HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2015 June 13.

Published in final edited form as:

J Am Stat Assoc. 2015 March 1; 110(509): 159–174. doi:10.1080/01621459.2014.896806.

Bayesian Inference of Multiple Gaussian Graphical Models

Christine B. Peterson[*], **Francesco C. Stingo**[†], and **Marina Vannucci**[‡]

[*]Department of Health Research and Policy, Stanford University

[†]Department of Biostatistics, The University of Texas MD Anderson Cancer Center

[‡]Department of Statistics, Rice University

Abstract

In this paper, we propose a Bayesian approach to inference on multiple Gaussian graphical models. Specifically, we address the problem of inferring multiple undirected networks in situations where some of the networks may be unrelated, while others share common features. We link the estimation of the graph structures via a Markov random field (MRF) prior which encourages common edges. We learn which sample groups have a shared graph structure by placing a spike-and-slab prior on the parameters that measure network relatedness. This approach allows us to share information between sample groups, when appropriate, as well as to obtain a measure of relative network similarity across groups. Our modeling framework incorporates relevant prior knowledge through an edge-specific informative prior and can encourage similarity to an established network. Through simulations, we demonstrate the utility of our method in summarizing relative network similarity and compare its performance against related methods. We find improved accuracy of network estimation, particularly when the sample sizes within each subgroup are moderate. We also illustrate the application of our model to infer protein networks for various cancer subtypes and under different experimental conditions.

Keywords

Gaussian graphical model; Markov random field; Bayesian inference; *G*-Wishart prior; protein network

1 Introduction

Graphical models, which describe the conditional dependence relationships among random variables, have been widely applied in genomics and proteomics to infer various types of networks, including co-expression, gene regulatory, and protein interaction networks (Friedman, 2004; Dobra et al., 2004; Mukherjee and Speed, 2008; Stingo et al., 2010; Telesca et al., 2012). Here we address the problem of inferring multiple undirected networks in situations where some networks may be unrelated, while others may have a similar structure. This problem relates to applications where we observe data collected under various conditions. In such situations, using the pooled data as the basis for inference of a single network may lead to the identification of spurious relationships, while performing inference separately for each group effectively reduces the sample size. Instead, we propose a joint inference method that infers a separate graphical model for each group but allows for shared structures, when supported by the data. Our approach not only allows estimation of a

graphical model for each sample group, but also provides insights on how strongly the graph structures for any two sample groups are related.

Some approaches for inferring graphical models for two or more sample groups have been proposed in recent years. Guo et al. (2011) extend the graphical lasso to multiple undirected graphs by expressing the elements of the precision matrix for each group as a product of common and group-specific factors. In their optimization criterion, they incorporate an ℓ_1 penalty on the common factors, to create a sparse shared structure, and a second ℓ_1 penalty on the group-specific factors, to allow edges included in the shared structure to be set to zero for specific groups. Danaher et al. (2013) propose a more general framework that uses convex penalties and explore in detail the properties of two specific penalty structures: the fused graphical lasso, which encourages both shared structure and shared edge values, and the group graphical lasso, which results in shared graph structures but not shared edge values. As for Bayesian approaches, Yajima et al. (2012) propose a Bayesian method to estimate Gaussian directed graphs for related samples. Focusing mainly on the case of two sample groups, the authors treat one group as the baseline and express the strength of association between two variables in the differential group as the sum of the strength in the baseline group plus a differential parameter.

In this paper, we formulate an alternative Bayesian approach to the problem of multiple network inference. We link estimation of the graph structures via a Markov random field (MRF) prior which encourages common structures. This prior favors the inclusion of an edge in the graph for a particular group if the same edge is included in the graphs of related sample groups. Unlike the approaches mentioned above, we do not assume that all subgroups are related. Instead, we learn which sample groups have a shared graph structure by placing a spike-and-slab prior on parameters that measure network relatedness. The posterior probabilities of inclusion for these parameters summarize the networks' similarity. This formulation allows us to share information between sample groups only when appropriate. Our framework also allows for the incorporation of relevant prior knowledge through an edge-specific informative prior. This approach enables borrowing of strength across related sample groups and can encourage similarity to an established network. Through simulations, we demonstrate the utility of our method in summarizing relative network similarity and compare its performance against related methods. We find improved accuracy of network estimation, particularly when the sample sizes within each subgroup are moderate. We also illustrate the application of our model to infer protein networks for various cancer subtypes and under different experimental conditions. In such applications, a measure of network similarity helps determine if treatments that are successful for one subtype are likely to be effective in another, while the differential edges between networks highlight potential targets for treatments specific to each group.

The rest of the paper is organized as follows. Section 2 below provides background on graphical models and on Bayesian methods for estimation. Section 3 presents the model and the construction of the priors. Section 4 addresses posterior inference, including the Markov chain Monte Carlo method. Section 5 includes the simulations and Section 6 demonstrates the application of our method on two case studies on protein networks. Section 7 concludes the paper.

2 Background

2.1 Graphical models

Graphical models use a graph G to represent conditional dependence relationships among random variables. A graph $G = (V, E)$ specifies a set of vertices $V = \{1, 2, \dots, p\}$ and a set of edges $E \subset V \times V$. In a directed graph, edges are denoted by ordered pairs $(i, j) \in E$. In an undirected graph, $(i, j) \in E$ if and only if $(j, i) \in E$. For an overview of graphical models in statistics, see Lauritzen (1996). We focus here on undirected graphical models, also known as Markov random fields. In this class of models, each vertex in the graph G corresponds to a random variable. The absence of an edge between two vertices means that the two corresponding variables are conditionally independent given the remaining variables, while an edge is included whenever the two variables are conditionally dependent.

In Gaussian graphical models (GGMs), also known as covariance selection models (Dempster, 1972), the conditional independence relationships correspond to constraints on the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ of the multivariate normal distribution

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Omega}^{-1}), \quad i = 1, \dots, n, \quad (2.1)$$

with $\mathbf{x}_i \in \mathbb{R}^p$ the vector of observed data for subject i, $\boldsymbol{\mu} \in \mathbb{R}^p$ the mean vector, and $\mathbf{\Omega} \in \mathbb{R}^p \times \mathbb{R}^p$ a positive definite symmetric matrix. The multivariate normal is parametrized here in terms of the precision matrix $\mathbf{\Omega}$ rather than the covariance matrix $\mathbf{\Sigma}$ since there is a correspondence between the conditional dependence graph G and the structure of $\mathbf{\Omega}$. Specifically, the precision matrix $\mathbf{\Omega}$ is constrained to the cone of symmetric positive definite matrices with off-diagonal entry ω_{ij} equal to zero if there is no edge in G between vertex i and vertex j.

Many of the estimation techniques for GGMs rely on the assumption of sparsity in the precision matrix, which is a realistic assumption for many real-world applications including inference of biological networks. Regularization methods are a natural approach to inference of a sparse precision matrix. The most popular of these is the graphical lasso (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008), which uses an ℓ_1 penalty on the off-diagonal entries of the precision matrix to achieve sparsity in estimation of the graph structure. Among Bayesian approaches, the Bayesian graphical lasso, proposed as the Bayesian analogue to the graphical lasso, places double exponential priors on the off-diagonal entries of the precision matrix (Wang, 2012; Peterson et al., 2013). Estimation of a sparse graph structure using the Bayesian graphical lasso is not straightforward, however, since the precision matrices sampled from the posterior distribution do not contain exact zeros.

2.2 *G*-Wishart prior framework

Bayesian approaches to graphical models which enforce exact zeros in the precision matrix have been proposed by Roverato (2002), Jones et al. (2005), and Dobra et al. (2011). In Bayesian analysis of multivariate normal data, the standard conjugate prior for the precision matrix $\mathbf{\Omega}$ is the Wishart distribution. Equivalently, one can specify that the covariance matrix

$\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ follows the Inverse-Wishart distribution. Early work (Dawid and Lauritzen, 1993; Giudici and Green, 1999) focused on restrictions of the Inverse-Wishart to decomposable graphs, which have the special property that all prime components are complete. The assumption of decomposability greatly simplifies computation, but is artificially restrictive for the inference of real world networks. To address this limitation, Roverato (2002) proposed the *G*-Wishart prior as the conjugate prior for arbitrary graphs. The *G*-Wishart is the Wishart distribution restricted to the space of precision matrices with zeros specified by a graph *G* which may be either decomposable or non-decomposable. The *G*-Wishart density $W_G(b, D)$ can be written as

$$p(\mathbf{\Omega}|G, b, D) = I_G(b, D)^{-1}|\mathbf{\Omega}|^{(b-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{\Omega}D)\right\}, \quad \mathbf{\Omega} \in P_G$$

where $b > 2$ is the degrees of freedom parameter, D is a $p \times p$ positive definite symmetric matrix, I_G is the normalizing constant, and P_G is the set of all $p \times p$ positive definite symmetric matrices with $\omega_{ij} = 0$ if and only if $(i, j) \notin E$. Although this formulation is more flexible for modeling, it introduces computational difficulties because both the prior and the posterior normalizing constants are intractable. Jones et al. (2005) and Lenkoski and Dobra (2011) simplify the problem by integrating out the precision matrix. Dobra et al. (2011) propose a reversible jump algorithm to sample over the joint space of graphs and precision matrices that does not scale well to large graphs. Wang and Li (2012) propose a sampler which does not require proposal tuning and circumvents computation of the prior normalizing constant through the use of the exchange algorithm, improving both the accuracy and efficiency of computation.

3 Proposed model

Our goal is to infer a graph structure and obtain an estimate of the precision matrix describing the relationships among variables within each of *K* possibly related sample groups. These networks are complex systems and may be difficult to infer using separate estimation procedures when the sample size for any of the subgroups is small. Our approach addresses this issue by allowing the incorporation of relevant prior knowledge and the sharing of information across subgroups, when appropriate. In addition, our method allows comparison of the relative network similarity across the groups, providing a pairwise assessment of graph relatedness.

3.1 Likelihood

We let $\mathbf{X}_k$ represent the $n_k \times p$ matrix of observed data for sample group k, where $k = 1, 2, \ldots, K$. We assume that the same p random variables are measured across all groups, but allow the sample sizes n_k to differ. Assuming that the samples are independent and identically distributed within each group, the likelihood of the data for subject i in group k can be written as

$$\mathbf{x}_{k,i} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}), \quad i = 1, \ldots, n_k, \quad (3.1)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the mean vector for the kth group, and the precision matrix for the kth group $\boldsymbol{\Omega}_k$ is a symmetric positive definite matrix constrained by a graph G_k specific to that group. The graph G_k for sample group k can be represented as a symmetric binary matrix where the off-diagonal entry $g_{k,ij}$ indicates the inclusion of edge (i, j) in G_k. The inclusion of edge (i, j) in graphs 1, … , K is represented by the binary vector $\mathbf{g}_{ij} = (g_{1,ij}, \ldots, g_{K,ij})^T$.

3.2 Markov random field prior linking graphs

We define a Markov random field (MRF) prior on the graph structures that encourages the selection of the same edges in related graphs. This prior does not require the assumption of Gaussianity, and it is sufficiently general that it could be applied to models using any type of undirected graph.

MRF priors have previously been used to model the relationships among covariates in the context of Bayesian variable selection (Li and Zhang, 2010; Stingo and Vannucci, 2011). Our MRF prior follows a similar structure, but replaces indicators of variable inclusion with indicators of edge inclusion. The probability of the binary vector of edge inclusion indicators $\mathbf{g}_{ij}$, where $1 \le i < j \le p$, is given by

$$p(\mathbf{g}_{ij} | \nu_{ij}, \boldsymbol{\Theta}) = C(\nu_{ij}, \boldsymbol{\Theta})^{-1} \exp(\nu_{ij} \mathbf{1}^T \mathbf{g}_{ij} + \mathbf{g}_{ij}^T \boldsymbol{\Theta} \mathbf{g}_{ij}), \quad (3.2)$$

where $\mathbf{1}$ is the unit vector of dimension K, ν_{ij} is a parameter specific to each set of edges $\mathbf{g}_{ij}$, and $\boldsymbol{\Theta}$ is a $K \times K$ symmetric matrix representing the pairwise relatedness of the graphs for each sample group. The diagonal entries of $\boldsymbol{\Theta}$ are set to zero, and the off-diagonal entries which are nonzero represent connections between related networks. To help visualize the model formulation, Figure 1 shows a supergraph $\boldsymbol{\Theta}$ for three sample groups.

The normalizing constant in equation (3.2) is defined as

$$C(\nu_{ij}, \boldsymbol{\Theta}) = \sum_{\mathbf{g}_{ij} \in \{0,1\}^K} \exp(\nu_{ij} \mathbf{1}^T \mathbf{g}_{ij} + \mathbf{g}_{ij}^T \boldsymbol{\Theta} \mathbf{g}_{ij}). \quad (3.3)$$

From equation (3.2), we can see that the prior probability that edge (i, j) is absent from all K graphs simultaneously is

$$p(\mathbf{g}_{ij} = \mathbf{0} | \nu_{ij}, \boldsymbol{\Theta}) = \frac{1}{C(\nu_{ij}, \boldsymbol{\Theta})}.$$

Although the normalizing constant involves an exponential number of terms in K, for most settings of interest the number of sample groups K is reasonably small and the computation is straightforward. For example, if $K = 2$ there are $2^K = 4$ possible values that $\mathbf{g}_{ij}$ can take and equation (3.2) then simplifies to

$$p(\mathbf{g}_{ij}|\nu_{ij},\theta_{12})=\frac{\exp(\nu_{ij}(g_{1,ij}+g_{2,ij})+2\theta_{12}g_{1,ij}g_{2,ij})}{\exp(2\nu_{ij}+2\theta_{12})+2\exp(\nu_{ij})+1}. \quad (3.4)$$

The joint prior on the graphs (G_1, G_2, … G_K) is the product of the densities for each edge:

$$p(G_1,\ldots,G_K|\boldsymbol{\nu},\boldsymbol{\Theta})=\prod_{i<j}p(\mathbf{g}_{ij}|\nu_{ij},\boldsymbol{\Theta}), \quad (3.5)$$

where $\boldsymbol{\nu}=\{\nu_{ij}|1\le i<j\le p\}$. Under this prior, the conditional probability of the inclusion of edge (i, j) in G_k, given the inclusion of edge (i, j) in the remaining graphs, is

$$p(g_{k,ij}|\{g_{m,ij}\}_{m\neq k},\nu_{ij},\boldsymbol{\Theta})=\frac{\exp(g_{k,ij}(\nu_{ij}+2\sum_{m\neq k}\theta_{km}g_{m,ij}))}{1+\exp(\nu_{ij}+2\sum_{m\neq k}\theta_{km}g_{m,ij})}. \quad (3.6)$$

Parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\nu}$ influence the prior probability of selection for edges in the graphs G_1, … , G_K. In the variable selection setting, Scott and Berger (2010) find that a fixed prior probability of variable inclusion offers no correction for multiple testing. Although we are selecting edges rather than variables, a similar idea holds here. We therefore impose prior distributions on $\boldsymbol{\nu}$ and $\boldsymbol{\Theta}$ to reduce the false selection of edges. This approach is also more informative since we obtain posterior estimates of these parameters which reflect information learned from the data.

3.3 Selection prior on network similarity

As previously discussed, the matrix $\boldsymbol{\Theta}$ represents a supergraph with nonzero off-diagonal entries θ_{km} indicating that the networks for sample group k and sample group m are related. The magnitude of the parameter θ_{km} measures the pairwise similarity between graphs G_k and G_m. A complete supergraph reflects that all the inferred networks are related. For other cases, some of the networks will be related while others may be different enough to be considered independent. We learn the structure of this supergraph from the data. Our approach has the flexibility to share information between groups when appropriate, but not enforce similarity when the networks are truly different.

We place a spike-and-slab prior on the off-diagonal entries θ_{km}. See George and Mc-Culloch (1997) for a discussion of the properties of this prior. Here we want the “slab” portion of the mixture to be defined on a positive domain since θ_{km} takes on positive values for related networks. Given this restriction on the domain, we want to choose a density which allows good discrimination between zero and nonzero values of θ_{km}. Johnson and Rossell (2010, 2012) demonstrate improved model selection performance when the alternative prior is non-local in the sense that the density function for the alternative is identically zero for null values of the parameter. Since the probability density function Gamma($x|\alpha$, β) with $\alpha > 1$ is equal to zero at the point $x = 0$ and is nonzero on the domain $x > 0$, an appropriate choice for the “slab” portion of the mixture prior is the Gamma($x|\alpha$, β) density with $\alpha > 1$.

We formalize our prior by using a latent indicator variable γ_{km} to represent the event that graphs k and m are related. The mixture prior on θ_{km} can then be written in terms of the latent indicator as

$$p(\theta_{km}|\gamma_{km}) = (1-\gamma_{km}) \cdot \delta_0 + \gamma_{km} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_{km}^{\alpha-1} e^{-\beta\theta_{km}}, \quad (3.7)$$

where $\Gamma(\cdot)$ represents the Gamma function and α and β are fixed hyperparameters. As there are no constraints on the structure of $\boldsymbol{\Theta}$ (such as positive definiteness), the θ_{km}'s are variation independent and the joint prior on the off-diagonal entries of $\boldsymbol{\Theta}$ is the product of the marginal densities:

$$p(\boldsymbol{\Theta}|\boldsymbol{\gamma}) = \prod_{k<m} p(\theta_{km}|\gamma_{km}). \quad (3.8)$$

We place independent Bernoulli priors on the latent indicators

$$p(\gamma_{km}|w) = w^{\gamma_{km}} (1-w)^{(1-\gamma_{km})}, \quad (3.9)$$

where w is a fixed hyperparameter in [0, 1]. We denote the joint prior as

$$p(\boldsymbol{\gamma}) = \prod_{k<m} p(\gamma_{km}|w). \quad (3.10)$$

3.4 Edge-specific informative prior

The parameter $\boldsymbol{\nu}$ from the prior on the graphs given in equation (3.5) can be used both to encourage sparsity of the graphs $G_1, \ldots, G_K$ and to incorporate prior knowledge on particular connections. Equation (3.2) shows that negative values of ν_{ij} reduce the prior probability of the inclusion of edge (i, j) in all graphs G_k. A prior which favors smaller values for $\boldsymbol{\nu}$ therefore reflects a preference for model sparsity, an attractive feature in many applications since it reduces the number of parameters to be estimated and produces more interpretable results.

Since larger values of ν_{ij} make edge (i, j) more likely to be selected in each graph k regardless of whether it has been selected in other graphs, prior network information can be incorporated into the model through an informative prior on ν_{ij}. Given a known reference network G_0, we define a prior that encourages higher selection probabilities for edges included in G_0. When θ_{km} is 0 for all $m \neq k$ or no edges $g_{m,ij}$ are selected for nonzero θ_{km}, then the probability of inclusion of edge (i, j) in G_k can be written as

$$p(g_{k,ij}|\nu_{ij}) = \frac{e^{\nu_{ij}}}{1+e^{\nu_{ij}}} = q_{ij}. \quad (3.11)$$

We impose a prior on q_{ij} that reflects the belief that graphs G_k which are similar to the reference network $G_0 = (V, E_0)$ are more likely than graphs which have many different edges,

$$q_{ij} = \begin{cases} \text{Beta}(1+c, 1) \text{ if } (i,j) \in E_0 \\ \text{Beta}(1, 1+c) \text{ if } (i,j) \notin E_0, \end{cases} \quad (3.12)$$

where $c > 0$. This determines a prior on ν_{ij} since $\nu_{ij} = \text{logit}(q_{ij})$. After applying a univariate transformation of variables to the Beta(a, b) prior on q_{ij}, the prior on ν_{ij} can be written as

$$p(\nu_{ij}) = \frac{1}{B(a,b)} \cdot \frac{e^{a\nu_{ij}}}{(1+e^{\nu_{ij}})^{a+b}}, \quad (3.13)$$

where $B(\cdot)$ represents the beta function.

In cases where no prior knowledge on the graph structure is available a prior that favors lower values, such as $q_{ij} \sim \text{Beta}(1, 4)$ for all edges (i, j), can be chosen to encourage overall sparsity. To account for the prior belief that most edges are missing in all graphs while the few edges that are present in any one graph tend to be present in all other graphs, a prior favoring even smaller values of ν_{ij} could be coupled with a prior favoring larger values for θ_{km}.

3.5 Completing the model

The prior on the mean vector $\boldsymbol{\mu}_k$ in model (3.1) is the conjugate prior

$$\boldsymbol{\mu}_k | \boldsymbol{\Omega}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, (\lambda_0 \boldsymbol{\Omega}_k)^{-1}), \quad (3.14)$$

where $\lambda_0 > 0$, for $k = 1, 2, \ldots, K$. For the prior on the precision matrix $\boldsymbol{\Omega}_k$ we choose the G-Wishart distribution $W_G(b, \mathbf{D})$,

$$\boldsymbol{\Omega}_k | G_k, b, \mathbf{D} \sim W_G(b, \mathbf{D}), \quad (3.15)$$

for $k = 1, 2, \ldots K$. This prior restricts $\boldsymbol{\Omega}_k$ to the cone of symmetric positive definite matrices with $\omega_{k,ij}$ equal to zero for any edge $(i, j) \notin G_k$, where G_k may be either decomposable or non-decomposable. In applications we use the noninformative setting $b = 3$ and $\mathbf{D} = \mathbf{I}_p$. Higher values of the degrees of freedom parameter b reflect a larger weight given to the prior, so a prior setting with $b > 3$ and $\mathbf{D} = c \cdot \mathbf{I}_p$ for $c > 1$ could be chosen to further enforce sparsity of the precision matrix.

4 Posterior inference

Let $\boldsymbol{\Psi}$ denote the set of all parameters and $\mathbf{X}$ denote the observed data for all sample groups. We can write the joint posterior as

$$p(\mathbf{\Psi}|\mathbf{X}) \propto \prod_{k=1}^{K} [p(\mathbf{X}_k|\boldsymbol{\mu}_k, \mathbf{\Omega}_k) \cdot p(\boldsymbol{\mu}_k|\mathbf{\Omega}_k) \cdot p(\mathbf{\Omega}_k|G_k)] \cdot \prod_{i<j} [p(\mathbf{g}_{ij}|\nu_{ij}, \mathbf{\Theta}) \cdot p(\nu_{ij})] \cdot p(\mathbf{\Theta}|\boldsymbol{\gamma}) \cdot p(\boldsymbol{\gamma}). \quad (4.1)$$

Since this distribution is analytically intractable, we construct a Markov chain Monte Carlo (MCMC) sampler to obtain a posterior sample of the parameters of interest.

4.1 MCMC sampling scheme

At the top level, our MCMC scheme is a block Gibbs sampler in which we sample the network specific parameters $\mathbf{\Omega}_k$ and G_k from their posterior full conditionals. As described in Section 2, a joint search over the space of graphs and precision matrices poses computational challenges. To sample the graph and precision matrix for each group, we adapt the method of Wang and Li (2012), which does not require proposal tuning and circumvents the computation of the prior normalizing constant. We then sample the graph similarity and selection parameters $\mathbf{\Theta}$ and $\boldsymbol{\gamma}$ from their conditional posterior distributions by using a Metropolis-Hastings approach that incorporates both between-model and within-model moves, similar in spirit to the sampler proposed in Gottardo and Raftery (2008). This step is equivalent to a reversible jump. Finally, we sample the sparsity parameters $\boldsymbol{\nu}$ from their posterior conditional distribution using a standard Metropolis-Hastings step.

Our MCMC algorithm, which is described in detail in Appendix A, can be summarized as follows. At iteration t:

- Update the graph $G_k^{(t)}$ and precision matrix $\mathbf{\Omega}_k^{(t)}$ for each group $k = 1, \ldots, K$
- Update the parameters for network relatedness θ_{km}^t and $\gamma_{km}^{(t)}$ for $1 \leq k < m \leq K$
- Update the edge-specific parameters $\nu_{ij}^{(t)}$ for $1 \leq i < j \leq p$

4.2 Posterior inference and model selection

One approach for selecting the graph structure for each group is to use the maximum a posteriori (MAP) estimate, which represents the mode of the posterior distribution of possible graphs for each sample group. This approach, however, is not generally feasible since the space of possible graphs is quite large and any particular graph may be encountered only a few times in the course of the MCMC sampling. A more practical solution is to select the edges marginally. Although networks cannot be reconstructed just by looking at the marginal edge inclusion probabilities, this approach provides an effective way to communicate the uncertainty over all possible connections in the network.

To carry out edge selection, we estimate the posterior marginal probability of edge inclusion for each edge $g_{k,ij}$ as the proportion of MCMC iterations after the burn-in in which edge (i, j) was included in graph G_k. For each sample group, we then select the set of edges that appear with marginal posterior probability (PPI) > 0.5. Although this rule was proposed by Barbieri and Berger (2004) in the context of prediction rather than structure discovery, we found that it resulted in a reasonable expected false discovery rate (FDR). Following Newton et al.

(2004), we let $\xi_{k,ij}$ represent 1 - the marginal posterior probability of inclusion for edge (i, j) in graph k. Then the expected FDR for some bound

$$\text{FDR} = \frac{\sum_k \sum_{i<j} (\xi_{k,ij}) \mathbf{1}[\xi_{k,ij} \leq \kappa]}{\sum_k \sum_{i<j} \mathbf{1}[\xi_{k,ij} \leq \kappa]}, \quad (4.2)$$

where **1** is the indicator function. In the current work, we found that $\kappa = 0.5$ resulted in a reasonable posterior expected FDR, so we retain this fixed threshold. An alternative approach is to select κ so that the posterior expected FDR is below a desired level, often 0.05. Since the FDR is a monotone function of κ, this selection process is straightforward. We also compute the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) to examine the selection performance of the model under varying PPI thresholds.

Since comparison of edges across graphs is an important focus of our model, we also consider the problem of learning differential edges. We consider an edge to be differential if the true value of $|g_{k,ij} - g_{m,ij}|$ is 1, which reflects that edge (i, j) is included in either G_k or G_m but not both. We compute the posterior probability of difference $P(|g_{k,ij} - g_{m,ij}| = 1|\mathbf{X})$ as the proportion of MCMC iterations after the burn-in in which edge (i, j) was included in graph G_k or graph G_m but not both. In addition to the inference focusing on individual edges and their differences, the posterior probability of inclusion of the indicator γ_{km} provides a broad measure of the similarity of graphs k and m which reflects the utility of borrowing of strength between the groups.

The posterior estimates of ν_{ij} provide another interesting summary as they reflect the preference for edge (i, j) in a given graph based on both the prior distribution for ν_{ij} and the sampled values for $g_{k,ij}$ for $k = 1, \ldots K$. As discussed in the prior construction given in Section 3.4, the parameter q_{ij}, defined in equation (3.11) as the inverse logit of ν_{ij}, may be reasonably interpreted as a lower bound on the marginal probability of edge (i, j) in a given graph, since the MRF prior linking graphs can only increase edge probability. The utility of posterior estimates of q_{ij} in illustrating the uncertainty around inclusion of edge (i, j) is demonstrated in Section 5.1.

5 Simulations

We include two simulation studies which highlight key features of our model. In the first simulation, we illustrate our approach to inference of graphical models across sample groups and demonstrate estimation of all parameters of interest. In the second simulation, we show that our method outperforms competing methods in learning graphs with related structure.

5.1 Simulation study to assess parameter inference

In this simulation, we illustrate posterior inference using simulated data sets with both related and unrelated graph structures. We construct four precision matrices Ω_1, Ω_2, Ω_3, and Ω_4 corresponding to graphs G_1, G_2, G_3 and G_4 with different degrees of shared structure. We include $p = 20$ nodes, so there are $p \cdot (p - 1)/2 = 190$ possible edges. The precision

matrix $\mathbf{\Omega}_1$ is set to the $p \times p$ symmetric matrix with entries $\omega_{i,i} = 1$ for $i = 1, \ldots, 20$, entries $\omega_{i,i+1} = \omega_{i+1,i} = 0.5$ for $i = 1, \ldots, 19$, and $\omega_{i,i+2} = \omega_{i+2,i} = 0.4$ for $i = 1, \ldots, 18$. This represents an AR(2) model. To construct $\mathbf{\Omega}_2$, we remove 5 edges at random by setting the corresponding nonzero entries in $\mathbf{\Omega}_1$ to 0, and add 5 edges at random by replacing zeros in $\mathbf{\Omega}_1$ with values sampled from the uniform distribution on $\{[-0.6, -0.4] \cup [0.4, 0.6]\}$. To construct $\mathbf{\Omega}_3$, we remove 10 edges in both $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, and add 10 new edges present in neither $\mathbf{\Omega}_1$ nor $\mathbf{\Omega}_2$ in the same manner. To construct $\mathbf{\Omega}_4$, we remove the remaining 22 original edges shared by $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_3$ and add 22 edges which are present in none of the first three graphs. The resulting graph G_4 has no edges in common with G_1. In order to ensure that the perturbed precision matrices are positive definite, we use an approach similar to that of Danaher et al. (2013) in which we divide each off-diagonal element by the sum of the off-diagonal elements in its row, and then average the matrix with its transpose. This procedure results in $\mathbf{\Omega}_2$, $\mathbf{\Omega}_3$ and $\mathbf{\Omega}_4$ which are symmetric and positive definite, but include entries of smaller magnitude than $\mathbf{\Omega}_1$, and therefore somewhat weaker signal.

The graph structures for the four groups are shown in Figure 2. All four graphs have the same degree of sparsity, with 37/190 = 19.47% of possible edges included, but different numbers of overlapping edges. The proportion of edges shared pairwise between graphs is

$$\text{Proportion of edges shared} = \begin{pmatrix} 0.86 & 0.59 & 0.00 \\ & 0.73 & 0.14 \\ & & 0.41 \end{pmatrix}.$$

We generate random normal data using $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_4$ as the true precision matrices by drawing a random sample $\mathbf{X}_k$ of size $n = 100$ from the distribution $\mathcal{N}(0, \mathbf{\Omega}_k^{-1})$ for $k = 1, \ldots, 4$. In the prior specification, we use a Gamma(a, β) density with $a = 2$ and $\beta = 5$ for the slab portion of the mixture prior defined in equation (3.7). As discussed in Section 3.3, the choice of $a > 1$ results in a non-local prior. We would not only like the density to be zero at $\theta_{km} = 0$ to allow better discrimination between zero and nonzero values, but would also like to avoid assigning weight to large values of θ_{km}. As discussed in Li and Zhang (2010), Markov random field priors exhibit a phase transition in which larger values of parameter rewarding similarity lead to a sharp increase in the size of the selected model. For this reason, $\beta = 5$, which results in a prior with mean 0.4 such that $P(\theta_{km} \leq 1) = 0.96$, is a reasonable choice. To reflect a strong prior belief that the networks are related, we set the hyperparameter $w = 0.9$ in the Bernoulli prior on the latent indicator of network relatedness γ_{km} given in equation (3.9). We fix the parameters a and b in the prior on ν_{ij} defined in equation (3.13) to $a = 1$ and $b = 4$ for all pairs (i, j). This choice of a and b leads to a prior probability of edge inclusion of 20%, which is close to the true sparsity level.

To obtain a sample from the posterior distribution, we ran the MCMC sampler described in Section 4 with 10,000 iterations as burn-in and 20,000 iterations as the basis of inference. Figure 3 shows the traces of the number of edges included in the graphs $G_1, \ldots, G_4$. These

plots show good mixing around a stable model size. Trace plots for the remaining parameters (not shown) also showed good mixing and no strong trends.

The marginal posterior probability of inclusion (PPI) for the edge $g_{k,ij}$ can be estimated as the percentage of MCMC samples after the burn-in period where edge (i, j) was included in graph k. The heat maps for the marginal PPIs of edge inclusion in each of the four simulated graphs are shown in Figure 4. The patterns of high-probability entries in these heat maps clearly reflect the true graph structures depicted in Figure 2. To assess the accuracy of graph structure estimation, we computed the true positive rate (TPR) and false positive rate (FPR) of edge selection using a threshold of 0.5 on the PPIs. The TPR is 1.00 for group 1, 0.78 for group 2, 0.68 for group 3, and 0.57 for group 4. The FPR is 0.00 for group 1, 0.01 for group 2, 0.01 for group 3, and 0.01 for group 4. The TPR is highest in group 1 because the magnitudes of the nonzero entries in $\mathbf{\Omega}_1$ are greater than those of the other precision matrices due to the way these matrices were generated. The overall expected FDR for edge selection is 0.051. The TPR of differential edge selection is 0.73, and the FPR is 0.04. The expected FDR for differential edge selection is 0.13.

The ROC curves showing the performance of edge selection for each group under varying thresholds for the marginal PPI are shown in Figure 5. The AUC was a perfect 1.00 for group 1, 0.996 for group 2, 0.96 for group 3, and 0.94 for group 4. The overall high AUC values demonstrate that the marginal posterior probabilities of edge inclusion provide an accurate basis for graph structure learning. The lower AUC for group 4 reflects the fact that G_4 has the least shared network structure and does not benefit as much from the prior linking the graph estimation across the groups. The AUC for differential edge detection is 0.94. This result demonstrates that although our model favors shared structure across graphs, it is reasonably robust to the presence of negative association.

To assess estimation of the precision matrices $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_4$, we computed the 95% posterior credible intervals (CIs) for each entry based on the quantiles of the MCMC samples. Overall, 96.7% of the CIs for the elements $\omega_{k,ij}$ where $i \leq j$ and $k = 1, \ldots, 4$ contained the true values.

To illustrate posterior inference of the parameter ν_{ij} in equation (3.5), in Figure 6 we provide empirical posterior distributions of q_{ij}, the inverse logit of ν_{ij} defined in equation (3.11), for edges included in different numbers of the true graphs $G_1, \ldots, G_4$. Each curve represents the pooled sampled values of q_{ij} for all edges (i, j) included in the same number of graphs. Since there are no common edges between G_1 and G_4, any edge is included in at most 3 graphs. As discussed in Section 3.4, the values of q_{ij} are a lower bound on the marginal probability of edge inclusion. From this plot, we can see that the inclusion of an edge in a larger number of the simulated graphs results in a posterior density for q_{ij} shifted further away from 0, as one would expect. The means of the sampled values for q_{ij} for edges included in 0, 1, 2 or 3 simulated graphs are 0.11, 0.18, 0.25, and 0.35, respectively.

We can also obtain a Rao-Blackwellized estimate of the marginal probability of the inclusion of edge (i, j) in a graph k by computing the probabilities $p(\mathbf{g}_{ij} \mid \nu_{ij}, \mathbf{\Theta})$ defined in equation (3.2) given the sampled values of ν_{ij} and $\mathbf{\Theta}$. This results in marginal edge inclusion

probabilities for edges included in 0, 1, 2 or 3 simulated graphs of 0.13, 0.22, 0.31, and 0.44. By comparing these estimates to the values for q_{ij} given above, we can see the impact of the prior encouraging shared structure in increasing the marginal edge probabilities. A more direct estimate of the number of groups in which in edge (i, j) is present is the MCMC average of $\Sigma_k g_{k,ij}$. For edges included in either 0, 1, 2, or 3 simulated graphs, the corresponding posterior estimates of $\Sigma_k g_{k,ij}$ are 0.08, 0.77, 1.52 and 2.49. Together these summaries illustrate how varying marginal probabilities of edge inclusion translate into different numbers of selected edges across graphs.

The marginal PPIs for the elements of $\boldsymbol{\Theta}$ can be estimated as the percentages of MCMC samples with $\gamma_{km} = 1$, or equivalently with $\theta_{km} > 0$, for $1 \leq k < m \leq K$. These estimates are

$$\mathrm{PPI}(\boldsymbol{\Theta}) = \begin{pmatrix} & 1.00 & 0.88 & 0.27 \\ & & 0.84 & 0.28 \\ & & & 0.53 \\ & & & \end{pmatrix}, \quad (5.1)$$

and reflect the degree of shared structure, providing a relative measure of graph similarity across sample groups. In addition, these probabilities show that common edges are more strongly encouraged when the underlying graphs have more shared structure, since in iterations where $\theta_{km} = 0$ common edges between graphs k and m are not rewarded. The marginal posterior mean of θ_{km} conditional on inclusion, estimated as the MCMC average for iterations where $\gamma_{km} = 1$, is consistent with the inclusion probabilities in that entries with smaller PPIs also have lower estimated values when selected. The posterior conditional means are

$$\mathrm{Mean}(\theta_{km} | \gamma_{km} = 1) = \begin{pmatrix} & 0.32 & 0.28 & 0.09 \\ & & 0.20 & 0.11 \\ & & & 0.16 \\ & & & \end{pmatrix}. \quad (5.2)$$

To assess uncertainty about our estimation results, we performed inference for 25 simulated data sets, each of size $n = 100$, generated using the same procedure as above. The average PPIs and their standard errors (SE) are

$$\mathrm{Mean}(\mathrm{PPI}(\boldsymbol{\Theta})) = \begin{pmatrix} & 0.97 & 0.92 & 0.30 \\ & & 0.80 & 0.35 \\ & & & 0.60 \\ & & & \end{pmatrix}, \quad \mathrm{SE}(\mathrm{PPI}(\boldsymbol{\Theta})) = \begin{pmatrix} & 0.03 & 0.05 & 0.02 \\ & & 0.06 & 0.03 \\ & & & 0.05 \\ & & & \end{pmatrix}.$$

The small standard errors demonstrate that the results are stable for data sets with moderate sample sizes. The performance of the method in terms of graph structure learning was consistent across the simulated data sets as well. Table 1 gives the average TPR, FPR, and AUC for edge selection within each group and for differential edge selection, along with the

associated standard error (SE). The average expected FDR for edge selection was 0.07, with standard error 0.01. The expected FDR for differential edge detection was 0.14, with standard error 0.01.

5.2 Simulation study for performance comparison

In this simulation, we compare the performance of our method against competing methods in learning related graph structures given sample sizes which are fairly small relative to the possible number of edges in the graph.

We begin with the precision matrix $\mathbf{\Omega}_1$ as in Section 5.1, then follow the same procedure to obtain $\mathbf{\Omega}_2$. To construct $\mathbf{\Omega}_3$, we remove 5 edges in both $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, and add 5 new edges present in neither $\mathbf{\Omega}_1$ nor $\mathbf{\Omega}_2$ in the same manner. Finally, the nonzero values in $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_3$ are adjusted to ensure positive definiteness. In the resulting graphs, the proportion of shared edges between G_1 and G_2 and between G_2 and G_3 is 86.5%, and the proportion of shared edges between G_1 and G_3 is 73.0%.

We generate random normal data using $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_3$ as the true precision matrices by creating a random sample $\mathbf{X}_k$ of size n from the distribution $\mathcal{N}(0, \mathbf{\Omega}_k^{-1})$, for $k = 1, 2, 3$. We report results on 25 simulated data sets for sample sizes $n = 50$ and $n = 100$.

For each data set, we estimate the graph structures within each group using four methods. First, we apply the fused graphical lasso and joint graphical lasso, available in the R package JGL (Danaher, 2012). To select the penalty parameters λ_1 and λ_2, we follow the procedure recommended in Danaher et al. (2013) to search over a grid of possible values and find the combination which minimizes the AIC criterion. Next, we obtain separate estimation with G-Wishart priors using the sampler from Wang and Li (2012) with prior probability of inclusion 0.2. Finally, we apply our proposed joint estimation using G-Wishart priors with the same parameter settings as in the simulation given in Section 5.1. For both Bayesian methods, we used 10,000 iterations of burn-in followed by 20,000 iterations as the basis for posterior inference. For posterior inference, we select edges with marginal posterior probability of inclusion > 0.5.

Results on structure learning are given in Table 2. The accuracy of graph structure learning is given in terms of the true positive rate (TPR), false positive rate (FPR), and the area under the curve (AUC). The AUC estimates for the joint graphical lasso methods were obtained by varying the sparsity parameter for a fixed similarity parameter. The results reported here are the maximum obtained for the sequence of similarity parameter values tested. The corresponding ROC curves are shown in Figure 7. These curves demonstrate that the proposed joint Bayesian approach outperforms the competing methods in terms of graph structure learning across models with varying levels of sparsity.

Results show that the fused and group graphical lassos are very good at identifying true edges, but tend to have a high false positive rate. The Bayesian methods, on the other hand, have very good specificity, but tend to have lower sensitivity. Our joint estimation improves this sensitivity over separate estimation, and achieves the best overall performance as measured by the AUC for both n settings.

Results on differential edge selection are given in Table 3. For the fused and group graphical lasso, a pair of edges is considered to be differential if the edge is included in the estimated adjacency matrix for one group but not the other. In terms of TPR and FPR, the fused and group graphical lasso methods perform very similarly since we focus on differences in inclusion rather than in the magnitude of the entries in the precision matrix. The Bayesian methods have better performance of differential edge detection than the graphical lasso methods, achieving both a higher TPR and lower FPR. Relative to separate estimation with *G*-Wishart priors, the proposed joint estimation method has somewhat lower TPR and FPR. This difference reflects the fact that the joint method encourages shared structure, so the posterior estimates of differential edges are more sparse.

It is not possible to compute the AUC of differential edge detection for the fused and group graphical lasso methods since even when there is no penalty placed on the difference across groups, the estimated adjacency matrices share a substantial number of entries. Therefore, we cannot obtain a full ROC curve for these methods. The ROC curves for the Bayesian methods are given in Figure 8. Since the proposed joint estimation method is designed to take advantage of shared structure, detection of differential edges is not its primary focus. Nevertheless, it still shows slightly better overall performance than separate estimation.

5.3 Sensitivity

In assessing the prior sensitivity of the model, we observe that the choice of a and b in equation (3.13), which affects the prior probability of edge inclusion, has an impact on the posterior probabilities of both edge inclusion and graph similarity. Specifically, setting a and b so that the prior probability of edge inclusion is high results in higher posterior probabilities of edge inclusion and lower probabilities of graph similarity. This effect is logical because the MRF prior increases the probability of an edge if that edge is included in related graphs, which has little added benefit when the probability for that edge is already high. As a general guideline, a choice of a and b which results in a prior probability of edge inclusion smaller than the expected level of sparsity is recommended. Further details on the sensitivity of the results to the choice of a and b are given in Appendix B.

Smaller values of the prior probability of graph relatedness w defined in equation (3.9) result in smaller posterior probabilities for inclusion of the elements of Θ. For example, in the simulation setting of Section 5.1, using a probability of $w = 0.5$ leads to the following posterior probabilities of inclusion for the elements of Θ:

$$\mathrm{PPI}(\Theta) = \begin{pmatrix} 1.00 & 0.57 & 0.15 \\ & 0.48 & 0.15 \\ & & 0.22 \end{pmatrix}. \quad (5.3)$$

These values are smaller than those given in equation (5.1), which were obtained using $w = 0.9$, but the relative ordering is consistent.

6 Case studies

We illustrate the application of our method to inference of real-world biological networks across related sample groups. In both case studies presented below, we apply the proposed joint estimation method using the same parameter settings as the simulations in Section 5. The MCMC sampler was run for 10,000 iterations of burn-in followed by 20,000 iterations used as the basis for inference. For posterior inference, we select edges with marginal posterior probability of inclusion > 0.5.

6.1 Protein networks for subtypes of acute myeloid leukemia

Key steps in cancer progression include dysregulation of the cell cycle and evasion of apoptosis, which are changes in cellular behavior that reflect alterations to the network of protein relationships in the cell. Here we are interested in understanding the similarity of protein networks in various subtypes of acute myeloid leukemia (AML). By comparing the networks for these groups, we can gain insight into the differences in protein signaling that may affect whether treatments for one subtype will be effective in another.

The data set analyzed here, which includes protein levels for 213 newly diagnosed AML patients, is provided as a supplement to Kornblau et al. (2009) and is available for download from the MD Anderson Department of Bioinformatics and Computational Biology at http://bioinformatics.mdanderson.org/Supplements/Kornblau-AML-RPPA/aml-rppa.xls. The measurements of the protein expression levels were obtained using reverse phase protein arrays (RPPA), a high-throughout technique for protein quantification (Tibes et al., 2006). Previous work on inference of protein networks from RPPA data includes Telesca et al. (2012) and Yajima et al. (2012).

The subjects are classified by subtype according to the French-American-British (FAB) classification system. The subtypes, which are based on criteria including cytogenetics and cellular morphology, have varying prognosis. It is therefore reasonable to expect that the protein interactions in the subtypes differ. We focus here on 18 proteins which are known to be involved in apoptosis and cell cycle regulation according to the KEGG database (Kanehisa et al., 2012). We infer a network among these proteins in each of the four AML subtypes for which a reasonable sample size is available: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects). Our prior construction, which allows sharing of information across groups, is potentially beneficial in this setting since all groups have small to moderate sample sizes.

The resulting graphs from the proposed joint estimation method are shown in Figure 9, with edges shared across all subgroups in red and differential edges dashed. The edge counts for each of the four graphs and the number of overlapping edges between each pair of graphs are given below, along with the posterior probabilities of inclusion for the elements of Θ:

$$\text{Shared edge count} = \begin{pmatrix} 17 & 11 & 14 & 12 \\ & 21 & 14 & 11 \\ & & 26 & 13 \\ & & & 22 \end{pmatrix}, \quad \text{PPI}(\Theta) = \begin{pmatrix} & 0.81 & 0.83 & 0.87 \\ & & 0.91 & 0.85 \\ & & & 0.90 \\ & & & \end{pmatrix}.$$

The estimated graphs have a fair amount of overlapping structure, with 9 edges common to all four groups. This highlights the fact that our joint estimation procedure is able to account for the presence of shared structure.

6.2 Protein-signaling networks under various perturbations

The data for this case study, provided as a supplement to Sachs et al. (2005), include the levels of 11 phosphorylated proteins and phospholipids quantified using flow cytometry under 9 different experimental conditions. The sample sizes for each condition are large (in the range 700–1000) since each observation corresponds to a single cell. Sachs et al. (2005) use the 9 perturbation conditions to infer a single DAG. Subsequently, Friedman et al. (2008) use the pooled data across all perturbations to infer a single undirected graph.

We use our method to infer an undirected graph for each of the 9 conditions allowing for the possibility of shared structure. We would like to note that as the number of groups increases, the prior probability that a given edge will be shared across all groups declines. If there is a preference for shared structure across all groups, for increasing numbers of groups the prior probability of shared structure could be increased by setting the parameter w from equation (3.9) closer to 1. Since the prior formulation and posterior summaries used here are primarily focused on pairwise comparison, we retain the previous parameter settings for consistency. The resulting graph structures are shown in Figure 10, with edges shared across all subgroups in red and differential edges dashed.

The number of edges included in each graph and the number of edges shared between each pair of graphs are

$$\begin{pmatrix} 8 & 7 & 7 & 8 & 5 & 8 & 8 & 8 & 8 \\ & 9 & 7 & 8 & 6 & 8 & 7 & 9 & 9 \\ & & 8 & 8 & 5 & 8 & 7 & 8 & 8 \\ & & & 9 & 5 & 9 & 8 & 9 & 9 \\ & & & & 6 & 5 & 5 & 6 & 6 \\ & & & & & 10 & 8 & 9 & 9 \\ & & & & & & 8 & 8 & 8 \\ & & & & & & & 10 & 10 \\ & & & & & & & & 10 \end{pmatrix}.$$

The posterior probabilities of inclusion for the elements of Θ are

$$\mathrm{PPI}(\boldsymbol{\Theta}) = \begin{pmatrix} 0.82 & 0.83 & 0.87 & 0.73 & 0.86 & 0.87 & 0.86 & 0.87 \\ & 0.82 & 0.84 & 0.80 & 0.85 & 0.80 & 0.91 & 0.91 \\ & & 0.86 & 0.74 & 0.85 & 0.80 & 0.85 & 0.85 \\ & & & 0.72 & 0.90 & 0.86 & 0.89 & 0.89 \\ & & & & 0.71 & 0.74 & 0.77 & 0.78 \\ & & & & & 0.85 & 0.88 & 0.88 \\ & & & & & & 0.85 & 0.85 \\ & & & & & & & 0.94 \end{pmatrix}.$$

These probabilities reflect that group 5 is the most different from the other groups. In Figure 10, we see that it has the sparsest network, a difference that is ignored when inference is performed on the pooled data. Although some inferred connections (such as Mek–Raf and Jnk–P38) are also selected in Friedman et al. (2008), treating the data as a single group does not account for the heterogeneity across the groups and therefore results in inference of a different graph structure.

7 Discussion

In this work, we have developed a novel modeling approach to inference of multiple graphs and illustrated its important features. The proposed model utilizes a Markov random field prior to encourage shared edges between related groups and a selection prior on the parameters that describe the similarity of the networks. This approach allows us to share information between sample groups, when appropriate, as well as to obtain a measure of relative network similarity across groups. A key difference of our approach from previous work on inference of multiple graphs is that we do not assume the networks for all subgroups are related, but rather infer the relationships among them from the data.

Through simulations, we have shown that the posterior probabilities of network similarity provide a reasonable summary of network relatedness across sample groups. We have also demonstrated that our joint estimation approach increases sensitivity and enables the selection of edges that would have been missed with separate estimation procedures. Finally, we have illustrated the utility of our method in inference of protein networks across various subtypes of acute myeloid leukemia and in estimation of signaling networks under different experimental interventions.

The results reported in this paper rely on the median model for selection. As noted in Section 4.2, an alternative approach to fixing the selection threshold on the posterior probabilities would be select this threshold so that the posterior expected FDR is controlled to a desired level, typically 0.05. Applying this alternative criterion to the simulation of Section 5.1 has minimal impact on the results for edge selection since the posterior expected FDR of edge selection is already close to 0.05. For differential edge detection, however, controlling the posterior expected FDR to 0.05 results in a much higher threshold on the posterior probabilities of difference and a correspondingly lower TPR and FPR. The reason for this is that our model favors shared edges, so the posterior probabilities of edges that are

not selected in related networks are not always very close to zero, and consequently few posterior probabilities of difference are relatively large.

The approach developed here links the dependence structures within each group, but does not enforce similarity of the nonzero elements of the precision matrices. This modeling decision, which reflects our interest in network inference, was also influenced by the mathematical and computational difficulties entailed in the development of priors which not only enforce common zeros but also shrink nonzero elements toward a common mean. In the context of covariance estimation, Hoff (2009) proposes encouraging similarity of covariance matrices across groups through a hierarchical model relating their eigenvectors. This approach, however, does not enforce sparsity of the covariance or precision matrices. An extension to inference of Gaussian graphical models is not straightforward, but would be of interest for future research.

The *G*-Wishart prior framework utilized in this paper enforces exact zeros in the precision matrix corresponding to missing edges in the graph *G*. Off-diagonal entries, however, may still be arbitrarily small. Although it would be interesting to pursue a non-local prior on the precision matrices to encourage better differentiation between zero and nonzero entries, a challenge in developing such an approach is that the entries in the precision matrix are dependent due to the constraint of positive definiteness.

To integrate group-specific prior information, the model could be extended to include a parameter $\nu_{k,ij}$ for each group $k = 1, \ldots, K$. This would give additional flexibility to allow groups to have different degrees of sparsity or favor particular edges only in certain groups. In the current model formulation where the parameter ν_{ij} is shared across groups, its posterior is shaped by the observed data for each group, as illustrated in the simulation results given in Section 5.1. This implies that information can still be shared across graphs even when $\Theta = \mathbf{0}$.

Our approach provides a flexible modeling framework which can be extended to new sampling approaches or other types of data. In particular, the proposed model can be integrated with any type of *G*-Wishart sampler. Although the Wang and Li (2012) algorithm works well in practice, it has potential drawbacks. Specifically, the proposed double Metropolis-Hastings approach relies on an approximation to the posterior and requires that moves in the graph space are constrained to edge-away neighbors. The recently proposed direct sampler of Lenkoski (2013), which resolves these limitations, could be considered as an alternative. In addition, although we have focused on normally distributed data, the approach can be extended to other types of graphical models, such as Ising or log-linear models.

Acknowledgments

Christine Peterson's research has been funded by the NIH/NCI T32 Pre-Doctoral Training Program in Biostatistics for Cancer Research (NIH Grant NCI T32 CA096520), and by a training fellowship from the Keck Center of the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093. Francesco Stingo is partially supported by a Cancer Center Support Grant (NCI Grant P30 CA016672). Marina Vannucci's research is partially funded by NIH/NHLBI P01-HL082798 and NSF/DMS 1007871.

We would like to thank the two anonymous reviewers, whose feedback substantially improved this work.

References

Barbieri M, Berger J. Optimal predictive model selection. Ann Stat. 2004; 32(3):870–897.

Danaher, P. JGL: Performs the joint graphical lasso for sparse inverse covariance estimation on multiple classes. R package version 2.2. 2012. URL: http://CRAN.R-project.org/package=JGL

Danaher P, Wang P, Witten D. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc B. 2013

Dawid A, Lauritzen S. Hyper Markov laws in the statistical analysis of decomposable graphical models. Ann Stat. 1993; 21(3):1272–1317.

Dempster A. Covariance selection. Biometrics. 1972; 28:157–175.

Dobra A, Hans C, Jones B, Nevins J, Yao G, West M. Sparse graphical models for exploring gene expression data. J Multivariate Anal. 2004; 90(1):196–212.

Dobra A, Lenkoski A, Rodriguez A. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. J Am Stat Assoc. 2011; 106(496):1418–1433.

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. [PubMed: 18079126]

Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004; 303(5659):799–805. [PubMed: 14764868]

George E, McCulloch R. Approaches for Bayesian variable selection. Stat Sinica. 1997; 7:339–374.

Giudici P, Green P. Decomposable graphical Gaussian model determination. Biometrika. 1999; 86(4): 785–801.

Gottardo R, Raftery A. Markov chain Monte Carlo with mixtures of mutually singular distributions. J Comput Graph Stat. 2008; 17(4):949–975.

Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011; 98(1):1–15. [PubMed: 23049124]

Hoff P. A hierarchical eigenmodel for pooled covariance estimation. J Roy Stat Soc B. 2009; 71(5): 971–992.

Johnson V, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. J Roy Stat Soc B. 2010; 72(Part 2):143–170.

Johnson V, Rossell D. Bayesian model selection in high-dimensional settings. J Am Stat Assoc. 2012; 107(498):649–660.

Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M. Experiments in stochastic computation for high-dimensional graphical models. Stat Sci. 2005; 20(4):388–400.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 2012; 40:D109–D114. [PubMed: 22080510]

Kornblau S, Tibes R, Qiu Y, Chen W, Kantarjian H, Andreeff M, Coombes K, Mills G. Functional proteomic profiling of AML predicts response and survival. Blood. 2009; 113(1):154–164. [PubMed: 18840713]

Lauritzen, S. Graphical models. Clarendon Press; Oxford: 1996.

Lenkoski A. A direct sampler for G-Wishart variates. Stat. 2013; 2:119–128.

Lenkoski A, Dobra A. Computational aspects related to inference in Gaussian graphical models with the *G*-Wishart prior. J Comput Graph Stat. 2011; 20(1):140–157.

Li F, Zhang N. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J Am Stat Assoc. 2010; 105(491):1202–1214.

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Statist. 2006; 34(3):1436–1462.

Mukherjee S, Speed T. Network inference using informative priors. P Natl Acad Sci. 2008; 105(38): 14313–14318.

Newton M, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5(2):155–176. [PubMed: 15054023]

Peterson C, Vannucci M, Karakas C, Choi W, Ma L, Maletić-Savatić M. Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. Statistics and Its Interface. 2013; 6(4):547–558. [PubMed: 24533172]

Roverato A. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scand J Statist. 2002; 29(3):391–411.

Sachs K, Perez O, Pe'er D, Lauffenburger D, Nolan G. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005; 308(5721):523–529. [PubMed: 15845847]

Scott J, Berger J. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann Stat. 2010; 38(5):2587–2619.

Stingo F, Chen Y, Vannucci M, Barrier M, Mirkes P. A Bayesian graphical modeling approach to microRNA regulatory network inference. Ann Appl Stat. 2010; 4(4):2024–2048. [PubMed: 23946863]

Stingo F, Vannucci M. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics. 2011; 27(4):495–501. [PubMed: 21159623]

Telesca D, Müller P, Kornblau S, Suchard M, Ji Y. Modeling protein expression and protein signaling pathways. J Am Stat Assoc. 2012; 107(500):1372–1384.

Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills G, Kornblau S. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther. 2006; 5(10):2512–2521. [PubMed: 17041095]

Wang H. Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis. 2012; 7(2):771–790.

Wang H, Li S. Efficient Gaussian graphical model determination under *G*-Wishart prior distributions. Electron J Stat. 2012; 6:168–198.

Yajima M, Telesca D, Ji Y, Muller P. Differential patterns of interaction and Gaussian graphical models. COBRA Preprint Series. 2012 (Paper 91).

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

Appendix A: Details of MCMC sampling

A.1 Updating of Ω_k and G_k

For simplicity, we assume that the data for each group are column centered. The likelihood for each group is then

$$\mathbf{X}_k \sim \mathcal{N}(0, \mathbf{\Omega}_k^{-1}) \quad k=1,\ldots,K. \quad \text{(A.1)}$$

Since the *G*-Wishart distribution is conjugate to the likelihood, the posterior full conditional of $\mathbf{\Omega}_k$ is the *G*-Wishart density

$$\mathbf{\Omega}_k | \mathbf{X}_k, G_k \sim W_G(n_k+b, \mathbf{S}_k+\mathbf{D}) \quad \text{(A.2)}$$

where $\mathbf{S}_k=\mathbf{X}_k^T\mathbf{X}_k$.

Sampling from the *G*-Wishart distribution requires MCMC methods even when the graph *G* is known. In this case, we want to learn the graph structure as well, so we need to search over the joint posterior space of graphs $G_1, \ldots, G_k$ and precision matrices $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_k$ conditional on the remaining parameters. To accomplish this, we use a sampling scheme based on Algorithm 2 from section 5.2 of Wang and Li (2012). We prefer this approach over

other recent proposals since it avoids computation of prior normalizing constants and does not require tuning of proposals.

The only modification required to use the algorithm from Wang and Li (2012) to sample from the conditional distribution $p(\Omega_k, G_k|\nu, \Theta, \{G_m\}_{m\neq k})$ is to use the conditional probability $p(G_k|\nu, \Theta, \{G_m\}_{m\neq k})$ for each graph rather than the unconditional $p(G_k)$. Following their notation, when proposing a new graph G_k' which differs from the current graph G_k in that edge (i, j) is included in G_k but not in G_k', given the MRF prior on the graph structure we have

$$\frac{p(G_k'|\nu_{ij}, \Theta, \{G_m\}_{m\neq k})}{p(G_k|\nu_{ij}, \Theta, \{G_m\}_{m\neq k})} = \exp\{-(\nu_{ij} + 2\sum_{m\neq k}\theta_{km} g_{m,ij})\}. \quad \text{(A.3)}$$

At each MCMC iteration, we apply this move successively to each (i, j) for $i < j$.

A.2 Updating of θ_{km} and γ_{km}

We sample θ_{km} and γ_{km} from their joint posterior full conditional distribution. The terms in the joint prior on the graphs $G_1, \ldots, G_k$ that include θ_{km} are

$$\begin{aligned} p(G_1, \ldots, G_K|\nu, \Theta) &= \prod_{i<j} C(\nu_{ij}, \Theta)^{-1} \exp(\nu_{ij}\mathbf{1}^T\mathbf{g}_{ij} + \mathbf{g}_{ij}^T\Theta\mathbf{g}_{ij}) \\ &\propto \prod_{i<j} C(\nu_{ij}, \Theta)^{-1} \exp(2\theta_{km} g_{k,ij} g_{m,ij}), \end{aligned}$$

considering only the terms that include θ_{km}. Given the prior on θ_{km} from equation (3.7) and the prior on γ_{km} from equation (3.9), the posterior full conditional of θ_{km} and γ_{km} can be written

$$\begin{aligned} p(\theta_{km}, \gamma_{km}|\cdot) \propto &\left(\prod_{i<j} C(\nu_{ij}, \Theta)^{-1} \exp(2\gamma_{km} g_{k,ij} g_{m,ij})\right) \\ &\cdot \left((1-\gamma_{km})\cdot\delta_0 + \gamma_{km}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_{km}^{\alpha-1} e^{-\beta\theta_{km}}\right)\cdot\left(w^{\gamma_{km}}(1-w)^{(1-\gamma_{km})}\right). \end{aligned} \quad \text{(A.4)}$$

Since the normalizing constant for this mixture is not analytically tractable, we use Metropolis-Hastings steps to sample θ_{km} and γ_{km} from their joint posterior full conditional distribution for each pair (k, m) where $1 \leq k < m \leq K$. Our construction is based on the MCMC approach described in Gottardo and Raftery (2008) for sampling from mixtures of mutually singular distributions. At each iteration we perform two steps: a between-model and a within-model move. As discussed in Gottardo and Raftery (2008), this type of sampler is effectively equivalent to reversible jump Markov chain Monte Carlo (RJMCMC).

For the between-model move, if in the current state $\gamma_{km} = 1$, we propose $\gamma_{km}^* = 0$ and $\theta_{km}^* = 0$. If in the current state $\gamma_{km} = 0$, we propose $\gamma_{km}^* = 1$ and sample θ_{km}^* from the proposal density

$q(\theta^*_{km})=\text{Gamma}(\theta^*_{km}|\alpha^*,\beta^*)$. When moving from γ_{km} = 1 to $\gamma^*_{km}=0$, the Metropolis-Hastings ratio is

$$r=\frac{p(\theta^*_{km},\gamma^*_{km}|\cdot)\cdot q(\theta_{km})}{p(\theta_{km},\gamma_{km}|\cdot)} =\frac{\Gamma(\alpha)}{\Gamma(\alpha^*)}\cdot\frac{(\beta^*)^{\alpha^*}}{\beta^{\alpha}}\cdot(\theta_{km})^{\alpha^*-\alpha}\cdot e^{(\beta-\beta^*)\theta_{km}}\cdot\prod_{i<j}\frac{C(\nu_{ij},\boldsymbol{\Theta})\cdot\exp(-2\theta_{km}g_{k,ij}g_{m,ij})}{C(\nu_{ij},\boldsymbol{\Theta}^*)}\cdot\frac{1-w}{w}, \quad \text{(A.5)}$$

where Θ^* represents the matrix Θ with entry $\theta_{km}=\theta^*_{km}$. When moving from γ_{km} = 0 to $\gamma^*_{km}=1$, the Metropolis-Hastings ratio is

$$r=\frac{p(\theta^*_{km},\gamma^*_{km}|\cdot)}{p(\theta_{km},\gamma_{km}|\cdot)\cdot q(\theta^*_{km})} =\frac{\Gamma(\alpha^*)}{\Gamma(\alpha)}\cdot\frac{\beta^{\alpha}}{(\beta^*)^{\alpha^*}}\cdot(\theta^*_{km})^{\alpha-\alpha^*}\cdot e^{(\beta^*-\beta)\theta^*_{km}}\cdot\prod_{i<j}\frac{C(\nu_{ij},\boldsymbol{\Theta})\cdot\exp(2\theta^*_{km}g_{k,ij}g_{m,ij})}{C(\nu_{ij},\boldsymbol{\Theta}^*)}\cdot\frac{w}{1-w}. \quad \text{(A.6)}$$

We then perform a within-model move whenever the value of γ_{km} sampled from the between-model move is 1. For this step, we propose a new value of θ_{km} using the same proposal density as before. The Metropolis-Hastings ratio for this step is

$$r=\frac{p(\theta^*_{km},\gamma^*_{km}|\cdot)\cdot q(\theta_{km})}{p(\theta_{km},\gamma_{km}|\cdot)\cdot q(\theta^*_{km})} =\left(\frac{\theta^*_{km}}{\theta_{km}}\right)^{\alpha-\alpha^*}\cdot e^{(\beta^*-\beta)(\theta^*_{km}-\theta_{km})}\cdot\prod_{i<j}\frac{C(\nu_{ij},\boldsymbol{\Theta})\cdot\exp(2(\theta^*_{km}-\theta_{km})g_{k,ij}g_{m,ij})}{C(\nu_{ij},\boldsymbol{\Theta}^*)} \quad \text{(A.7)}$$

A.3 Updating of ν_{ij}

To find the posterior full conditional distribution of ν_{ij}, we consider the terms in the joint prior on the graphs $G_1, \ldots, G_k$ that include ν_{ij}:

$$p(G_1,\ldots,G_K|\nu,\boldsymbol{\Theta})=\prod_{i<j}C(\nu_{ij},\boldsymbol{\Theta})^{-1}\exp(\nu_{ij}\mathbf{1}^T\mathbf{g}_{ij}+\mathbf{g}_{ij}^T\boldsymbol{\Theta}\mathbf{g}_{ij}) \propto C(\nu_{ij},\boldsymbol{\Theta})^{-1}\exp(\nu_{ij}\mathbf{1}^T\mathbf{g}_{ij}),$$

considering only the terms that include ν_{ij}. Given the prior from equation (3.13), the posterior full conditional of ν_{ij} given the data and all remaining parameters is proportional to

$$p(\nu_{ij}|\cdot)\propto\frac{\exp(a\nu_{ij})}{(1+e^{\nu_{ij}})^{a+b}}\cdot C(\nu_{ij},\boldsymbol{\Theta})^{-1}\exp(\nu_{ij}\mathbf{1}^T\mathbf{g}_{ij}) =\frac{\exp(\nu_{ij}(a+\mathbf{1}^T\mathbf{g}_{ij}))}{C(\nu_{ij},\boldsymbol{\Theta})\cdot(1+e^{\nu_{ij}})^{a+b}} \quad \text{(A.8)}$$

For each pair (i, j) where $1 \le i < j \le p$, we propose a value q^* from the density Beta(2, 4), then set ν^* = logit(q^*). The proposal density can be written in terms of ν^* as

$$q(\nu^*)=\frac{1}{B(a^*,b^*)}\cdot\frac{e^{a^*\nu^*}}{(1+e^{\nu^*})^{a^*+b^*}}. \quad \text{(A.9)}$$

For the simulation given in Section 5.1, this proposal resulted in an average acceptance rate of 38.8%, which is a reasonable proportion. Although the use of a fixed proposal may result in low acceptance rates in some situations, the efficiency of this step is not a pressing concern since we require many iterations to search the graph space, so we can obtain a reasonable sample of ν_{ij} even if the mixing is slow. The Metropolis-Hastings ratio is

$$r=\frac{p(\nu^*|\cdot)}{p(\nu_{ij}|\cdot)}\frac{q(\nu_{ij})}{q(\nu^*)} =\frac{\exp((\nu^*-\nu_{ij})\cdot(a-a^*+\mathbf{1}^T\mathbf{g}_{ij}))\cdot C(\nu_{ij},\boldsymbol{\Theta})\cdot(1+e^{\nu_{ij}})^{a+b-a^*-b^*}}{C(\nu^*,\boldsymbol{\Theta})\cdot(1+e^{\nu^*})^{a+b-a^*-b^*}} \quad \text{(A.10)}$$

Appendix B: Details of sensitivity analysis

Here we provide more details of the sensitivity analysis summarized in Section 5.3.

B.1 Sensitivity to prior parameters *a* and *b*

The parameters *a* and *b* are the shape and scale parameters of the Beta prior on the parameter q_{ij} defined in equation (3.11). The parameter q_{ij} can be interpreted as a lower bound on the prior probability of inclusion for edge (*i*, *j*) which may be increased by the effect of the prior encouraging shared structure across groups.

To assess the impact of the choice of *a* and *b* on posterior inference, we applied the proposed joint estimation method at a range of (*a*, *b*) settings to a single fixed data set generated following the setup of the simulation given in Section 5.1. The results given in Section 5.1 were obtained using the setting $a = 1$ and $b = 4$, which reflects a Beta prior on q_{ij} with mean 0.2. To examine the effect of varying *a* and *b*, we performed inference for 6 additional settings chosen so that mean of the Beta prior ranged from 0.05 to 0.35 while the variance of the Beta prior remained fixed. The effect on the average edge PPIs and on the average PPI for the entries of Θ is summarized in Figure 11.

The average edge PPIs showed a steady increase from just over 0.17 for prior means in the range 0.05 – 0.10 to around 0.19 for prior mean 0.35. The direction of the effect is logical, and the overall difference in levels is not strong. The average PPIs for the elements of Θ are relatively stable for prior means up 0.25, just above the true sparsity level of 0.20. Beyond this point, they decline sharply, demonstrating that shared structure is no longer rewarded when the prior on q_{ij} results in a prior probability of edge inclusion much greater than the true level before factoring in the impact of the sharing of information across graphs.

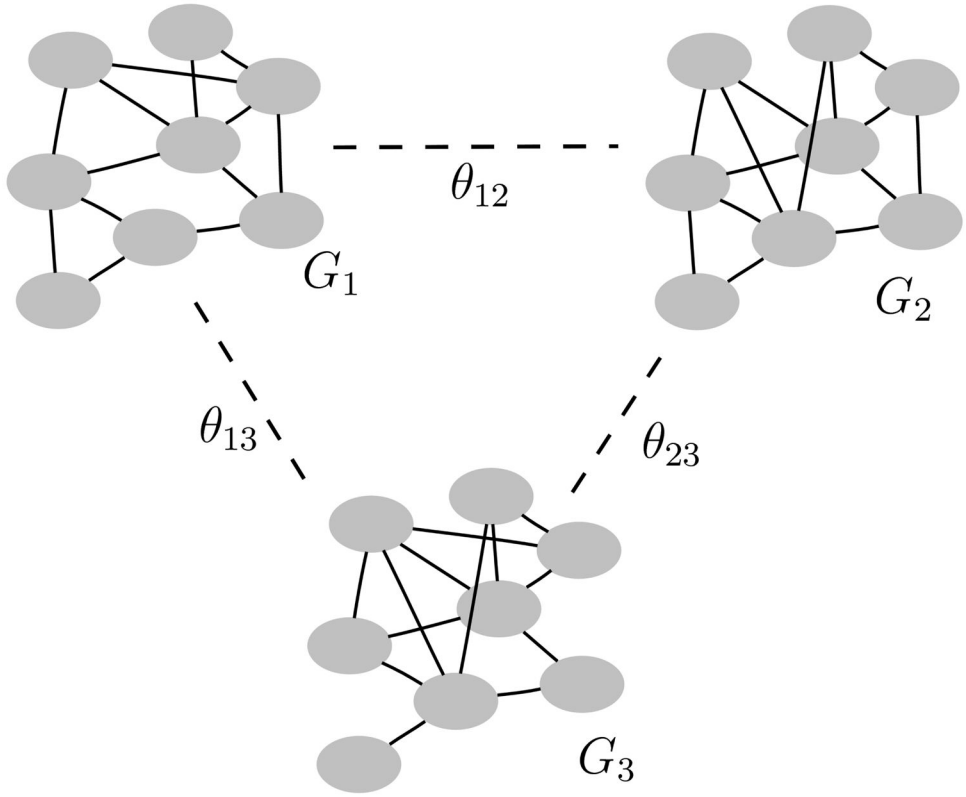

Figure 1.
Illustration of the model for three sample groups. The parameters θ_{12}, θ_{13}, and θ_{23} reflect the pairwise similarity between the graphs G_1, G_2, and G_3.

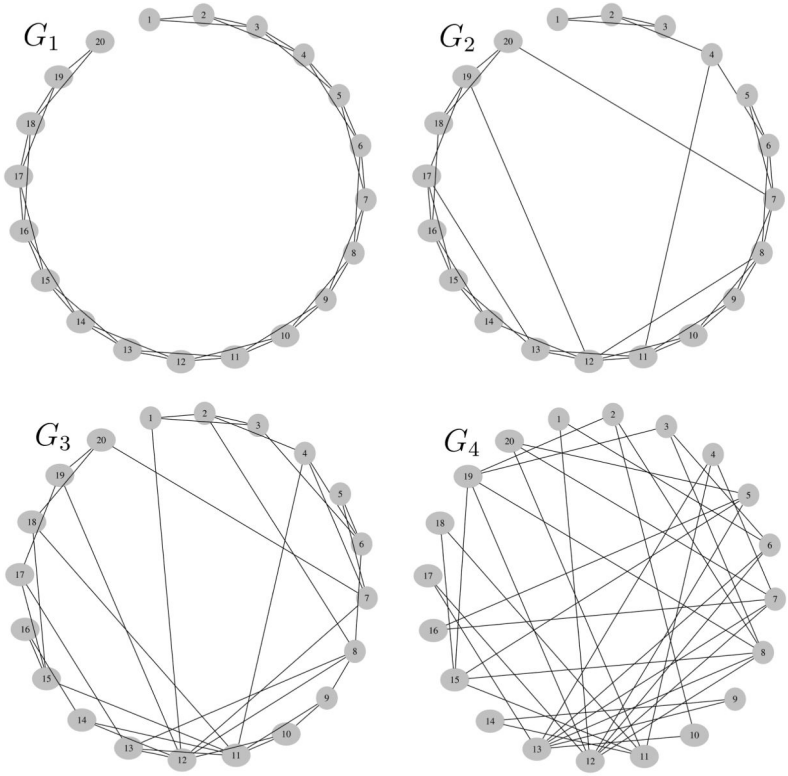

Figure 2.
Simulation of Section 5.1. True graph structures for each simulated group.

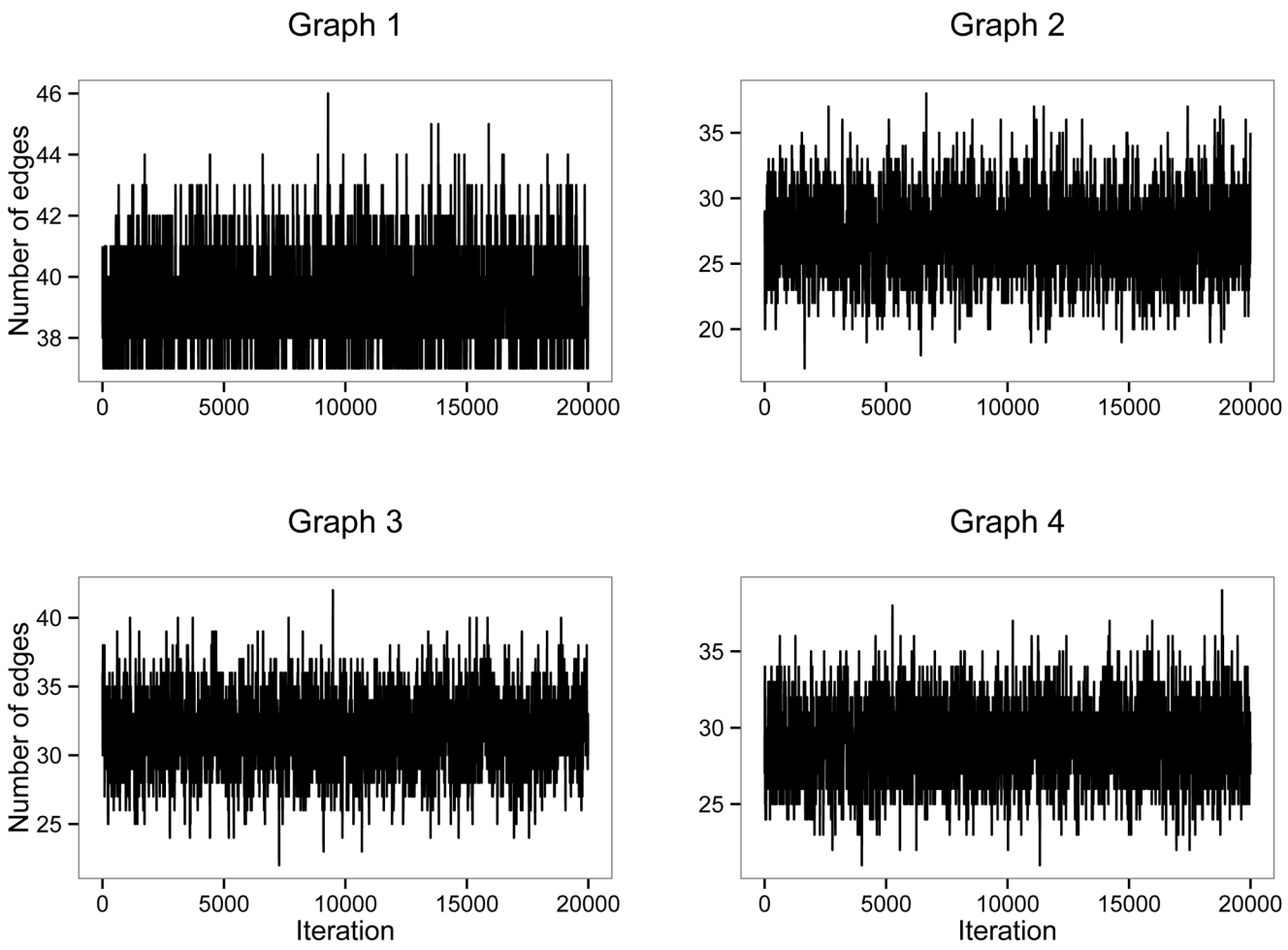

Figure 3.
Simulation of Section 5.1. Trace plots of the number of edges included in each graph, thinned to every fifth iteration for display purposes.

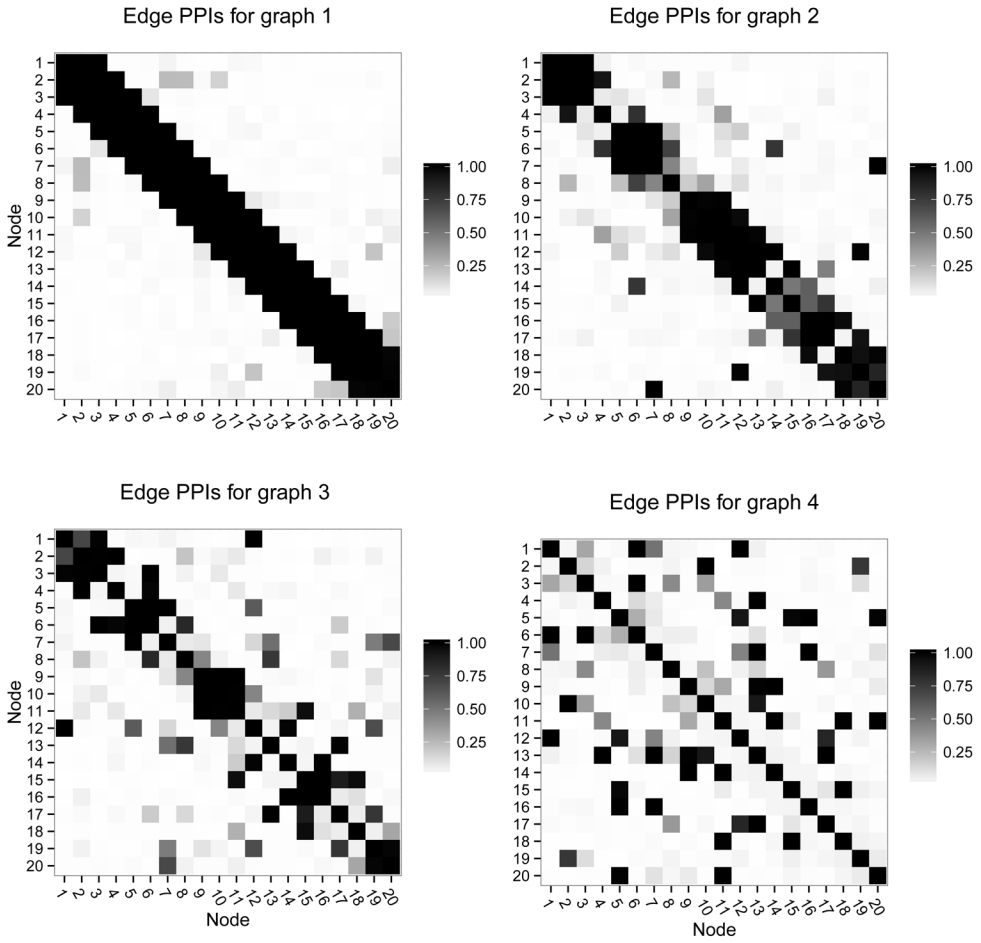

Figure 4.

Simulation of Section 5.1. Heat maps of the posterior probabilities of edge inclusion (PPIs) for the four simulated graphs.

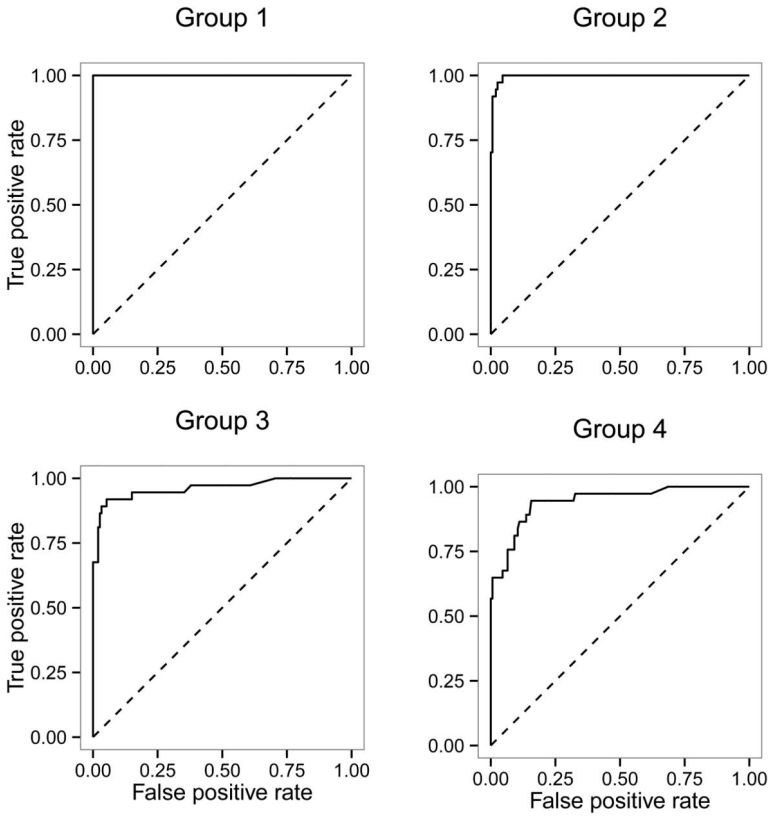

Figure 5.

Simulation of Section 5.1. ROC curves for varying thresholds on the posterior probability of edge inclusion for each of the simulated groups. The corresponding AUCs are 1.00 for group 1, 0.996 for group 2, 0.96 for group 3 and 0.94 for group 4.

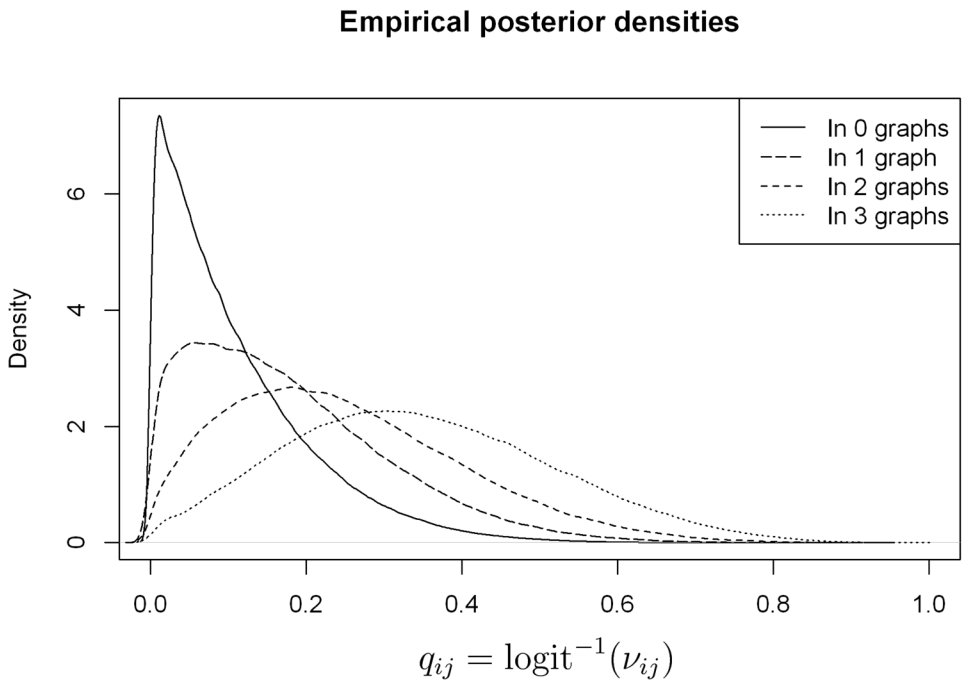

Figure 6.

Simulation of Section 5.1. Empirical posterior densities of edge-specific parameters q_{ij} for edges included in 0, 1, 2 or 3 of the simulated graphs.

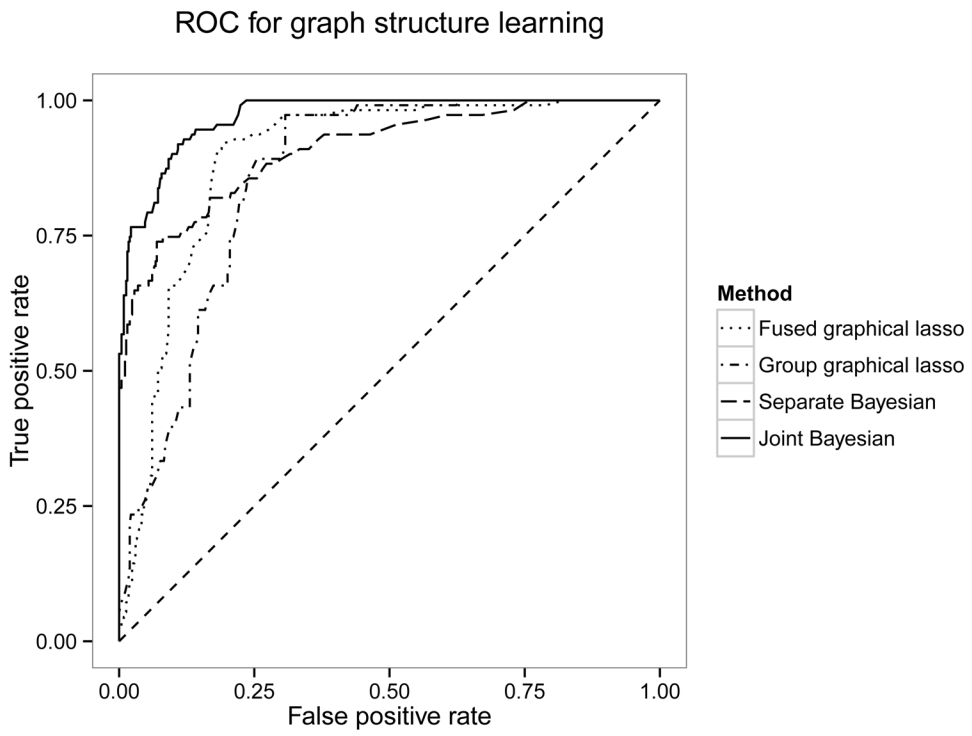

Figure 7.
Simulation of Section 5.2. ROC curves for graph structure learning for sample size $n = 50$.

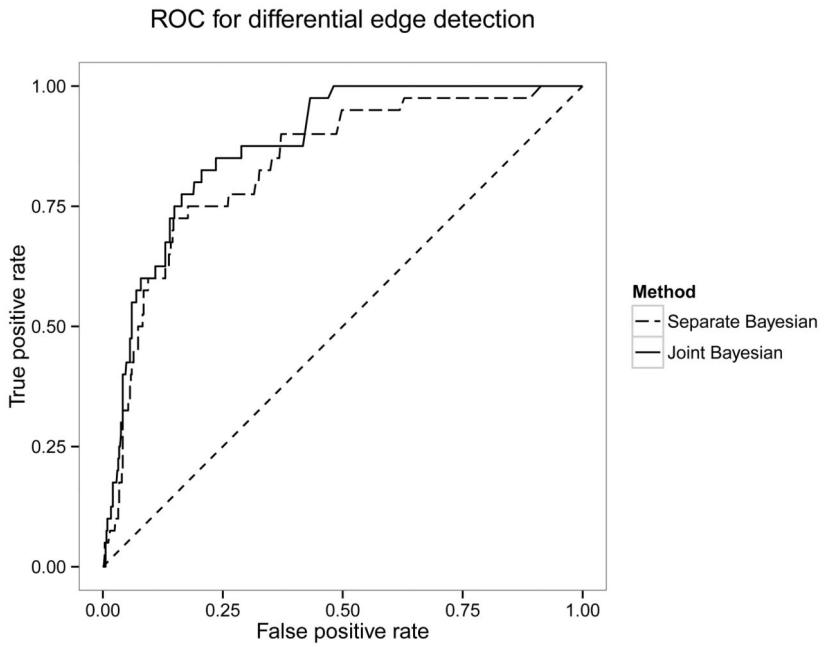

Figure 8.
Simulation of Section 5.2. ROC curves for differential edge detection for sample size $n = 50$.

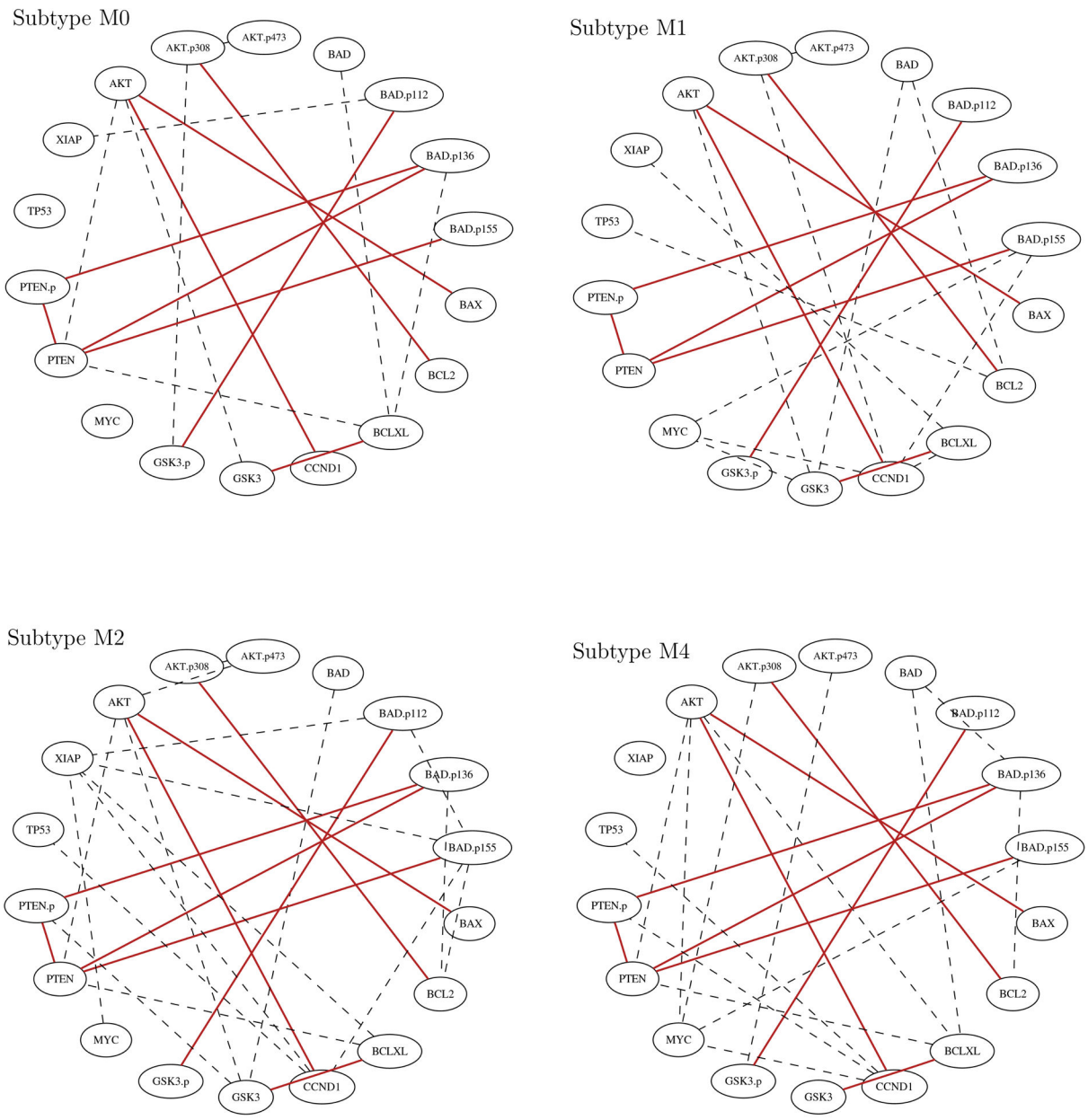

Figure 9.

Case study of Section 6.1. Inferred protein networks for the AML subtypes M0, M1, M2, and M4, with edges shared across all subgroups in red and differential edges dashed.

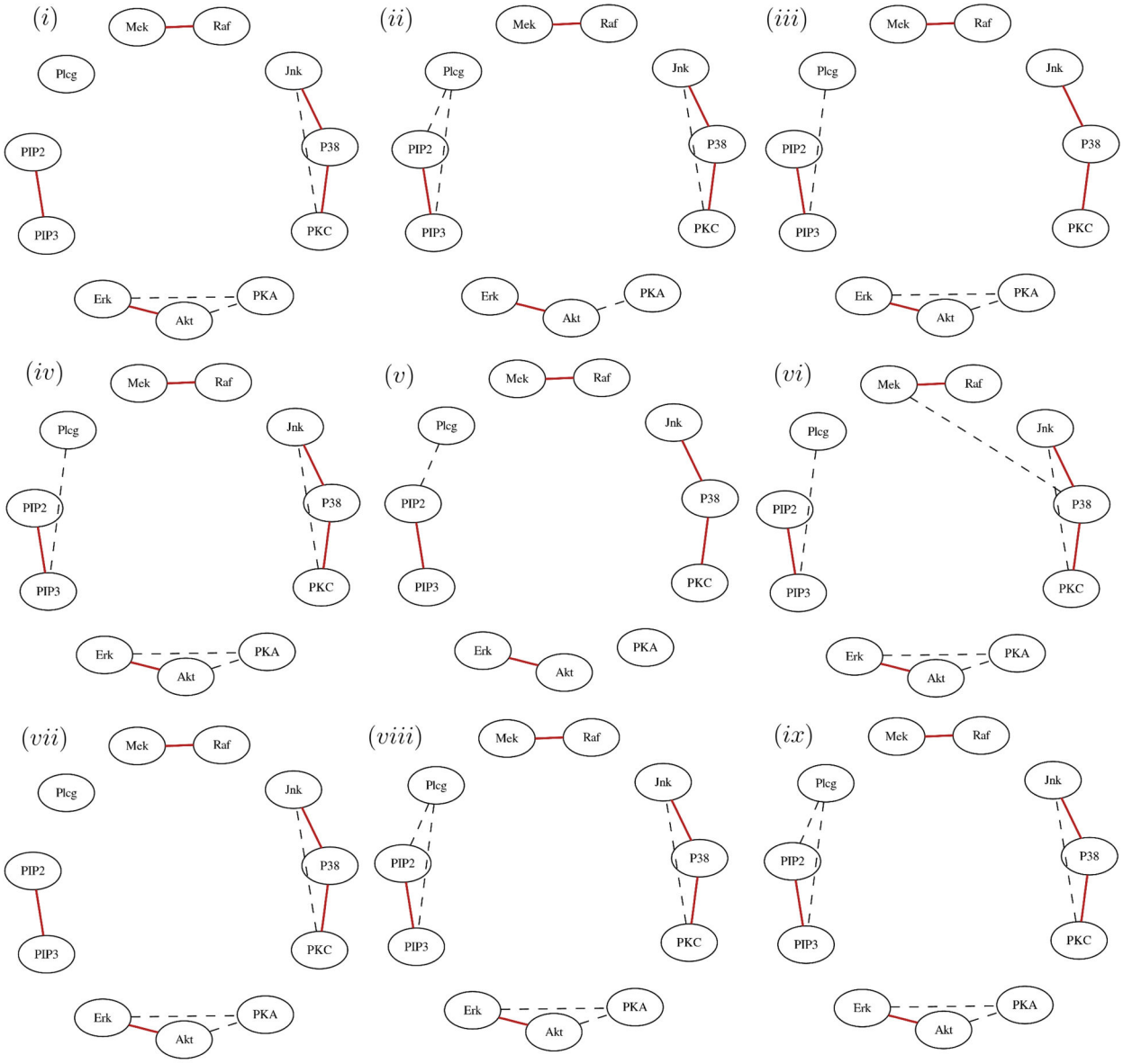

Figure 10.

Case study of Section 6.2. Inferred protein signaling networks, with edges shared across all subgroups in red and differential edges dashed.

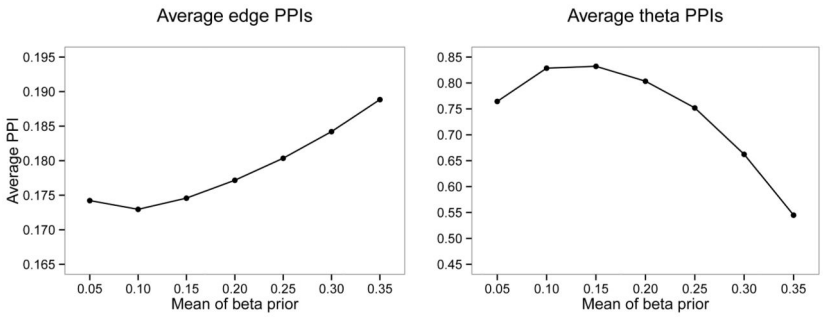

Figure 11.

Simulation of Section B.1. Sensitivity of the average edge PPIs (left) and average PPIs for the elements of Θ (right) to the parameters a and b in the prior $q_{ij} \sim \text{Beta}(a, b)$.

Table 1

Simulation of Section 5.1. Average true positive rate (TPR), false positive rate (FPR), and area under curve (AUC) with associated standard error (SE) across 25 simulated data sets.

	TPR (SE)	FPR (SE)	AUC (SE)
Group 1	1.00 (0.01)	0.002 (0.003)	1.00 (0.002)
Group 2	0.61 (0.08)	0.007 (0.006)	0.98 (0.01)
Group 3	0.73 (0.05)	0.007 (0.008)	0.98 (0.01)
Group 4	0.63 (0.06)	0.006 (0.005)	0.94 (0.02)
Differential	0.71 (0.03)	0.039 (0.006)	0.94 (0.01)

Table 2

Simulation of Section 5.2. Results for graph structure learning, with a comparison of true positive rate (TPR), false positive rate (FPR), and area under the curve (AUC) with standard errors (SE) over 25 simulated datasets.

	n = 50			n = 100		
	TPR (SE)	**FPR (SE)**	**AUC (SE)**	**TPR (SE)**	**FPR (SE)**	**AUC (SE)**
Fused graphical lasso	0.93 (0.03)	0.52 (0.10)	0.91 (0.01)	0.99 (0.01)	0.56 (0.10)	0.93 (0.01)
Group graphical lasso	0.93 (0.03)	0.55 (0.07)	0.88 (0.02)	0.99 (0.01)	0.63 (0.05)	0.91 (0.01)
Separate estimation with *G*-Wishart priors	0.52 (0.03)	0.010 (0.006)	0.91 (0.01)	0.68 (0.03)	0.004 (0.002)	0.97 (0.01)
Joint estimation with *G*-Wishart priors	0.58 (0.04)	0.008 (0.004)	0.97 (0.01)	0.78 (0.05)	0.003 (0.002)	0.99 (0.003)

Table 3

Simulation of Section 5.2. Results for differential edge detection, with a comparison of true positive rate (TPR), false positive rate (FPR), and area under the curve (AUC) with standard errors (SE) over 25 simulated datasets.

	n = 50			n = 100		
	TPR (SE)	**FPR (SE)**	**AUC (SE)**	**TPR (SE)**	**FPR (SE)**	**AUC (SE)**
Fused graphical lasso	0.46 (0.11)	0.43 (0.03)	n/a	0.44 (0.13)	0.41 (0.02)	n/a
Group graphical lasso	0.45 (0.11)	0.43 (0.04)	n/a	0.45 (0.13)	0.41 (0.02)	n/a
Separate estimation with *G*-Wishart priors	0.59 (0.07)	0.11 (0.01)	0.85 (0.02)	0.80 (0.06)	0.09 (0.01)	0.93 (0.02)
Joint estimation with *G*-Wishart priors	0.56 (0.08)	0.09 (0.01)	0.88 (0.02)	0.78 (0.08)	0.06 (0.01)	0.95 (0.02)