteva, F. J., Sahin, A. A., Klos, K. S., Li, P., Monia, B. P., Nguyen, N. T., Hortobagyi, G. N., Hung, M.-C., and Yu, D. (2004), "PTEN Activation Contributes to Tumor Inhibition by Trastuzumab, and Loss of PTEN Predicts Trastuzumab Resistance in Patients," *Cancer Cell*, 6, 117 – 127.

Neeley, E. S., Kornblau, S. M., Coombes, K. R., and Baggerly, K. A. (2009), "Variable Slope Normalization of Reverse Phase Protein Arrays," *Bioinformatics*, 25, 1384–1389.

Neve, R., Chin, K., Fridlyand, J., Yeh, J., Baehner, F., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P., DeVries, S., Lapuk, A., Wang, N., Kuo, W.-L., Stilwell, J., Pinkel, D., Albertson, D., Waldman, F., McCormick, F., Dickson, R.,

Johnson, M., Lippman, M., Ethier, S., Gazdar, A., and Gray, J. (2006), "A Collection of Breast Cancer Cell Lines for the Study of Functionally Distinct Cancer Subtypes," *Cancer Cell*, 10, 515–527, cited By (since 1996) 261.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176.

Newton, M. A. and Raftery, A. E. (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, pp. 3–48.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W. C., Waltham, M., Kouros-Mehr, H., Bussey, K. J., Lee, J. K., Espina, V., Munson, P. J., Petricoin, E., Liotta, L. A., and Weinstein, J. N. (2003), "Proteomic Profiling of the NCI-60 Cancer Cell Lines Using New High-density Reverse-phase Lysate Microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 14229–14234.

O'Reilly, K. E., Warycha, M., Davies, M. A., Rodrik, V., Zhou, X. K., Yee, H., Polsky, D., Pavlick, A. C., Rosen, N., Bhardwaj, N., Mills, G., and Osman, I. (2009), "Phosphorylated 4E-BP1 Is Associated with Poor Survival in Melanoma," *Clinical Cancer Research*, 15, 2872–2878.

Park, E. S., Rabinovsky, R., Carey, M., Hennessy, B. T., Agarwal, R., Liu, W., Ju, Z., Deng, W., Lu, Y., Woo, H. G., Kim, S.-B., Cheong, J.-H., Garraway, L. A., Weinstein, J. N., Mills, G. B., Lee, J.-S., and Davies, M. A. (2010), "Integrative Analysis of Proteomic Signatures, Mutations, and Drug Responsiveness in the NCI 60 Cancer Cell Line Set," *Molecular Cancer Therapeutics*, 9, 257–267.

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

Paweletz, C., Charboneau, L., Bichsel, V., Simone, N., Chen, T., Gillespie, J., Emmert-Buck, M., Roth, M., Petricoin, E., and Liotta, L. (2001), "Reverse Phase Protein Microarrays which Capture Disease Progression Show Activation of Pro-survival Pathways at the Cancer Invasion Front," *ONCOGENE*, 20, 1981–1989.

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.

Rahnenfhrer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004), "Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data," *Statistical Applications in Genetics and Molecular Biology*, 3, 16.

Rao, C. R. (1948), "Tests of Significance in Multivariate Analysis," *Biometrika*, 35, 58–79.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007), "Classification of Microarray Data Using Gene Networks," *BMC Bioinformatics*, 8, 1–15.

Roverato, A. (2000), "Cholesky Decomposition of a Hyper Inverse Wishart Matrix," *Biometrika*, 87, 99–112.

Roverato, A. (2002), "Hyper Inverse Wishart Distribution for Non-Decomposable Graphs and Its Application to Bayesian Inference for Gaussian Graphical Models," *Scandinavian Journal of Statistics*, 29, 391–411.

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.

Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N. (2007), "Transcript and Protein Expression Profiles of the NCI-60 Cancer Cell Panel: An Integromic Microarray Study," *Molecular Cancer Therapeutics*, 6, 820–832.

Sheehan, K. M., Calvert, V. S., Kay, E. W., Lu, Y., Fishman, D., Espina, V., Aquino, J., Speer, R., Araujo, R., Mills, G. B., Liotta, L. A., Petricoin, E. F., and Wulfkuhle, J. D. (2005), "Use of Reverse Phase Protein Microarrays and Reference Standard Development for Molecular Network Analysis of Metastatic Ovarian Carcinoma," *Molecular & Cellular Proteomics*, 4, 346–355.

Siena, S., Sartore-Bianchi, A., Di Nicolantonio, F., Balfour, J., and Bardelli, A. (2009), "Biomarkers Predicting Clinical Outcome of Epidermal Growth Factor Receptor-Targeted Therapy in Metastatic Colorectal Cancer," *Journal of the National Cancer Institute*, 101, 1308–1324.

Sivachenko, A., Yuryev, A., Daraselia, N., and Mazo, I. (2002), "Identifying Local Gene Expression Patterns in Biomolecular Networks," *Bioinformatics*, 18, S145–54.

Speed, T. P. and Kiiveri, H. T. (1986), "Gaussian Markov Distributions over Finite Graphs," *The Annals of Statistics*, 14, 138–150.

Stemke-Hale, K., Gonzalez-Angulo, A. M., Lluch, A., Neve, R. M., Kuo, W.-L., Davies, M., Carey, M., Hu, Z., Guan, Y., Sahin, A., Symmans, W. F., Pusztai, L., Nolden, L. K., Horlings, H., Berns, K., Hung, M.-C., van de Vijver, M. J., Valero, V., Gray, J. W., Bernards, R., Mills, G. B., and Hennessy, B. T. (2008), "An Integrative Genomic and Proteomic Analysis of PIK3CA, PTEN, and AKT Mutations in Breast Cancer," *Cancer Research*, 68, 6084–6091.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 9440–9445.

Tabus, I., Hategan, A., Mircean, C., Rissanen, J., Shmulevich, I., Zhang, W., and Astola, J. (2006), "Nonlinear Modeling of Protein Expressions in Protein Arrays," *IEEE Transactions on Signal Processing*, 54, 2394–2407.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2006), "Reverse Phase Protein Array: Validation of a Novel Proteomic Technology and Utility for Analysis of Primary Leukemia Specimens and Hematopoietic Stem Cells," *Molecular Cancer Therapeutics*, 5, 2512–2521.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Varambally, S., Yu, J., Laxman, B., Rhodes, D. R., Mehra, R., Tomlins, S. A., Shah, R. B., Chandran, U., Monzon, F. A., Becich, M. J., Wei, J. T., Pienta, K. J., Ghosh, D., Rubin, M. A., and Chinnaiyan, A. M. (2005), "Integrative Genomic and Proteomic Analysis of Prostate Cancer Reveals Signatures of Metastatic Progression," *Cancer Cell*, 8, 393 – 406.

Vasudevan, K. M., Barbie, D. A., Davies, M. A., Rabinovsky, R., McNear, C. J., Kim, J. J., Hennessy, B. T., Tseng, H., Pochanard, P., Kim, S. Y., Dunn, I. F., Schinzel, A. C., Sandy, P., Hoersch, S., Sheng, Q., Gupta, P. B., Boehm, J. S., Reiling, J. H., Silver, S., Lu, Y., Stemke-Hale, K., Dutta, B., Joy, C., Sahin, A. A., Gonzalez-Angulo, A. M., Lluch, A., Rameh, L. E., Jacks, T., Root, D. E., Lander, E. S., Mills, G. B., Hahn, W. C., Sellers, W. R., and Garraway, L. A. (2009), "AKT-Independent Signaling Downstream of Oncogenic PIK3CA Mutations in Human Cancer," *Cancer Cell*, 16, 21–32.

Vert, J. P., and Kanehisa, M. (2003), "Extracting Active Pathways from Gene Expression Data," *Bioinformatics*, 19, 238–244.

Vivanco, I., and Sawyers, C. L. (2002), "The Phosphatidylinositol 3-Kinase AKT Pathway in Human Cancer." *Nature Reviews. Cancer*, 2, 489 – 501.

Whittaker, J. (1990), *Graphical models in applied multivariate statistics*, New York: Wiley.

Wong, F., Carter, C. K., and Kohn, R. (2003), "Efficient Estimation of Covariance Selection Models," *Biometrika*, 90, 809–830.

Yang, R., and Berger, J. O. (1994), "Estimation of a Covariance Matrix Using the Reference Prior," *The Annals of Statistics*, 22, 1195–1211.

Yuan, M., and Lin, Y. (2005), "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models," *Journal of the American Statistical Association*, 100, 1215–1225.

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35.

Yuan, T., and Cantley, L. (2008), "PI3K Pathway Alterations in Cancer: Variations on a Theme," *Oncogene*, 27, 5497–5510.

Zhang, L., Wei, Q., Mao, L., Liu, W., Mills, G. B., and Coombes, K. (2009), "Serial Dilution Curve: A New Method for Analysis of Reverse Phase Protein Array Data," *Bioinformatics*, 25, 650–654.

# APPENDIX A

## CHECKING FOR POSITIVE DEFINITENESS

The drawing of a particular $R_{ij}(i > j)$ given the other correlations and $S$ (as well as whatever other parameters are in the model) is complicated by the requirement that $\boldsymbol{R}$ be positive definite. We need to know what values of $R_{ij}$ keep $\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{R}$ positive definite given that the other correlations are fixed. It should be noted that $\boldsymbol{R}$ and $\boldsymbol{C}$ are equivalent in the MCMC sampling as $R_{ij} = 0$ when $A_{ij} = 0$. We follow the approach of Barnard et al. (2000), as shown below.

Start with a correlation matrix $\boldsymbol{R}$, which is positive definite. Assume $\boldsymbol{R}(r)$ as the matrix obtained by replacing $i, j^{th}$ element of $\boldsymbol{R}$ by $r$ and let $f(r) = |\boldsymbol{R}(r)|$ which is the determinant of $\boldsymbol{R}$. $f(r) > 0$ is a necessary and sufficient condition for $\boldsymbol{R}(r)$ to be positive definite. The determinant of $\boldsymbol{R}$ is a quadratic function in $r$ which is $f(r) = ar^2 + br + c$. The coefficients $a, b$ and $c$ can be calculated from the value of the determinant for different values of $r$. By finding the range of r in which the matrix is positive definite we continue to keep the correlation matrix positive definite in subsequents iterations of the MCMC.

# APPENDIX B

## COMPUTING BIC VALUES FOR THE GRAPHS

The Bayesian information criterion (BIC) is widely used for model selection problems. BIC penalizes the complex models in favor of balanced models. BIC can be computed as

$$-2\log p(Y|\mathcal{G}) + const \approx -2L(Y,\hat{\theta}) + m_{\mathcal{G}} log(n) \equiv BIC,$$

where $p(Y|\mathcal{G})$ is the likelihood of the data for the model $\mathcal{G}$, $L(Y,\hat{\theta})$ is the maximized log likelihood for the model, $m_{\mathcal{G}}$ is the number of independent parameters to be estimated in the model, and $n$ is the number of samples. Given any two estimated models, $\mathcal{G}_1$ and $\mathcal{G}_2$, the model with the lower value of BIC is the preferred model. The number of parameters to be estimated in the model is considered to be the number of non-zero edges and all the other parameters in the model. In the finite mixture model the number of clusters is not considered an independent parameter for the purpose of computing the BIC. If each model is equally likely *a priori*, then $p(Y|\mathcal{G})$ is proportional to the posterior probability that the data conform to the model $\mathcal{G}$.

# VITA

Rajesh Talluri was born in Guntur, India. He graduated from the Indian Institute of Technology (IIT) at Guwahati, India in May 2005 with a Bachelor of Technology (B.Tech.) in electronics and communications engineering. He received his Ph.D. in statistics from Texas A&M University in August of 2011 under the direction of Dr. Bani K. Mallick and Dr. Veerabhadran Baladandayuthapani.

Address: Dept of Statistics
Texas A&M University
3143 TAMU
College Station, TX 77840

Email Address: rajeshstat@gmail.com

The typist for this dissertation is Rajesh Talluri