Published in final edited form as:

Stat. 2014 January 1; 3(1): 109–125. doi:10.1002/sta4.49.

Bayesian sparse graphical models and their mixtures

Rajesh Talluri^{a,*}, Veerabhadran Baladandayuthapani^a, and Bani K. Mallick^b

^aThe University of Texas MD Anderson Cancer Center, Houston, TX, USA

^bTexas A&M University, College Station TX, USA

Abstract

We propose Bayesian methods for Gaussian graphical models that lead to sparse and adaptively shrunk estimators of the precision (inverse covariance) matrix. Our methods are based on lasso-type regularization priors leading to parsimonious parameterization of the precision matrix, which is essential in several applications involving learning relationships among the variables. In this context, we introduce a novel type of selection prior that develops a sparse structure on the precision matrix by making most of the elements exactly zero, in addition to ensuring positive definiteness – thus conducting model selection and estimation simultaneously. More importantly, we extend these methods to analyze clustered data using finite mixtures of Gaussian graphical model and infinite mixtures of Gaussian graphical models. We 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 result from the restriction of positive definiteness on the correlation matrix. We evaluate the operating characteristics of our method via several simulations and demonstrate the application to real data examples in genomics.

Keywords

bayesian; covariance selection; finite mixtures; gaussian graphical models; infinite mixtures; sparse modeling

1. Introduction

Consider the p dimensional random vector $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(p)})$, 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 modeling of the covariance matrix, $\boldsymbol{\Sigma}$, or equivalently the precision matrix, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, is one of the most important tasks in analyzing Gaussian multivariate data. Furthermore, it has a direct relationship to constructing Gaussian graphical models (GGMs)

Copyright © 2012 John Wiley & Sons, Ltd.

*rtalluri@mdanderson.org.

Supporting Information

Additional information for this article is available:

Section S1: Posterior sampling schemes for the Bayesian lasso selection model

Section S2: Posterior inference and conditional distributions for finite mixture of Gaussian graphical models

Section S3: Posterior inference and conditional distributions for infinite mixture of Gaussian graphical models

Section S4: Simulations for Bayesian lasso selection model

by identifying the significant edges. Of particular interest in this structure is the identification of zero entries in the precision matrix Ω . 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 & Kiiveri, 1986; Wong et al., 2003), which provides a framework for the exploration of multivariate dependence patterns. GGMs are tools for modeling conditional independence relationships. Among the practical advantages of using GGMs in high-dimensional problems are 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.

There have been many approaches to Gaussian graphical modeling. In a Bayesian setting, modeling is based on hierarchical specifications for the covariance matrix (or precision matrix) using global conjugate priors on the space of positive-definite matrices, such as inverse Wishart priors or its equivalents. Dawid & Lauritzen (1993) introduced an equivalent form as the hyper-inverse Wishart (HIW) distribution. Although that construction enjoys many advantages, such as computational efficiency due to its conjugate formulation and exact calculation of marginal likelihoods (Scott & 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 examples, but a prohibitively large number of graphs for a moderately large p . Moreover, extension to a nondecomposable graph is nontrivial and computationally expensive using reversible-jump algorithms (Giudici & Green, 1999; Brooks et al., 2003). G-Wishart prior distributions has been proposed as an generalization of HIW priors for nondecomposable graphs (Roverato, 2002; Atay-Kayis & Massam, 2005). Although it is convenient to use HIW prior due to its conjugate nature, it has several limitations. First, due to global nature of the hyper-prior specifications, HIW priors become more restrictive. Second, sampling from the non-decomposable HIW distribution is challenging due to the presence of an unknown normalizing constant. Computationally challenging Monte Carlo based techniques have been proposed (Roverato, 2002; Dellaportas et al., 2003; Jones et al., 2004; Carvalho & Scott, 2009; Lenkoski & Dobra, 2011; Wang et al., 2010; Mitsakakis et al., 2011; Dobra et al., 2011; Wang et al., 2012). Due to this computational burden, extension of these models in more complex frameworks (like mixtures of graphical models) makes it a daunting task.

Alternate approaches for more adaptive estimation and/or selection in graphical models are based on methods that enforce sparsity either via variable (edge) selection or regularization/penalization approaches. In a regression context for variable selection problems, such priors have been proposed by (George & McCulloch, 1993, 1997; Kuo & Mallick, 1998; Dellaportas et al., 2000, 2002; Fan & Li, 2001; Hans, 2010, 2009; Tibshirani et al., 2005; Griffin et al., 2013). 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 in the order of p^2 instead of p . Regularization based covariance estimation has been done in a frequentist framework (Rothman et al., 2009; Peng et al., 2009; Levina et al., 2008). 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 graphical model context, in a frequentist setting Meinshausen & Bühlmann (2006), Yuan & Lin (2007), Friedman et al. (2008) and Anjum et al. (2009) 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 & Mallick, 2004; Park & 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 (Wang, 2012).

There have been several attempts to shrink the covariance/precision matrix via matrix factorizations for unrestricted search over the space of both decomposable and nondecomposable 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 modeling schemes using decompositions of the correlation matrix. We adopt this strategy in this paper and propose Bayesian methods for GGMs that allow for simultaneous model selection and parameter estimation. We introduce a novel type of prior in Section 2 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, 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 certain types of correlated 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. The main scientific question 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 modeling and inferential challenges are related to determining the number of components, as well as estimating the underlying graph for each component. In Section 3 we extend the Bayesian lasso selection model to model finite mixtures of Gaussian graphical models. In Section 4 we further generalize the finite mixture model to infinite mixtures using Dirichlet process priors, where the number of mixture components are unknown. In

Section 5 the practical applicability of the proposed model is demonstrated through analyzing a gene expression dataset and the models validated using appropriate simulations. We provide a discussion and conclusion in Section 6. A detailed supplementary material is provided which includes posterior inference and sampling schemes for all the models along with additional validation of the models through a variety of simulation scenarios.

2. Bayesian graphical lasso selection model

To formalize notations, let $\mathbf{Y}_{p \times n} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ be a $p \times n$ matrix with n independent samples and p variates, where each sample $\mathbf{Y}_i = (Y_i^{(1)}, \dots, Y_i^{(p)})$ is a p dimensional vector corresponding to the p variates; each sample $\mathbf{Y}_i$ comes from a multivariate normal distribution with mean $\boldsymbol{\mu}$; and the covariance matrix between the p variates is Σ . We have n samples with covariance matrix $\sigma^2 \mathbf{I}_n$, which implies n independent samples with variance σ^2 . This can be formalized as follows: $\mathbf{Y}$ follows a matrix normal distribution $N(\boldsymbol{\mu}, \Sigma, \sigma^2 \mathbf{I}_n)$ with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix Σ between the p variates ($Y^{(1)}, \dots, Y^{(p)}$) and σ^2 is the variance of the samples. This is identical to the multivariate normal distribution $MVN(\boldsymbol{\mu}_v, \Sigma \otimes \sigma^2 \mathbf{I}_n)$ where $\boldsymbol{\mu}_v$ is the vectorized form of matrix $\boldsymbol{\mu}$ and $\otimes$ is the Kronecker product. Given a random sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, we wish to estimate the precision matrix $\Omega = \Sigma^{-1}$. The maximum likelihood estimator of $(\boldsymbol{\mu}, \Sigma)$ is $(\bar{\mathbf{Y}}, \bar{\mathbf{V}})$ where

$\bar{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$. The commonly used sample covariance matrix is $\hat{\mathbf{S}} = n\bar{\mathbf{V}}/(n-1)$. The precision matrix Ω can be estimated by $\bar{\mathbf{V}}^{-1}$ or $\hat{\mathbf{S}}^{-1}$. 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 Ω . In addition, given that our main aim is to explore the conditional relationships among the variables, our main interest is the identification of zero entries in the precision matrix because a zero entry $\Omega_{ij} = 0$ indicates conditional independence between the two covariates $Y^{(i)}$ and $Y^{(j)}$, given all other covariates. We propose different priors over Ω to explore these zero entries. Here and throughout the paper we follow the notation that $\theta_1 | \theta_2$ represents the conditional distribution of the random variable θ_1 given θ_2 so that the likelihood of the Gaussian graphical model is written as

$\mathbf{Y} | G \sim (2\pi\sigma^2)^{-\frac{np}{2}} |\Omega|^{\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2} \text{tr}\{\Omega \mathbf{Y} \mathbf{Y}^T\}\}$. Instead of modeling the entire $p \times p$ precision matrix, Ω , we explore local dependencies by breaking the model down into components. In our modeling framework, we work directly with standard deviations and a correlation matrix following the *separation* strategy of Barnard et al. (2000) that do not correspond to any particular type of parameterization (e.g., the Cholesky decomposition). Specification of a reasonable prior for the entire precision matrix is complicated because Wishart priors (or equivalents) are not fully general, but restrict the degrees of freedom of the partial standard deviations. Using this decomposition, prior beliefs on the partial standard deviations and correlations can be easily accommodated.

To this end, we can parameterize the precision matrix as $\Omega = \mathbf{S} \mathbf{C} \mathbf{S}$, where $\mathbf{S}$ is the diagonal matrix of standard deviations and $\mathbf{C}$ is the correlation matrix. The partial correlation

coefficients ρ_{ij} are related to C_{ij} as $\rho_{ij} = \frac{-\Omega_{ij}}{(\Omega_{ii}\Omega_{jj})^{\frac{1}{2}}} = -C_{ij}$. In this setting, our primary construct of interest is $\mathbf{C}$, which we wish to model in an adaptive manner. One could model

the elements of $\mathbf{C}$ directly using shrinkage priors such as Laplace priors, which gives us the formulation of the graphical lasso models explored by Meinshausen & Bühlmann (2006), Yuan & Lin (2007), Friedman et al. (2008) and Wang (2012). However, in the Bayesian framework the Laplace priors leads to shrinkage of the partial correlation but does not set them exactly to zero, i.e., does not explicitly carry out selection, which is of essence in graphical models. To this end, we decompose the correlation matrix $\mathbf{C}$ as $\mathbf{C} = \mathbf{A} \odot \mathbf{R}$, where $\odot$ is the Hadamard operator that conducts the element-wise multiplication. This parameterization helps us to divide the problem into two parts: shrinkage and selection, using the respective elements of $\mathbf{R}$ and $\mathbf{A}$, with the following properties: (a) both $\mathbf{A}$ and $\mathbf{R}$ are symmetric matrices; (b) the diagonal elements of both matrices are ones and, (c) the off-diagonal elements of the *selection matrix*: $\mathbf{A}$ consist of binary random variables (0 or 1), whereas the off-diagonal elements of the *shrinkage matrix*: $\mathbf{R}$ model the partial correlations between the variables with elements that lie between $[-1, 1]$. Moreover, $\mathbf{R}$ can thought of as a *pre-correlation matrix* which may not be a positive definite matrix but we constrain the convoluted correlation matrix $\mathbf{C} = \mathbf{A} \odot \mathbf{R}$ to be positive definite. We do so by jointly modeling $\mathbf{A}$ and $\mathbf{R}$ as detailed below.

2.1. Joint prior specification of shrinkage and selection matrices

We propose joint prior specifications on the shrinkage matrix $\mathbf{R}$ and selection matrix $\mathbf{A}$ as follows. We use a Laplace prior on the elements of the shrinkage matrix defined as

$f(R_{ij}|\tau_{ij}) \propto \frac{1}{2\tau_{ij}} \exp(-\frac{|R_{ij}|}{\tau_{ij}})$, with each individual element ij having its own scale parameter, τ_{ij} , that controls the level of sparsity. As discussed previously, Laplace priors have been widely used for shrinkage applications. Since $\mathbf{A}$ is the selection matrix that performs the variable selection on the elements of the matrix $\mathbf{R}$, 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 $\mathbf{A}$, 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.

To specify a joint prior for $\mathbf{A}$ and $\mathbf{R}$, we have to satisfy the constraint that $\mathbf{C} = \mathbf{A} \odot \mathbf{R}$ is positive definite. Hence, the joint prior can be expressed as

$$R_{ij}, A_{ij} | \tau_{ij}, q_{ij} \sim \text{Laplace}(0, \tau_{ij}) \text{Bernoulli}(q_{ij}) I(\mathbf{C} \in \mathbb{C}_p),$$

where $-1 \leq R_{ij} \leq 1$, $0 \leq q_{ij} \leq 1$ and $I(\mathbf{C} \in \mathbb{C}_p) = 1$ if $\mathbf{C}$ is a correlation matrix and is 0 otherwise. Therefore, we ensure the positive definiteness of $\mathbf{C}$ using plausible values of $\mathbf{A}$ and $\mathbf{R}$. That way, R_{ij} and A_{ij} are dependent through the indicator function. The joint prior of $\mathbf{A}$ and $\mathbf{R}$ can be specified as

$$\mathbf{A}, \mathbf{R} | \boldsymbol{\tau}, \mathbf{q} \sim \prod_{i < j} \text{Laplace}(0, \tau_{ij}) \text{Bernoulli}(q_{ij}) I(\mathbf{C} \in \mathbb{C}_p)$$

where τ and q are the vectors containing τ_{ij} and q_{ij} values respectively. The full specification of the constraints on the C_{ij} 's to ensure the positive definiteness are discussed in Section S1 (Supplementary Material).

R_{ij} 's are not independent under this prior specification due to the constraint of positive definiteness of C through the indicator function. To observe the marginal prior distribution of R_{ij} , we simulate from this joint prior distribution by fixing $A_{ij} = 1$ for all i, j so that C is identical to R . Figure 1 shows marginal distributions of some of the r_{ij} 's with different values of τ_{ij} 's when $p = 3, 10$ and 20 . In addition, we have also plotted the marginal prior distribution arising from a constrained uniform distribution as suggested by Barnard et al. (2000). It is clear from the figures that the marginal prior on individual correlations under the joint Laplace prior shrinks more tightly towards zero compared to the joint uniform prior. Furthermore, the effect of dimension p is negligible on Laplace prior distribution for small τ_{ij} values. In this setting, the shrinkage parameter τ_{ij} controls the degree of sparsity, i.e., determines the degree to which the ij^{th} element of R will be shrunk towards zero. Accordingly, we treat this τ_{ij} as an unknown parameter and estimate them adaptively using the data. 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 $A = 1_n$ (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 & Lin (2007), where the single penalty parameter (τ) 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 & Brown (2011) in the variable selection context.

Furthermore, q_{ij} is assigned a beta prior as $q_{ij} \sim \text{Beta}(a, b)$, $i < j$. In this construction the hyper-parameters 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 A and R above gives us the graphical lasso selection model 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 : $S_i \sim IG(g, h)$, $i = 1, 2, \dots, p$. We use MCMC techniques for estimation and inference for the above specified models that includes hybrid Gibbs and Metropolis-Hastings strategies to explore the graph space. A critical step in our moves is the evaluation of normalizing constants which result from the positive-definite constraint on the correlation matrices. Detailed posterior sampling schemes are detailed in Section S1 of supplementary material.

3. Mixtures of Gaussian graphical models

3.1. Motivating example

A strength of the proposed method is that it can be employed in a more complex modeling framework in a hierarchical manner. In this section we use the proposed method to develop finite mixture graphical models, where each mixture component is assumed to follow a GGM with an adaptive covariance structure. Our motivation for this model arises from the analysis of a high-throughput genomics data set. Suppose we have a gene expression data set with n samples and g genes, and are interested in detecting K subtypes of cancer among the n samples. We assume a different network (graph) structure of these g genes for each cancer subtype and have a primary goal of using this information efficiently to cluster the samples into the correct subtype of cancer. In addition, we wish to identify biologically significant differences among the networks to explain the variations between the cancer subtypes. To this end, we extended the Bayesian lasso selection models outlined in the previous sections in a finite mixture set-up. This framework allows for additional flexibility over that of regular GGMs by allowing a heterogeneous population to be divided into groups that are more homogeneous and which share similar connectivity or graphs.

3.2. The hierarchical finite mixture GGM model

Let $\mathbf{Y}_{p \times n} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ 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 $\mathbf{Y}_i$ follows a multivariate normal distribution $\mathcal{N}(\theta_j, \Sigma_j)$ if it belongs to the j th group. Given a random sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, we wish to estimate the number of mixtures K as well as the precision

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

$$\mathbf{Y}_i | L_i = j, \theta, \Omega \sim \mathcal{N}(\theta_j, \Omega_j^{-1}).$$

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

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

and the associated class probabilities follow a Dirichlet distribution as

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

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

$$\theta_j | \mathbf{B} \sim N(\mathbf{0}, \mathbf{B}),$$

We assign a common inverse Wishart prior for covariance matrix $\mathbf{B}$ across groups as $\mathbf{B} \sim IW(\mathbf{v}_0, \mathbf{B}_0)$, where $\mathbf{v}_0$ is the shape parameter and $\mathbf{B}_0$ is the scale matrix. The hierarchical specification of the GGM structure for each group Ω_j parallels the development of the previous section, 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:

$Y_{i L_i=j}, \boldsymbol{\theta}, \boldsymbol{\Omega}$	$\sim$	$N(\boldsymbol{\theta}_j, \boldsymbol{\Omega}_j^{-1})$
L_i	$\sim$	$\text{Multinomial}(n, [p_1, p_2, \dots, p_K])$
$p_1, p_2, \dots, p_K \alpha$	$\sim$	$\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_K)$
$\boldsymbol{\theta}_j \mathbf{B}$	$\sim$	$N(\mathbf{0}, \mathbf{B})$
$\mathbf{B}$	$\sim$	$IW(\mathbf{v}_0, \mathbf{B}_0)$
$\boldsymbol{\Omega}_j$	$=$	$S_j(A_j \odot \mathbf{R}_j)S_j$
$A_{j(lm)} q_{j(lm)}$	$\sim$	$\text{Bernoulli}(q_{j(lm)}), i < j$
$\mathbf{R}_j A_j$	$\sim$	$\prod_{l < m} \text{Laplace}(0, \tau_{j(l,m)}) / (C_j \in \mathbb{C}_p)$
$\tau_{j(lm)}$	$\sim$	$IG(e, f)$
$S_{j(l)}$	$\sim$	$IG(g, f),$

where i denotes the sample, j denotes the mixture component, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, K$. In addition, $A_{j(lm)}$ and $\tau_{j(lm)}$ denote the lm^{th} component of the A_j and the τ_j , $l = 1, 2, \dots, p$, $m = 1, \dots, p$. Posterior inference and conditional distributions are detailed in section S2 of the supplementary material. 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 Section S2 (Supplementary Material). Alternative ways of model determination are to find the Bayes factor using the MCMC samples (Chib & Jeliazkov, 2001).

4. Infinite Mixtures of Gaussian Graphical Models

We have extended our modeling framework to infinite mixtures of Gaussian graphical models in cases where the number of mixture components are unknown. The added advantage of this procedure is the number of mixtures is treated as a model unknown and will be determined adaptively by data rather than using BIC. Infinite mixtures of Gaussian graphical models using hyper inverse wishart priors were previously proposed by Rodriguez et al. (2011). The proposed model improves upon it by incorporating shrinkage and selection priors to estimate mixtures of sparse graphical models.

Following previous section, $\mathbf{Y}_{p \times n} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ is a $p \times n$ matrix with n samples and p covariates and the likelihood function is

$$Y_i|\theta_i, \Omega_i \sim N(\theta_i, \Omega_i^{-1}).$$

Let $\gamma_i = (\theta_i, \Omega_i^{-1})$. We propose the Dirichlet process prior (Ferguson, 1973; Dey et al., 1998; Müller & Quintana, 2004) on $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n)$ which can be written as

$$\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n) \sim DP(\alpha, H_\phi),$$

where α is the precision parameter H_ϕ is the base distribution of the Dirichlet process (DP) prior. In a sequence of draws $\gamma_1, \gamma_2, \dots$ from the Polya urn representation of the Dirichlet process (Blackwell & 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 $\mathbf{f}_{-n} = \{\gamma_1, \dots, \gamma_n\} - \{\gamma_n\}$ and d_{n-1} = number of preexisting clusters of tied samples in γ_{-n} at the n th draw, then we have

$$f(\gamma_n | \mathbf{f}_{-n}, \alpha, \phi) = \frac{\alpha}{\alpha + n - 1} H_\phi + \sum_{j=1}^{d_{n-1}} \frac{n_j}{\alpha + n - 1} \delta_{\gamma_j}, \quad (1)$$

where H_ϕ is the base prior, and the j th cluster has n_j tied samples that are commonly expressed by γ_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 $\{\gamma_1, \dots, \gamma_{d_n}\}$, where d_n is essentially the number of clusters. We induce sparsity into the model using the base distribution H_ϕ which defines the cluster configuration. We allow the individual means of each group to follow a Normal distribution as

$$\theta_j | \Omega_j \sim N(0, \Omega_j^{-1}).$$

The hierarchical specification of the GGM structure for each group Ω_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. Hence, $H_\phi = N_{\theta|\Omega}(\cdot) F_\Omega$ where F_Ω is the baseline prior for the precision matrix. The hierarchical model for the base prior can be summarized as follows:

$$\begin{aligned} H_\phi &\propto N_{\theta}(\cdot) F_\Omega \\ \Omega &= \mathbf{S}(\mathbf{A} \odot \mathbf{R}) \mathbf{S} \\ F_\Omega &\propto F_A(\cdot) F_R(\cdot) F_S(\cdot) \\ \mathbf{A}, \mathbf{R} | \boldsymbol{\tau}, \mathbf{Q} &\sim \prod_{i < j} \text{Laplace}(0, \tau_{ij}) \text{Bernoulli}(q_{ij}) I(\mathbf{C} \in \mathbb{C}_p) \\ \boldsymbol{\tau} &\sim IG(\nu_e, \nu_f) \\ \mathbf{Q} &\sim \text{Beta}(\nu_c, \nu_d) \\ \mathbf{S} &\sim IG(\nu_\alpha, \nu_\beta). \end{aligned}$$

The parametrization is similar to the models detailed in the previous sections. 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 & West, 1995). We need to use Metropolis Hastings algorithm to handle the non-conjugacy (Neal, 2000). The posterior inference and conditional distributions for this model are detailed in section S3 of the supplementary material.

5. Simulations and Real Data Analysis

5.1. Application to a genomics dataset

For the real data analysis, we used the leukemia data from Golub et al. (1999) as a case study to illustrate our mixtures of graphical models. 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 2 which shows that the expression profiles of these genes form distinct groups of genes that behave concordantly – hence warranting a more through investigation to explicitly explore the dependence patterns that vary by group.

We fit the finite mixtures of Gaussian graphical models to this data set to find the top graphs for the data. We ran the Markov Chain Monte Carlo simulation for 100000 samples and removed the first 20000 samples as burn-in. We obtained 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 3 and Figure 4. 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 play a role in encoding cell surface proteins which are 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 plays a role in producing an important protein important for cell adhesion. 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 remodeling (RBBP4 and SMARC4), transcription (GTF2E2), and cell adhesion (zyxin and ITGAX), or are known oncogenes (MYB, TCF3 and HOXA9) Golub et al. (1999). 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 remodeling 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 overexpression is responsible transformation in myeloid cells and causes 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.

5.2. Simulations

For validation of the Bayesian lasso selection model we compared our method with the glasso (Friedman et al., 2008) and MB (Meinshausen & Bühlmann, 2006) methods because both methods use the L1-regularization and are closest to our approach using Laplace priors. We assessed the performance of these methods in terms of the Kullback-Leibler (K-L) loss. The Kullback-Leibler loss is defined as $_{KL}(\hat{\Omega}, \hat{\Omega}) = \text{trace}(\hat{\Omega}\hat{\Omega}^{-1}) - \log|\hat{\Omega}\hat{\Omega}^{-1}| - p$. The simulations were designed using sparse, identity, banded, block and dense structures for the precision matrix. In terms of the K-L loss, the Bayesian method outperforms the glasso and MB methods for the sparse unstructured and banded diagonal matrices. All the methods performed equivalently for the identity matrix, the block diagonal, and the dense structure. The complete simulation details are presented in section S4 of the supplementary material.

For validation of the mixture model 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 $(\hat{\mu}_j, \hat{\Omega}_j^{-1})$ denote the estimates of the mean and precision matrices corresponding to the ALL ($j = 1$) and AML ($j = 2$) groups, respectively. We generated data under the convolution of the following multivariate normal likelihood, $\mathbf{Y}_j \sim N(\hat{\mu}_j, \hat{\Omega}_j^{-1})$, 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 & 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 Markov Chain Monte Carlo 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 5, 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 similar 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 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 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 6. 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 using MCLUST. 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 better in recovering the true sparse precision matrix. We found that our methods scale reasonably until we reach around 300 covariates but above that level the high computational complexity did not allow for a reasonable computation time. The computational time for the above simulation is approximately 20 hours for 100000 MCMC runs. 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.

6. Discussion and conclusions

In this article we develop a Bayesian framework for adaptive estimation of precision matrices in Gaussian graphical models. 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 outperform the existing methods.

Our simulations and analysis suggest that it is feasible to implement adaptive GGMs and mixtures of GGMs using Markov Chain Monte Carlo for a reasonable number of variables (p in a few hundreds). 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 modeling 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. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

RT research was supported in part by a cancer prevention fellowship supported by a grant from the National Institute of Drug Abuse (NIH R25 DA026120). VB research is partially supported by NIH grant R01 CA160736 and the Cancer Center Support Grant (CCSG) (P30 CA016672). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We also thank Lee Ann Chastain for editing the manuscript.

References

- Anjum S, Doucet A, Holmes CC. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*. 2009; 25(22):2929–2936. [PubMed: 19696047]
- Atay-Kayis A, Massam H. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*. 2005; 92(2):317.
- Bae K, Mallick B. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*. 2004; 20(18):3423–3430. [PubMed: 15256404]
- Barnard J, McCulloch R, Meng X. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*. 2000; 10(4):1281–1312.
- Blackwell D, MacQueen JB. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*. 1973; 1(2):353–355.
- Brooks S, Giudici P, Roberts G. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2003; 65(1):3–39.
- Carvalho CM, Scott JG. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*. 2009; 96(3):497–512.

- Chib S, Jeliazkov I. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*. 2001; 96(453):270–281.
- Dawid AP, Lauritzen SL. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*. 1993; 21(3):1272–1317.
- Dellaportas P, Forster J, Ntzoufras I. Bayesian variable selection using the Gibbs sampler. *Generalized Linear Models: A Bayesian Perspective*. Dey, D.; Ghosh, S.; Mallick, B., editors. Chemical Rubber Company Press; New York: 2000.
- Dellaportas P, Forster JJ, Ntzoufras I. On Bayesian model and variable selection using MCMC. *Statistics and Computing*. 2002; 12:27–36.
- Dellaportas P, Giudici P, Roberts G. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhya, Series A*. 2003; 65(1):43–55.
- Dempster AP. Covariance selection. *Biometrics*. 1972; 28(1):157–175.
- Dey, D.; Müller, P.; Sinha, D. *Practical nonparametric and semiparametric Bayesian statistics*. Springer; New York: 1998.
- Dobra A, Lenkoski A, Rodriguez A. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*. 2011; 106(496).
- Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*. 1995; 90(430):577–588.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*. 1973; 1(2):209–230.
- Figueiredo MAT. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis Machine Intelligence*. 2003; 25(9):1150–1159.
- Fraley C, Raftery AE. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*. 2007; 24(2):155–181.
- Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008; 9(3):432–441. [PubMed: 18079126]
- George, El; McCulloch, RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*. 1993; 88(423):881–889.
- George, El; McCulloch, RE. Approaches for Bayesian variable selection. *Statistica Sinica*. 1997; 7:339–373.
- Giudici, P. *Learning in graphical Gaussian models*. Oxford University Press; London: 1996.
- Giudici P, Green PJ. Decomposable graphical Gaussian model determination. *Biometrika*. 1999; 86(4):785–801.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286(5439):531–537. [PubMed: 10521349]
- Griffin JE, Brown PJ. Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*. 2011; 53(4):423–442.
- Griffin JE, Brown PJ, et al. Some priors for sparse regression modelling. *Bayesian Analysis*. 2013; 8(3):691–702.
- Hans C. Bayesian lasso regression. *Biometrika*. 2009; 96(4):835–845.
- Hans C. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*. 2010; 20(2):221–229.
- Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*. 2004; 20:388–400.
- Kuo L, Mallick B. Variable selection for regression models. *Sankhya, Series B*. 1998; 60(1):65–81.
- Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: A Bayesian variable selection approach. *Bioinformatics*. 2003; 19(1):90–97. [PubMed: 12499298]

- Lenkoski A, Dobra A. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*. 2011; 20(1):140–157.
- Levina E, Rothman A, Zhu J. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*. 2008; 2(1):245–263.
- Liechty JC, Liechty MW, Muller P. Bayesian correlation estimation. *Biometrika*. 2004; 91(1):1–14.
- Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*. 2006; 34(3):1436–1462.
- Mitsakakis N, Massam H, Escobar MD, et al. A Metropolis-Hastings based method for sampling from the G-Wishart distribution in Gaussian graphical models. *Electronic Journal of Statistics*. 2011; 5:18–30.
- Müller P, Quintana FA. Nonparametric Bayesian data analysis. *Statistical Science*. 2004; 19(1):95–110.
- Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*. 2000; 9(2):249–265.
- Park T, Casella G. The Bayesian lasso. *Journal of the American Statistical Association*. 2008; 103(482):681–686.
- Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*. 2009; 104(486):735–746. [PubMed: 19881892]
- Rodriguez A, Lenkoski A, Dobra A, et al. Sparse covariance estimation in heterogeneous samples. *Electronic Journal of Statistics*. 2011; 5:981–1014.
- Rothman A, Levina E, Zhu J. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*. 2009; 104(485):177–186.
- Roverato A. Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*. 2000; 87(1):99–112.
- Roverato A. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*. 2002; 29(3):391–411.
- Scott JG, Carvalho CM. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*. 2008; 17(4):790–808.
- Speed TP, Kiiveri HT. Gaussian Markov distributions over finite graphs. *The Annals of Statistics*. 1986; 14(1):138–150.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996; 58(1):267–288.
- Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67(1):91–108.
- Wang H. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*. 2012; 7(4):867–886.
- Wang H, Carvalho CM, et al. Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics*. 2010; 4:1470–1475.
- Wang H, Li SZ, et al. Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics*. 2012; 6:168–198.
- Wong F, Carter CK, Kohn R. Efficient estimation of covariance selection models. *Biometrika*. 2003; 90(4):809–830.
- Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007; 94(1):19–35.

Figure 1.

Shown here are the marginal priors on R_{ij} for different values of $\tau = 1, .5, .1$ and compared with the uniform prior.

Figure 2.
Heat map of top 50 genes in leukaemia data set.

Figure 3. Significant edges for the genes in the ALL cluster. The red (green) lines between the proteins indicate a negative (positive) correlation between the proteins.

Figure 4.

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

Figure 5.

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.

Figure 6.

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.