Bayesian Analysis (2012) **7**, Number 4, pp. 867–886

Bayesian Graphical Lasso Models and Efficient Posterior Computation

Hao Wang [*]

Abstract. Recently, the graphical lasso procedure has become popular in estimating Gaussian graphical models. In this paper, we introduce a fully Bayesian treatment of graphical lasso models. We first investigate the graphical lasso prior that has been relatively unexplored. Using data augmentation, we develop a simple but highly efficient block Gibbs sampler for simulating covariance matrices. We then generalize the Bayesian graphical lasso to the Bayesian adaptive graphical lasso. Finally, we illustrate and compare the results from our approach to those obtained using the standard graphical lasso procedures for real and simulated data. In terms of both covariance matrix estimation and graphical structure learning, the Bayesian adaptive graphical lasso appears to be the top overall performer among a range of frequentist and Bayesian methods.

Keywords: Adaptive graphical lasso, Block Gibbs sampler, Constrained parameter spaces, Covariance matrix estimation, Double-exponential distribution, Graphical lasso

1 Introduction

The graphical lasso (Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Banerjee et al. 2008; Friedman et al. 2008; Guo et al. 2011) is widely used for simultaneous graphical model determination and covariance matrix estimation. Let $\boldsymbol{S}$ be the sum of the products matrix such that $\boldsymbol{S} = \mathbf{Y}'\mathbf{Y}$ where $\mathbf{Y}(n \times p)$ is the data matrix of p variables and n samples. The graphical lasso problem is to maximize the penalized log-likelihood

$$\log(\det \boldsymbol{\Omega}) - \mathbf{tr}(\frac{\boldsymbol{S}}{n}\boldsymbol{\Omega}) - \rho||\boldsymbol{\Omega}||_1 \tag{1}$$

over the space of positive definite matrices M^+ with $\rho \geq 0$ being the shrinkage parameter. Here, $\boldsymbol{\Omega} = (\omega_{ij})$ is the $p \times p$ inverse of the covariance matrix and $||\boldsymbol{\Omega}||_1 = \sum_{1 \leq i,j \leq p} |\omega_{ij}|$ is the L_1 norm of $\boldsymbol{\Omega}$. Some authors omit the diagonal elements from the penalty. We discuss the graphical lasso that applies the penalty to all of the elements in $\boldsymbol{\Omega}$ unless otherwise noted. Equation (1) is a convex objective function and various algorithms have been proposed to solve it (Yuan and Lin 2007; Friedman et al. 2008; Rothman et al. 2008). The penalty parameter ρ is chosen by cross-validation or criteria similar to the Bayesian information criterion (BIC).

The graphical lasso has a Bayesian interpretation. The graphical lasso estimator is

[*] Department of Statistics, University of South Carolina, Columbia, SC, U.S.A. haowang@sc.edu

DOI:10.1214/12-BA729

equivalent to the maximum a posteriori estimation of the following model:

$$
\begin{aligned}
p(\mathbf{y}_i \mid \boldsymbol{\Omega}) &= \mathrm{N}(\mathbf{y}_i \mid 0, \boldsymbol{\Omega}^{-1}) \quad (i = 1, \ldots, n), \\
p(\boldsymbol{\Omega} \mid \lambda) &= C^{-1} \prod_{i<j} \Big\{ \mathrm{DE}(\omega_{ij} \mid \lambda) \Big\} \prod_{i=1}^{p} \Big\{ \textsc{Exp}(\omega_{ii} \mid \frac{\lambda}{2}) \Big\} 1_{\boldsymbol{\Omega} \in M^{+}},
\end{aligned}
\tag{2}
$$

where *(i)* $\mathrm{N}(\mathbf{y} \mid \mathbf{,} \boldsymbol{\Sigma})$ represents the density function, evaluated at $\mathbf{y}$, of a multivariate normal random vector with mean and covariance matrix $\boldsymbol{\Sigma}$; *(ii)* $\mathrm{DE}(x \mid \lambda)$ represents the double exponential density function of the form $p(x) = \lambda/2 \exp(-\lambda|x|)$; *(iii)* $\textsc{Exp}(x \mid \lambda)$ represents the exponential density function of the form $p(x) = \lambda \exp(-\lambda x) 1_{x>0}$; and *(iv)* C is the normalizing constant not involving λ as shown in Section (2.5). For any fixed values $\lambda \geq 0$, the posterior mode of $\boldsymbol{\Omega}$ is the graphical lasso estimate with $\rho = \lambda/n$. Bayesian interpretations based on the model (2) and its variants have been examined by Marlin et al. (2009) and Marlin and Murphy (2009), among others. The emphasis of these works has been placed on constructing flexible penalties to induce group and block structures in graphs and on efficiently finding the maximum a posteriori estimation of the corresponding posterior distribution, with little mention of the properties of the prior distribution and the inference that can address parameter estimation uncertainty and the choice of shrinkage parameters.

In this paper, our objective is to develop a framework for efficient Bayesian inference for the graphical lasso model (2) with low, i.e., p =10, to moderate, i.e., p =100–200, dimensions. We first investigate the properties of the graphical lasso prior, along with the discussion of the simulation-based Bayesian inference. We then provide a novel block Gibbs sampler for sampling $\boldsymbol{\Omega}$ from model (2). This block Gibbs sampler is remarkably efficient − it generates 1000 iterations with excellent mixing for $p = 100$ problems in about 1.2 minutes under a MATLAB implementation. The basic Bayesian graphical lasso then generalizes to the Bayesian adaptive graphical lasso to overcome the well-known shortcomings of double exponential priors. We empirically illustrate that the Bayesian adaptive graphical lasso is a very attractive method for both covariance matrix estimation and graphical structure learning.

Finally, we note that the work reported here was developed independently and concurrently by a recent paper of Khondker et al. (2012). Their work has substantial overlap as well as differences with ours. The main difference is that our algorithm is a block Gibbs sampler while Khondker et al. (2012)'s algorithm is a random walk Metropolis-Hasting. We expect our block Gibbs sampler to be much more efficient than their Metropolis-Hasting algorithm, especially for higher dimensional problems. Moreover, we have investigated the distributional properties of the graphical lasso priors while Khondker et al. (2012) only developed a sampling algorithm without exploring the properties of the prior distributions.

2 The Bayesian graphical lasso

2.1 The graphical lasso prior

The graphical lasso prior (2) has the form of the product of double exponential densities. However, due to the positive definite constraint, the resulting marginal distributions for individual ω_{ij}'s are not double-exponential. Figure 1 (a)–(c) display marginal distributions of one of the p diagonal elements, one of the $p(p-1)/2$ off-diagonal elements, and one of the $p(p-1)/2$ partial correlations, respectively, when $\lambda = 3$ and $p \in \{2, 10, 50\}$. These densities are based on Monte Carlo samples generated by the sampling schemes described in Section 2.4. Clearly, the marginal distribution of individual diagonal elements is not exponential and tends to have larger mean and variance as p increases. The marginal distribution of individual off-diagonal elements has more probability mass around 0 than the double-exponential distribution with the scale parameter $\lambda = 3$. The partial correlation becomes tighter around 0 as p increases. More interestingly, a change of variable of $\boldsymbol{\Omega}$ reveals that the marginal distribution of partial correlations does not depend on λ under the joint prior (2). This suggests that, regardless of λ, the graphical lasso prior increasingly favors the value of a partial correlation close to 0 as p grows. When structural learning is based on the posterior distributions of partial correlations, it is desirable to have a prior that increases its shrinkage of a partial correlation towards 0 as p increases in order to control the number of false positive signals.

We next examine the hierarchical representation of the graphical lasso prior (2). The double exponential distribution can be represented as a scale mixture of normals (Andrews and Mallows 1974; West 1987). In the Bayesian regression lasso, the assumption of prior independence allows the use of this hierarchical representation (Park and Casella 2008; Hans 2009). The use of a scale mixture of normals for the ω_{ij}'s in the graphical lasso prior seems to be a natural choice, however, the positive definite constraint implies that the normals for the ω_{ij}'s are no longer independent given the scale parameters. Let $\boldsymbol{\omega} = \{\omega_{ij}\}_{i\leq j}$ be the vector of the upper off-diagonal and diagonal entries of $\boldsymbol{\Omega}$, $\boldsymbol{\tau} = \{\tau_{ij}\}_{i<j}$ be the latent scale parameters and

$$p(\boldsymbol{\omega} \mid \boldsymbol{\tau}, \lambda) \quad = \quad C_{\boldsymbol{\tau}}^{-1} \prod_{i<j} \left\{ \frac{1}{\sqrt{2\pi\tau_{ij}}} \exp(-\frac{\omega_{ij}^2}{2\tau_{ij}}) \right\} \prod_{i=1}^{p} \left\{ \frac{\lambda}{2} \exp(-\frac{\lambda}{2}\omega_{ii}) \right\} 1_{\boldsymbol{\Omega}\in M^+}, \quad (3)$$

where the normalizing term, $C_{\boldsymbol{\tau}}$, depends on $\boldsymbol{\tau}$ and is analytically intractable. To induce the marginal distribution that is of the form (2), the following mixing density is proposed here for $\boldsymbol{\tau}$,

$$p(\boldsymbol{\tau} \mid \lambda) \quad \propto \quad C_{\boldsymbol{\tau}} \prod_{i<j} \frac{\lambda^2}{2} \exp(-\frac{\lambda^2}{2}\tau_{ij}). \quad (4)$$

Note that $p(\boldsymbol{\Omega} \mid \boldsymbol{\tau})$ and $p(\boldsymbol{\tau} \mid \lambda)$ are proper priors because

$$\begin{aligned} C_{\boldsymbol{\tau}} &= \int \prod_{i<j} \left\{ \frac{1}{\sqrt{2\pi\tau_{ij}}} \exp(-\frac{\omega_{ij}^2}{2\tau_{ij}}) \right\} \prod_{i=1}^{p} \left\{ \frac{\lambda}{2} \exp(-\frac{\lambda}{2}\omega_{ii}) \right\} 1_{\boldsymbol{\Omega} \in M^+} \mathrm{d}(\omega_{ij})_{i \le j} \\ &< \int \prod_{i<j} \left\{ \frac{1}{\sqrt{2\pi\tau_{ij}}} \exp(-\frac{\omega_{ij}^2}{2\tau_{ij}}) \right\} \prod_{i=1}^{p} \left\{ \frac{\lambda}{2} \exp(-\frac{\lambda}{2}\omega_{ii}) \right\} \mathrm{d}(\omega_{ij})_{i \le j} = 1. \end{aligned}$$

The intractable terms $C_{\boldsymbol{\tau}}$ in (3) and (4) cancel out so that the marginal distribution of the ω_{ij}'s follows (2). This hierarchical representation (3) and (4) is further exploited to construct a data-augmented block Gibbs sampling algorithm in Section 2.4. Our representation is related to the "shadow" prior of Liechty et al. (2004, 2009), although the "shadow" prior is motivated by the acceleration of the posterior computation rather than the introduction of new marginal priors. In particular, the shadow prior uses an additional level in the probability model to move the constraints on a particular parameter to another level, and to reduce the computational burden due to the intractable normalizing constant. It can be seen that our prior corresponds to a special case of the shadow prior when the variance parameter of the shadow prior is zero.

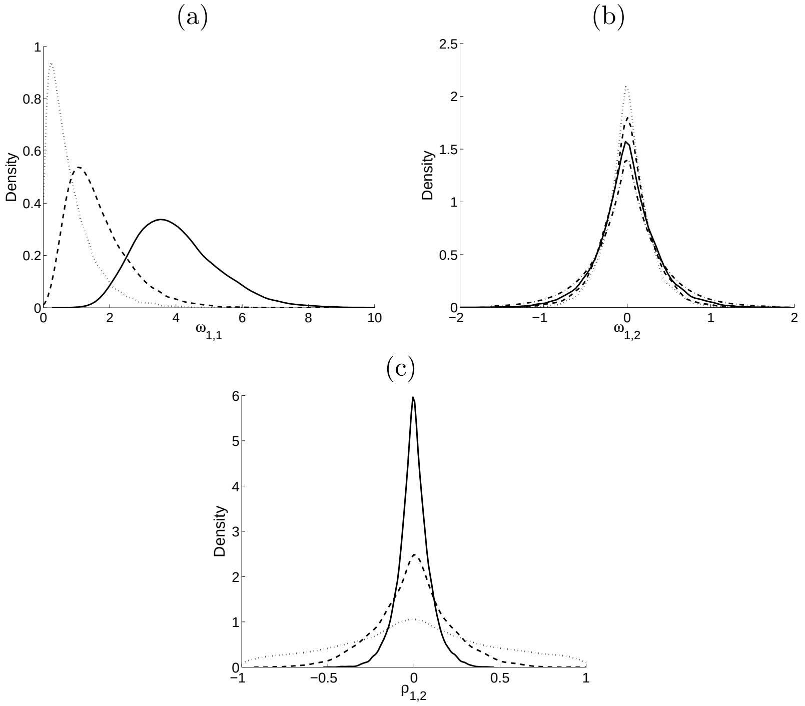


Figure 1: Marginal prior densities for diagonal (a), off-diagonal (b), and partial correlation (c) when $\lambda = 3$ and p =2 (dotted), 10 (dashed) and 50 (solid). For (b), double-exponential with $\lambda = 3$ is also shown (dashdot).

2.2 Bayesian inference on the covariance matrix

Bayes estimators of the covariance and precision matrices can be derived through a decision theoretic approach. For example, consider Stein's loss function for the covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$:

$$L_1(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \mathbf{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - \log\{\det(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1})\} - p \tag{5}$$

where $\hat{\mathbf{\Sigma}}$ denotes the estimator of $\mathbf{\Sigma}$. Then, the corresponding Bayes estimator derived by Yang and Berger (1994) is

$$\hat{\mathbf{\Sigma}}_1 = E(\mathbf{\Sigma}^{-1} \mid \mathbf{Y})^{-1}, \tag{6}$$

which can be estimated from a Monte Carlo sample from the posterior distribution of $\mathbf{\Omega}$. The standard graphical lasso estimator is sometimes referred to as a Bayes estimator because the posterior mode is the limit of a sequence of Bayes estimators under a sequence of loss functions (Hans 2009). Note that the covariance matrix is positive definite and also typically of high dimension. This suggests that the mode estimates and other Bayes estimates can be substantially different.

To investigate the difference, we consider a subset of $n = 10$ observations from the flow cytometry dataset of Friedman et al. (2008), which consists of $p = 11$ proteins and a total of 7466 cells. The sample size was chosen to be relatively small so that the posterior distribution is not a peak. We compare the posterior mode and the posterior mean estimates of $\mathbf{\Omega}$ which are produced by the standard graphical lasso procedure and the Bayesian graphical lasso procedure under Stein's loss, respectively. Figure 2 displays the path of the mean and mode estimates of two entries of $\mathbf{\Omega}$ together with their corresponding 95% credible intervals. Each estimate is plotted as a function of its relative L_1 norm defined as the ratio of the L_1 norm of the graphical lasso estimate of $\mathbf{\Omega}$ over the L_1 norm of the sample precision matrix estimate. Evidently, the width of the credible intervals increases as the relative L_1 norm increases. This important feature of the estimation uncertainty associated with λ is not taken into account by the point estimate of the graphical lasso. Furthermore, the posterior mean estimate is generally larger in absolute value than the posterior mode estimate; the majority of the posterior mass is generally far away from the mode.

2.3 Bayesian inference on graphical structures

The classical graphical lasso procedure is able to produce possible $\hat{\omega}_{ij} = 0$ for $i \neq j$ in the maximizer of (1), thus providing a method for graphical model determination. The Bayesian graphical lasso places zero probability on the event $\{\omega_{ij} = 0\}$, hence has zero posterior probability on the event $\{\omega_{ij} = 0\}$. When fully Bayesian posterior inferences about the event such as $\{\omega_{ij} = 0\}$ are desired, positive prior mass must be allocated to these events. This requires discrete and continuous mixture prior distributions such as the popular G-Wishart prior (Dawid and Lauritzen 1993; Roverato 2000) for $\mathbf{\Omega}$. However, under our absolutely continuous graphical lasso priors, the Bayesian

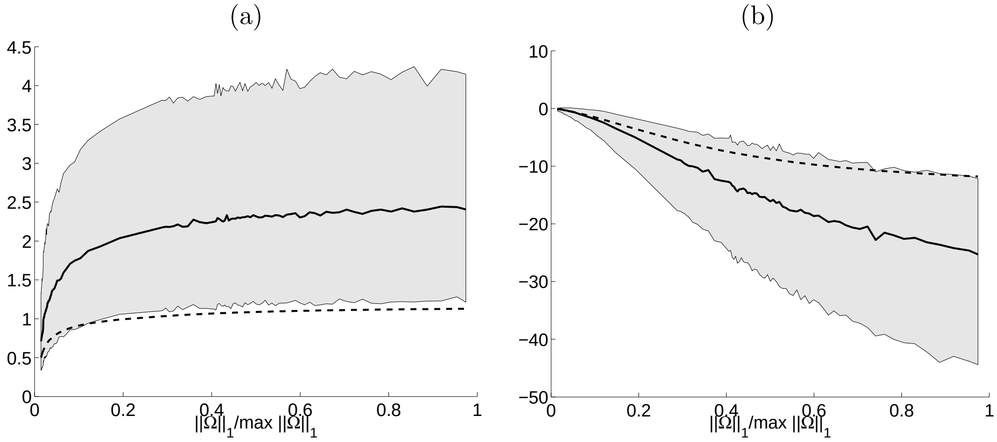


Figure 2: The graphical lasso and Bayesian graphical lasso estimates of one diagonal (a) element and one off-diagonal (b) element of $\mathbf{\Omega}$ as λ varies. Estimates are plotted versus the relative L_1 norm. For the Bayesian graphical lasso, the 95% credible intervals (light gray band), and the posterior mean (solid line within the light gray band) are shown. For the graphical lasso, the posterior mode (dashed line) is shown.

approaches need some heuristic treatments for graphical model determination. We describe a thresholding approach which is recommended by Carvalho et al. (2010) for classification under absolutely continuous priors.

Recall that, under the discrete and continuous mixture prior, the Bayesian posterior mean estimator of the partial correlation of ρ_{ij} is

$$\hat{\rho}_{ij} = \pi_{ij} E_g(\rho_{ij} \mid \mathbf{Y})$$

where π_{ij} is the posterior probability of the event $\{\omega_{ij} \neq 0\}$ and g is the continuous prior distribution for non-zero ω_{ij}. Here π_{ij} has dual roles: Posterior edge inclusion probability that forms the basis for graphical model determination and the amount of shrinkage that is performed on ρ_{ij}. Now, consider the graphical lasso prior which also shrinks ρ_{ij} towards zero. Its posterior mean estimator $\tilde{\rho}_{ij}$ of ρ_{ij} can be written as

$$\tilde{\rho}_{ij} = \tilde{\pi}_{ij} E_g(\rho_{ij} \mid \mathbf{Y}),$$

where $\tilde{\pi}_{ij}$ is the amount of shrinkage applied by the graphical lasso prior on $E_g(\rho_{ij} \mid \mathbf{Y})$. By linking the two shrinkage parameters, $\hat{\pi}_{ij}$ and $\tilde{\pi}_{ij}$, we may claim the event $\{\omega_{ij} \neq 0\}$ if and only if

$$\tilde{\pi}_{ij} = \frac{\tilde{\rho}_{ij}}{E_g(\rho_{ij} \mid \mathbf{Y})} > 0.5,$$

where $\tilde{\rho}_{ij}$ is the posterior sample mean estimate of the partial correlation under graphical lasso priors. As for the choice of g, we use the standard conjugate Wishart prior, $\boldsymbol{W}(3, \boldsymbol{I}_p)$ (Jones et al. 2005) in the simulation study of Section 4.

2.4 A data-augmented block Gibbs sampling scheme

We propose a block Gibbs sampler for posterior computation using the hierarchical representations in (3) and (4). The data-augmented target distribution can be expressed as follows:

$$p(\boldsymbol{\Omega}, \boldsymbol{\tau} \mid \mathbf{Y}, \lambda) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\{-\mathbf{tr}(\frac{1}{2}\boldsymbol{S}\boldsymbol{\Omega})\} \prod_{i<j} \Big\{ \tau_{ij}^{-\frac{1}{2}} \exp(-\frac{\omega_{ij}^2}{2\tau_{ij}}) \exp(-\frac{\lambda^2}{2}\tau_{ij}) \Big\} \times \prod_{i=1}^{p} \Big\{ \exp(-\frac{\lambda}{2}\omega_{ii}) \Big\} 1_{\boldsymbol{\Omega} \in M^+}. \quad (7)$$

Marginally, the original model in (2) is maintained. Note that, in (7), only the parameters in $\boldsymbol{\Omega}$ are constrained to be positive definite. The parameters in $\boldsymbol{\tau}$ are unrestricted. At first glance, the conditional distributions in (7) for subsets of $\boldsymbol{\Omega}$ are not standard distributions. However, an efficient block Gibbs sampler actually exists for (7) after appropriate reparametrization.

We show how to update $\boldsymbol{\Omega}$ one column and row at a time. Without loss of generality, we focus on the last column and row. Let $\boldsymbol{\Upsilon} = (\tau_{ij})$ be a $p \times p$ symmetric matrix with zeros in the diagonal entries and $\boldsymbol{\tau}$ in the upper diagonal entries. Partition $\boldsymbol{\Omega}, \boldsymbol{S}$ and $\boldsymbol{\Upsilon}$ as follows:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11}, \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}', \omega_{22} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11}, \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}', s_{22} \end{pmatrix}, \quad \boldsymbol{\Upsilon} = \begin{pmatrix} \boldsymbol{\Upsilon}_{11}, \boldsymbol{\tau}_{12} \\ \boldsymbol{\tau}_{12}', 0 \end{pmatrix}. \quad (8)$$

From (7), the conditional distribution of the last column in $\boldsymbol{\Omega}$ is

$$p(\boldsymbol{\omega}_{12}, \omega_{22} \mid \boldsymbol{\Omega}_{11}, \Upsilon, \mathbf{Y}, \lambda) \propto (\omega_{22} - \boldsymbol{\omega}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12})^{\frac{n}{2}} \times \exp\big[-\frac{1}{2}\{\boldsymbol{\omega}_{12}'\boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\omega}_{12} + 2\boldsymbol{s}_{12}'\boldsymbol{\omega}_{12} + (s_{22}+\lambda)\omega_{22}\}\big],$$

where $\boldsymbol{D}_{\boldsymbol{\tau}} = \mathbf{diag}(\boldsymbol{\tau}_{12})$. Make a change of variables

$$(\boldsymbol{\omega}_{12}, \omega_{22}) \to (\boldsymbol{\beta} = \boldsymbol{\omega}_{12}, \gamma = \omega_{22} - \boldsymbol{\omega}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}).$$

Then the Jacobian is a constant not involving $(\boldsymbol{\beta}, \gamma)$ and so

$$p(\boldsymbol{\beta}, \gamma \mid \boldsymbol{\Omega}_{11}, \Upsilon, \mathbf{Y}, \lambda) \propto \gamma^{\frac{n}{2}} \exp(-\frac{s_{22}+\lambda}{2}\gamma) \times \exp\Big(-\frac{1}{2}\big[\boldsymbol{\beta}'\{\boldsymbol{D}_{\boldsymbol{\tau}}^{-1} + (s_{22}+\lambda)\boldsymbol{\Omega}_{11}^{-1}\}\boldsymbol{\beta} + 2\boldsymbol{s}_{12}'\boldsymbol{\beta}\big]\Big).$$

This implies that:

$$(\gamma, \boldsymbol{\beta}) \mid (\boldsymbol{\Omega}_{11}, \Upsilon, \mathbf{Y}, \lambda) \sim \mathrm{G}\textsc{a}(\frac{n}{2}+1, \frac{s_{22}+\lambda}{2})\mathrm{N}(-\mathbf{C}\boldsymbol{s}_{21}, \mathbf{C}),$$

where $\text{Ga}(a, b)$ represents a gamma distribution with a shape parameter a and scale parameter b, and $\mathbf{C} = \{(s_{22} + \lambda)\boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\}^{-1}$. This leads to a block Gibbs procedure for iteratively sampling one column in $\boldsymbol{\Omega}$ at a time. Note that the positive definite constraint on $\boldsymbol{\Omega}$ is maintained in each iteration because of $\gamma > 0$. To see this, suppose at iteration t, the current sample, denoted by $\boldsymbol{\Omega}^{(t)}$, is positive definite. This is equivalent to the condition that all of its p leading principal minors are positive. Now, after updating the last column and row of $\boldsymbol{\Omega}$ using the above procedure, the new sample, denoted by $\boldsymbol{\Omega}^{(t+1)}$, has the same leading principal minors as $\boldsymbol{\Omega}^{(t)}$ except for the last one which is of order p. Clearly, this last leading principal minor is $\det(\boldsymbol{\Omega}^{(t+1)}) = \gamma \det(\boldsymbol{\Omega}_{11}^{(t)})$, where $\det(\boldsymbol{\Omega}_{11}^{(t)})$ is the $(p-1)$-th leading principal minor of $\boldsymbol{\Omega}^{(t)}$ and is positive. Therefore, $\{\gamma > 0\}$ implies that $\det(\boldsymbol{\Omega}^{(t+1)}) > 0$ and so all p leading principal minors of $\boldsymbol{\Omega}^{(t+1)}$ are positive, which yields that $\boldsymbol{\Omega}^{(t+1)}$ is positive definite.

To update $\boldsymbol{\tau}$ in (7), the conditional posterior distributions of the $1/\tau_{ij}$'s are clearly independently inverse Gaussian, $\text{Inv-Gau}(\mu', \lambda')$, with parameters

$$\mu' = \sqrt{(\lambda^2/\omega_{ij}^2)}, \quad \lambda' = \lambda^2,$$

in the inverse Gaussian density:

$$p(x) = (\frac{\lambda'}{2\pi x^3})^{1/2} \exp\left\{\frac{-\lambda'(x-\mu')^2}{2(\mu')^2 x}\right\}, \quad x > 0.$$

With this elaboration, we summarize the block Gibbs sampler as follows:

Block Gibbs sampler. Given the current value $\boldsymbol{\Omega} \in M^+$ and $\boldsymbol{\tau}$

1. For $i = 1, \ldots, p$,
   (a) Partition $\boldsymbol{\Omega}, \boldsymbol{S}$ and $\boldsymbol{\Upsilon}$ as in (8).
   (b) Sample $\gamma \sim \text{Ga}(n/2 + 1, (s_{22} + \lambda)/2)$ and $\beta \sim \text{N}(-\mathbf{C}\boldsymbol{s}_{21}, \mathbf{C})$ where $\mathbf{C} = \{(s_{22} + \lambda)\boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\}^{-1}$.
   (c) Update $\boldsymbol{\omega}_{21} = \beta, \boldsymbol{\omega}_{12} = \beta'$, $\omega_{22} = \gamma + \beta'\boldsymbol{\Omega}_{11}^{-1}\beta$.
2. For $i < j$, sample $u_{ij} \sim \text{Inv-Gau}(\mu', \lambda')$ where $\mu' = \sqrt{(\lambda^2/\omega_{ij}^2)}$ and $\lambda' = \lambda^2$, and then update $\tau_{ij} = 1/u_{ij}$.

The above data-augmented block Gibbs sampler is not the only Gibbs sampler for fitting model (2). Rather than working with the hierarchical representation (3) and (4), we may also consider the direct representation of the posterior distribution of (2) without relying the on the latent scale $\boldsymbol{\tau}$. Using the Cholesky decomposition of $\boldsymbol{\Omega}$, we have explored an alternative Gibbs sampler. The Cholesky-based Gibbs sampler updates one element at a time and is thus far less efficient than the block Gibbs sampler, although it does not use the latent variables of $\boldsymbol{\tau}$. The details of the Cholesky-based Gibbs sampler are given in supplement, together with instructions for the free MATLAB routines for implementing these two Gibbs samplers.

2.5 Choosing the shrinkage parameter λ

The graphical lasso requires the selection of the shrinkage parameter λ. Typically one can estimate this parameter by selecting a value for λ that maximizes the normal likelihood on the validation data with $\boldsymbol{\Omega}$ estimated from the training data, or by cross-validation (Friedman et al. 2008; Rothman et al. 2008). In the Bayesian framework, we can choose λ by placing an appropriate hyperprior on it and then extending the Markov chain Monte Carlo sampler to sample from the full conditional distribution of λ. This fully Bayesian method for choosing λ has been successfully applied in the context of Bayesian lasso regression models (Park and Casella 2008; Kyung et al. 2010; Hans 2009). In the graphical model, the positive definite constraint on the support of $\boldsymbol{\Omega}$ gives rise to complications in the computation of the prior normalization constant in (2). In the simplest case where a single λ is used for all elements of $\boldsymbol{\Omega}$, the prior normalizing constant in (2) can be computed as

$$\begin{aligned} C &= \int_{\boldsymbol{\Omega}\in M^+} \prod_{i<j} \Big\{ \mathrm{DE}(\omega_{ij} \mid \lambda) \Big\} \prod_{i=1}^{p} \Big\{ \textsc{Exp}(\omega_{ii} \mid \frac{\lambda}{2}) \Big\} \mathrm{d}\boldsymbol{\Omega} \\ &= \int_{\tilde{\boldsymbol{\Omega}}\in M^+} \prod_{i<j} \Big\{ \mathrm{DE}(\tilde{\omega}_{ij} \mid 1) \Big\} \prod_{i=1}^{p} \Big\{ \textsc{Exp}(\tilde{\omega}_{ii} \mid \frac{1}{2}) \Big\} \mathrm{d}\tilde{\boldsymbol{\Omega}}, \end{aligned} \tag{9}$$

where the last equality holds after applying the substitution $\tilde{\boldsymbol{\Omega}} = \lambda\boldsymbol{\Omega}$ and noticing that $\{\tilde{\boldsymbol{\Omega}} : \tilde{\boldsymbol{\Omega}} \in M^+\} = \{\boldsymbol{\Omega} : \boldsymbol{\Omega} \in M^+\}$ for $\lambda > 0$. Hence C is a constant term not involving λ, although it is unknown and intractable. We can then assign a gamma prior $\lambda \sim \textsc{Ga}(r, s)$, which leads to the conditional posterior $\lambda \sim \textsc{Ga}(r + p(p+1)/2, s + ||\boldsymbol{\Omega}||_1/2)$. However, the normalizing constant C will depend on λ_{ij} when we allow different λ_{ij}'s for different ω_{ij}'s, or even when we only penalize the off-diagonal elements of $\boldsymbol{\Omega}$. In these situations, the use of the standard priors on λ will complicate the posterior simulation because of the evaluation of this normalizing constant. Techniques such as those used in equations (3) and (4) can be useful. In Section 3, we illustrate a set of prior distributions for λ_{ij} when we allow different λ_{ij} for different ω_{ij}.

2.6 Computational speed and scaling experiments

We evaluated the computational speed and the scalability of the above block Gibbs sampler. We used subsets of monthly stock return data from a set of $n = 60$ samples on up to $p = 200$ stocks randomly selected from the population of domestic commonly traded stocks in the New York Stock Exchange. We first standardized the data and then applied the sampler under the hyperprior $\lambda \sim \textsc{Ga}(1,0.01)$. All the chains were initialized at the identity matrix. All computations were implemented on a six-core CPU 3.33GHz desktop running CentOS 5.0 Unix using MATLAB.

For each such data set, we measured the time it took the block Gibbs sampler to generate all of the entries of $\boldsymbol{\Omega}$, which we called one iteration. Figure 3 displays the number of minutes required to compute 1000 iterations versus p. The algorithm took about 1.2 and 9 minutes to generate 1000 iterations for $p = 100$ and 200. In addition, we also used the Markov chain Monte Carlo (MCMC) output to perform convergence diagnostics by calculating the inefficiency factor $1+2\sum_{k=1}^{\infty}\rho(k)$ where $\rho(k)$ is the sample autocorrelation at lag k. We used 3000 samples after 1000 burn-ins and 500 lags in the estimation of the inefficiency factors. The median inefficiency factor among all of the elements of $\boldsymbol{\Omega}$ was 1.1. This suggests that the MCMC mixes quite well. In summary, the block Gibbs sampler appears to be highly efficient because of the fast column-wise updating.

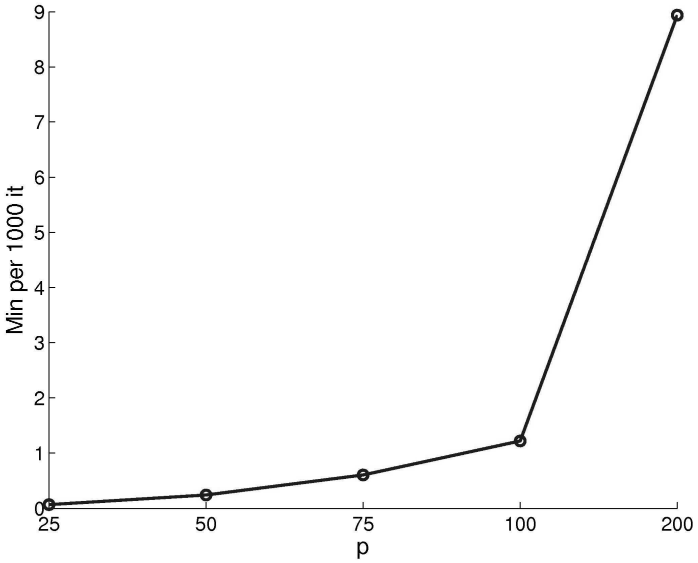
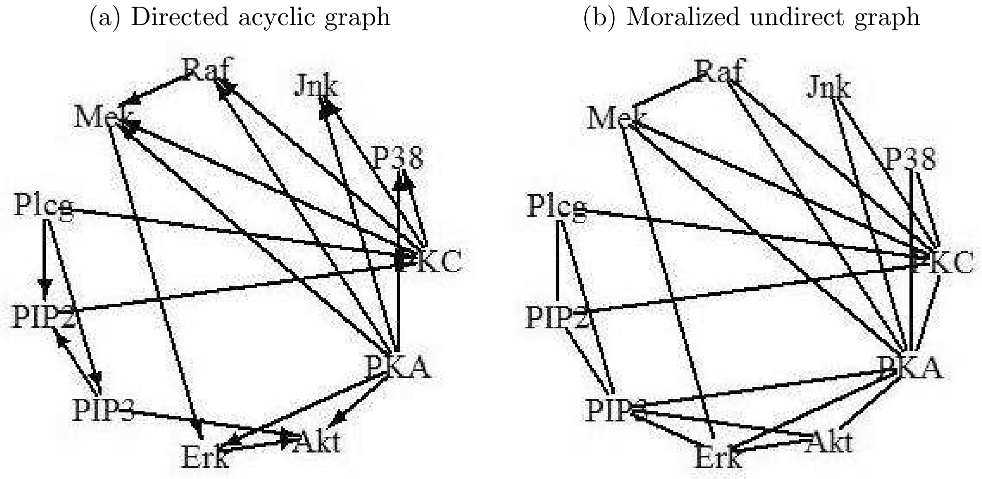


Figure 3: Computational cost as a function of p for the block Gibbs sampler.

2.7 Cell signaling example

The flow cytometry dataset of Friedman et al. (2008), described in Section 2.2, is used here to compare the Bayesian graphical lasso with the standard graphical lasso. The data consist of $p = 11$ proteins and $n = 7466$ cells. Sachs et al. (2005) fit a directed acyclic graph to the data. Figure 4 (a) and (b) display their inferred network and the moralized undirected graph, respectively. Friedman et al. (2008) applied the graphical lasso to these data to produce a set of undirected graphs for different values of the penalty parameter ρ. We applied the Bayesian graphical lasso and used the MCMC outputs from 10000 iterations after 5000 burn-ins.

We compare the parameter estimates from the Bayesian graphical lasso and the standard graphical lasso along the solution path. Figure 5 (a) and (b) show the standard graphical lasso estimates and the posterior mean estimates, respectively, of 10 out of all of the 55 off-diagonal ω_{ij}'s evolving as a function of their relative L_1 norm. They are remarkably similar in their values and patterns largely because of the large number of observations and the small number of variables. Figure 6 displays the 95% credible intervals for the zero off-diagonal elements estimated by the graphical lasso at two different values of λ. The Bayesian 95% credible intervals cover the zero point for all of the cases. However, the width of the credible interval appears to vary a great deal between and within each λ. Moreover, many of these credible intervals are not symmetric about zero.


(a) Directed acyclic graph

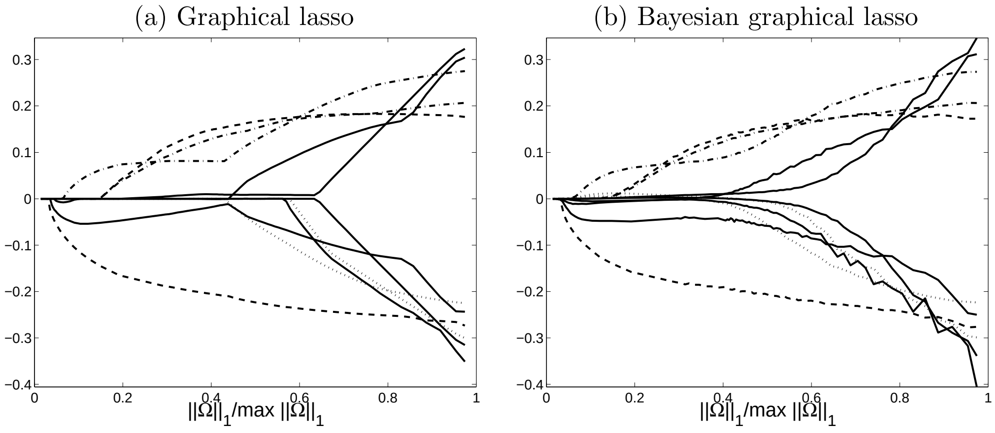
(b) Moralized undirect graph

Figure 4: Directed acyclic graph (a), the moralized undirected graph (b) in the cell-signaling data example from Sachs et al. (2005).

When modeling the shrinkage parameter λ, Friedman et al. (2008) reported choosing the unregularized model with a 10-fold cross-validation. We used the Monte Carlo simulation to estimate the λ with the hyperparameters $r = 1$ and s =0.01. The posterior median for λ was approximately 0.35 and the 95% posterior credible interval for λ was approximately (0.28,0.45).

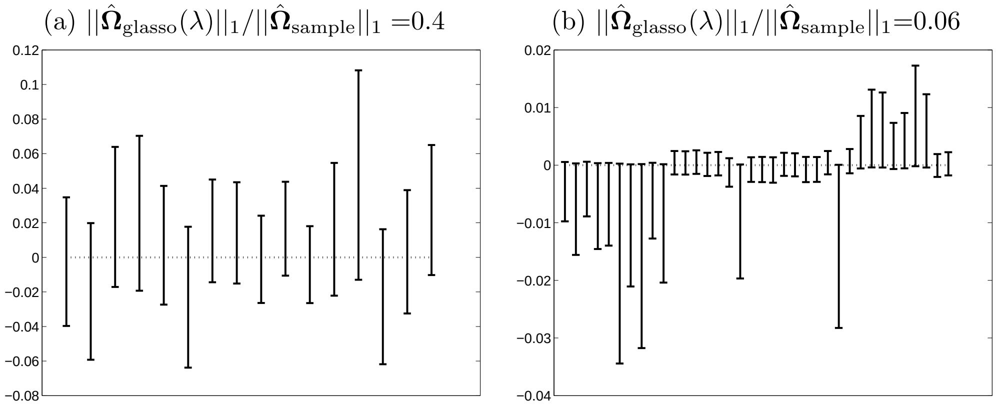
3 Extension

The double exponential prior in (2) shrinks off-diagonal elements of $\mathbf{\Omega}$ towards zero, however, it has a few well-known limitations: It may over-shrink large coefficients but under-shrink small ones. In the regression context, such properties of the double exponential prior have been well studied and many alternative priors have been proposed (e.g., Carvalho et al. 2010; Griffin and Brown 2010; Li and Lin 2010). In our graphical model framework, the hierarchical structure and the block Gibbs sampler in Section 2 allow us to consider generalizations of the graphical lasso prior to overcome its shortcomings.


(a) Graphical lasso

(b) Bayesian graphical lasso

Figure 5: The graphical lasso (a) and the Bayesian graphical lasso (b) estimates of 10 off-diagonal ω_{ij}'s as λ varies. The estimates are plotted versus the relative L_1 norm.


(a) $||\hat{\boldsymbol{\Omega}}_{\text{glasso}}(\lambda)||_1/||\hat{\boldsymbol{\Omega}}_{\text{sample}}||_1$ =0.4

(b) $||\hat{\boldsymbol{\Omega}}_{\text{glasso}}(\lambda)||_1/||\hat{\boldsymbol{\Omega}}_{\text{sample}}||_1$=0.06

Figure 6: Posterior 95% equal-tailed credible intervals with two different values of λ for the zero off-diagonal elements.

Before describing one generalization of the Bayesian graphical lasso, we briefly review two frequentist alternatives to the graphical lasso procedure: The adaptive graphical lasso and the graphical, smoothly clipped absolute deviation (SCAD) method (Fan et al. 2009). These two alternatives are based on the general setting of the following penalized likelihood:

$$\log(\det \boldsymbol{\Omega}) - \mathbf{tr}(\frac{\boldsymbol{S}}{n}\boldsymbol{\Omega}) - \sum_{1\le i\le p}\sum_{1\le j\le p} p_{\lambda_{ij}}(|\omega_{ij}|), \tag{10}$$

where $p(\cdot)$ is the generic penalty function on each element, and λ_{ij} is the tuning parameter for the (i,j)-element of $\boldsymbol{\Omega}$.

The adaptive graphical lasso uses the following penalized likelihood:

$$\log(\det \boldsymbol{\Omega}) - \mathbf{tr}(\frac{\boldsymbol{S}}{n}\boldsymbol{\Omega}) - \lambda \sum_{1\le i\le p}\sum_{1\le j\le p} w_{ij}|\omega_{ij}|, \tag{11}$$

where the adaptive weights are $w_{ij} = 1/|\tilde{\omega}_{ij}|^{\alpha}$ for some $\alpha > 0$. Fan et al. (2009) recommended to fix $\alpha = 0.5$ and the weighting matrix $\{\tilde{\omega}_{ij}\}$ to be the inverse of the sample covariance matrix for the case $p < n$ or the graphical lasso estimates of Ω for the case $p \ge n$, and λ to be chosen by cross-validation.

The graphical SCAD uses a SCAD penalty for $p(\cdot)$ whose first order derivative is given by:

$$p'_{\lambda_{ij}}(|\omega_{ij}|) = \lambda\left\{1_{\{|x|\le\lambda\}} + \frac{(a\lambda - |x|)_+}{(a-1)\lambda}1_{\{|x|>\lambda\}}\right\}, \tag{12}$$

where a and λ are two tuning parameters. Fan et al. (2009) recommended to fix $a = 3.7$ and choose λ via cross-validation.

Now, we describe a Bayesian analog of the adaptive graphical lasso (11):

$$\begin{aligned} p(\mathbf{y}_i \mid \boldsymbol{\Omega}) &= \mathrm{N}(\mathbf{y}_i \mid 0, \boldsymbol{\Omega}^{-1}) \quad (i = 1,\ldots,n), \\ p(\boldsymbol{\Omega} \mid \{\lambda_{ij}\}_{i\le j}) &= C^{-1}_{\{\lambda_{ij}\}_{i\le j}} \prod_{i<j}\Big\{\mathrm{DE}(\omega_{ij} \mid \lambda_{ij})\Big\}\prod_{i=1}^{p}\Big\{\textsc{Exp}(\omega_{ii} \mid \frac{\lambda_{ii}}{2})\Big\}1_{\boldsymbol{\Omega}\in M^+}, \\ p(\{\lambda_{ij}\}_{i<j} \mid \{\lambda_{ii}\}_{i=1}^{p}) &\propto C_{\{\lambda_{ij}\}_{i\le j}} \prod_{i<j}\textsc{Ga}(r,s), \end{aligned} \tag{13}$$

where $C_{\{\lambda_{ij}\}_{i\le j}}$ is the intractable normalizing constant and $\{\lambda_{ii}\}_{i=1}^{p}$ for the diagonal elements are hyperparameters. The two terms of $C_{\{\lambda_{ij}\}_{i\le j}}$ in (13) cancel out to ease the computations when updating λ_{ij}. Model fitting is straightforward using the block Gibbs sampler described in (2.4) with the modification to update different λ_{ij} from different gamma distributions.

In comparison with the Bayesian graphical lasso in (2), the Bayesian adaptive graphical lasso places different shrinkage parameters λ_{ij} on different off-diagonal elements ω_{ij}. When compared with frequentist adaptive graphical lasso (11), the Bayesian counterpart automatically chooses the amount of shrinkage according to the current value of

ω_{ij} as follows. Conditional on $\mathbf{\Omega}$, model (13) gives

$$\lambda_{ij} \mid \mathbf{\Omega} \sim \textsc{Ga}(1+r, |\omega_{ij}|+s),$$

which implies the conditional mean of λ_{ij} is $(1+r)/(|\omega_{ij}|+s)$. If the current ω_{ij} is small (large), then in the updating stage, the shrinkage parameter λ_{ij} will tend to be large (small).

The choice of hyperparameters (r, s) is fundamentally important in the performance of the Bayesian adaptive graphical lasso. To encourage the adaptiveness of λ_{ij} to ω_{ij}, the hyperparameter s must be small relative to ω_{ij}. To see this, suppose a current value of $\omega_{ij} = 0.01$ and a hyperparameter $s = 0.1$ which is small in raw scale but large relative to ω_{ij}, then in the updating stage, λ_{ij} will be around $(1+r)/0.11$ which essentially ignores the small value of $\omega_{ij} = 0.01$ and underestimates λ_{ij} by a factor of about 10 as compared with $1/|\omega_{ij}|$. Thus, a large s relative to ω_{ij} will not enjoy the adaptiveness. In the simulation study, we used $r = 10^{-2}$ to represent that the prior degrees of freedom is 0.01 when the sample size of each individual ω_{ij} is one, and chose $s = 10^{-6}$ to allow λ_{ij} to be adaptive to small ω_{ij}. The simulation study in Section 4 shows that this choice of prior hyperparameters provides excellent performance.

The Bayesian adaptive lasso has another interpretation under marginalization. Upon integrating over $(\lambda_{ij})_{i<j}$ in (13) , the marginal prior for Ω is

$$\begin{aligned} p(\mathbf{\Omega} \mid a, b, \{\lambda_{ii}\}_{i=1}^{p}) \propto & \Big[\int \prod_{i<j} \Big\{ \mathrm{DE}(\omega_{ij} \mid \lambda_{ij}) \textsc{Ga}(\lambda_{ij} \mid s, t) \Big\} \mathrm{d}(\lambda_{ij})_{i<j} \Big] \\ & \times \prod_{i=1}^{p} \Big\{ \textsc{Exp}(\omega_{ii} \mid \frac{\lambda_{ii}}{2}) \Big\} 1_{\mathbf{\Omega} \in M^+} \\ = & \prod_{i<j} \Big\{ \mathrm{GDP}(\lambda_{ij} \mid \xi = s/r, \alpha = r) \Big\} \prod_{i=1}^{p} \Big\{ \textsc{Exp}(\omega_{ii} \mid \frac{\lambda_{ii}}{2}) \Big\} 1_{\mathbf{\Omega} \in M^+}. \end{aligned} \tag{14}$$

where $\mathrm{GDP}(x \mid \xi, \alpha)$ denotes the generalized double Pareto distribution with a density function

$$p(x) = \frac{1}{2\xi}(1 + \frac{|x|}{\alpha\xi})^{-(1+\alpha)}.$$

Armagan et al. (2012) showed that the generalized double Pareto distribution is a useful shrinkage prior for linear regression models. Our Bayesian adaptive graphical lasso can be seen as a Bayesian GDP graphical lasso with fixed parameters ξ and α.

4 Simulated example

This simulation experiment was designed to test the performance of the Bayesian and the frequentist graphical lassos in terms of parameter estimation and structure learning. We considered 6 different models in our simulation:

- *Model 1:* An AR(1) model with $\sigma_{ij} = 0.7^{|i-j|}$.
- *Model 2:* An AR(2) model $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$ and $\omega_{i,i-2} = \omega_{i-2,i} = 0.25$.
- *Model 3:* A block model with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$ for $1 \leq i \neq j \leq p/2$, $\sigma_{ij} = 0.5$ for $p/2 + 1 \leq i \neq j \leq 10$ and $\sigma_{ij} = 0$ otherwise.
- *Model 4:* A star model with every node connected to the first node, with $\omega_{ii} = 1, \omega_{1,i} = \omega_{i,1} = 0.1$ and $\omega_{ij} = 0$ otherwise.
- *Model 5:* A circle model with $\omega_{ii} = 2, \omega_{i,i-1} = \omega_{i-1,i} = 1$, and $\omega_{1,p} = \omega_{p,1} = 0.9$.
- *Model 6:* A full model with $\omega_{ii} = 2$ and $\omega_{ij} = 1$ for $i \neq j$.

For each model, we generated samples of size $n = 50$ and dimension $p = 30$ or size $n = 200$ and dimension $p = 100$. For each generated sample, we fit the graphical lasso (1), the adaptive graphical lasso (11), the graphical SCAD (12), the Bayesian graphical lasso (2) with hyperparameters $r = 1$ and s =0.01 for the prior distribution of λ, and the Bayesian adaptive graphical lasso (13) with $r = 10^{-2}$ and $s = 10^{-6}$ for the prior distributions of λ_{ij} for $i < j$ and $\lambda_{ii} = 1$ for $i = 1, \ldots, p$. The tuning parameters of the three frequentist lassos were chosen by 10-fold cross-validation. The two Bayesian estimates were based on 10000 iterations of the Monte Carlo sampler after 5000 burn-in iterations.

To assess the performance of the covariance matrix estimation, we calculated Stein's loss in (5) using the Bayes estimator (6) for the two Bayesian procedures and the mode estimator for the three frequentist procedures. Table 1 reports the median and the standard error of Stein's loss for $p = 30$ and 100 in models 1–6 based on 50 replications. Under each scenario, the best two performances are boldfaced in the tables. Three things are worth noting from Table 1. First, except for the star case, the Bayesian adaptive graphical lasso always ranks among the best two methods. Second, except for the star and the block cases, the frequentist graphical lasso is always among the worst two methods perhaps because it over-shrinks large coefficients. The graphical lasso performs well in the star and the block cases perhaps because all signals are small in these two cases – the star matrix has a non-zero partial correlation of 0.1 and the block matrix has a non-zero partial correlation of -0.067 and -0.02 for $p = 30$ and $p = 100$ respectively. Third, the Bayesian procedures seem to have the advantage of reducing the standard errors. This is not surprising as the estimates of the covariance matrix in the Bayesian procedures are not based on a single and fixed value of penalty parameter but rather are based on all of them, which leads to a robust estimation of the covariance matrix.

To assess the performance of the graphical structure learning, we computed specificity, sensitivity and Matthews Correlation Coefficients (MCC), which have been used

	AR(1)	AR(2)	Block	Star	Circle	Full
			p = 30			
Glas	4.50(0.52)	7.05(0.90)	3.45(0.45)	1.67(0.30)	5.31(0.67)	31.43(2.71)
Adap	**3.58(0.52)**	5.62(0.83)	4.00(0.42)	**1.62(0.20)**	4.23(0.56)	28.53(3.13)
Scad	5.22(1.46)	5.51(0.71)	7.44(0.60)	**1.66(0.56)**	5.40(1.34)	26.99(4.54)
Bgla	3.82(0.33)	**4.99(0.31)**	**2.63(0.31)**	2.07(0.30)	**4.10(0.37)**	**15.23(0.51)**
Bada	**3.39(0.49)**	**4.59(0.40)**	**2.80(0.33)**	1.93(0.53)	**3.72(0.62)**	**15.37(1.13)**
			p = 100			
Glas	8.21(0.38)	15.31(0.75)	6.69(0.24)	**2.47(0.22)**	9.49(0.42)	106.73(2.77)
Adap	4.09(0.23)	10.05(0.46)	7.64(0.25)	**2.56(0.37)**	**5.24(0.24)**	101.08(2.90)
Scad	**1.75(0.40)**	**7.80(0.71)**	17.84(0.62)	8.20(0.43)	11.48(0.56)	87.75(4.68)
Bgla	8.92(0.33)	13.94(0.31)	**6.29(0.24)**	4.98(0.24)	9.56(0.28)	**59.42(0.70)**
Bada	**2.95(0.24)**	**5.85(0.34)**	**5.47(0.23)**	3.83(0.26)	**3.01(0.21)**	**70.33(1.35)**

Table 1: Summary of Stein's loss for the different models and different methods based on 50 replications. "Glas" refers to the frequentist graphical lasso with cross-validation;"Adap" refers to the frequentist adaptive graphical lasso with cross-validation; "Scad" refers to the SCAD method with cross-validation; "Bgla" refers to the Bayesian graphical lasso and "Bada" refers to the Bayesian adaptive graphical lasso. The medians are reported here, the standard errors are shown in parentheses, and the best two methods are boldfaced.

in Fan et al. (2009) and are defined as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \tag{15}$$

where TP,TN,FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively. MCC is generally regarded as a balanced measure of the classification because it takes into account TP, TN, FP, and FN. For all three metrics, the larger the values are, the better the classification is. We compared the results from the Bayesian thresholding of Section 2.3 and from the frequentist methods, where we claim $\{\omega_{ij} = 0\}$ if $\hat{\omega}_{ij} < 10^{-3}$ as in Fan et al. (2009). The results, averaged over 50 runs, are reported in Table 2. The standard deviations around the mean are within 10% of the mean values for all results. The Bayesian adaptive graphical lasso tends to provide higher specificity in all cases but lower specification in cases when signal is weak (e.g., block and star). Its overall performance seems to be good as its MCC ranks among the top three methods in all cases. All the other four methods appear to perform well in some cases, and fare badly in other cases. For example, when $p = 100$, the SCAD ranks among the top two for the cases of AR(1) and AR(2) but is in the bottom for the case of block and star.

	AR(1)			AR(2)			Block			Star			Circle		
	SP	SE	MC	SP	SE	MC	SP	SE	MC	SP	SE	MC	SP	SE	MC
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
							p = 30								
Glas	66	100	34	65	82	32	78	67	**46**	91	28	**16**	63	100	32
Adap	78	100	44	75	81	**42**	80	49	31	91	17	05	75	100	41
Scad	82	93	**45**	78	76	**43**	74	36	10	96	03	03	86	100	**53**
Bgla	74	100	38	71	33	05	68	42	10	80	28	04	58	100	30
Bada	96	100	**77**	90	43	33	93	34	**34**	94	14	**07**	95	100	**76**
							p = 100								
Glas	77	100	25	75	97	31	89	44	**37**	95	100	**54**	73	100	23
Adap	90	100	38	85	96	41	90	31	**27**	93	97	**44**	86	100	**33**
Scad	98	100	**74**	90	96	**50**	82	22	5	67	94	18	42	100	12
Bgla	79	100	27	80	36	7	78	27	7	80	92	24	73	100	23
Bada	96	100	**55**	94	90	**56**	93	20	19	91	85	36	97	100	**62**

Table 2: Percentage of specificity (SP%), sensitivity (SE%) and Matthews Correlation Coefficient (MC%) for the different models and different methods based on 50 replications. These three measures are defined in equation (15). The best two methods for MCC are boldfaced.

5 Discussion

The Bayesian graphical lassos are shown to be attractive for estimating covariance matrices for moderately large problems. In comparison with frequentist methods, the Bayesian graphical lassos offer MCMC outputs that can be summarized in any manner a modeler wants at a low computational cost (e.g., a few minutes for $p = 100$). There is also strong empirical evidence that the Bayesian adaptive graphical lasso has excellent performance. In comparison with many other Bayesian shrinkage estimation methods for covariance matrices (for example, Daniels and Kass 1999, 2001; Barnard et al. 2000; Wong et al. 2003; Liechty et al. 2004), our methods use the block Gibbs sampler which requires no tuning and runs quickly. To the best of our knowledge, the only existing permutation-invariant Bayesian method that uses block Gibbs sampler and so can easily scale up to hundreds of variables is the factor model. Thanks to the block Gibbs sampler, the proposed Bayesian graphical lassos can also easily handle hundreds of variables, hence adding a powerful set of new tools to the Bayesian toolbox for large covariance matrix estimation.

On the graphical structure learning, the proposed thresholding approach is ad hoc and lacks the formal Bayesian interpretation. A fully Bayesian treatment of structure learning should place point mass priors on the events $\{\omega_{ij} = 0\}$ to make a posterior inference about sparse structures. Using standard notation, let $\mathbf{\Gamma}$ be a $p(p-1)/2$-vector of edge inclusion indicators where $\gamma_{ij} = 1$ whenever $\omega_{ij} \neq 0$ for $i > j$. The prior distribution of $\mathbf{\Omega}$ for a particular structure $\mathbf{\Gamma}$ can be expressed as

$$p(\mathbf{\Omega} \mid \mathbf{\Gamma}) \quad = \quad C^{-1}_{\mathbf{\Gamma},\lambda} \prod_{\gamma_{ij}=1} \Big\{ \exp(-\lambda|\omega_{ij}|) \Big\} \prod_{i=1}^{p} \Big\{ \exp(-\frac{\lambda}{2}\omega_{ii}) \Big\}, \tag{16}$$

where $C_{\mathbf{\Gamma},\lambda}$ is the normalizing constant depending on $\mathbf{\Gamma}$ and λ. Clearly, posterior inferences about $\mathbf{\Gamma}$ and λ require the evaluation of $C_{\mathbf{\Gamma},\lambda}$ for each setting of $(\mathbf{\Gamma}, \lambda)$. This can be computationally challenging because $C_{\mathbf{\Gamma},\lambda}$ is intractable. A similar problem occurs in the Bayesian Gaussian graphical model literature (Atay-Kayis and Massam 2005; Wang and Li 2012) where the evaluation of the normalizing constant for a given graph is the biggest computational bottleneck. We have begun to explore methods that can potentially eliminate these computational burdens.

The Bayesian graphical lasso can be extended to other lasso-related methods for graphical models. Hierarchical models based on the normal scale mixtures have been well examined in the context of the regression analysis. However, theoretical and empirical studies of these priors in the context of Gaussian graphical models are less known and are needed to increase the flexibility of Bayesian graphical models. While the implied priors can be different, the block Gibbs sampler in this paper can offer ways to apply these methods to graphical models.

Supplementary materials

Computational details for the Cholesky-based Gibbs sampler, and instructions for the MATLAB routines implementing all frequentist and Bayesian procedures used in the paper, are available from the author's the web site of the paper.

References

Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1): pp. 99–102. 869

Armagan, A., Dunson, D., and Lee, J. (2012). "Generalized double Pareto shrinkage." *Statistica Sinica (forthcoming)*. 880

Atay-Kayis, A. and Massam, H. (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models." *Biometrika*, 92: 317–335. 884

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data." *The Journal of Machine Learning Research*, 9: 485–516. 867

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 10(4): 1281–1311. 883

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. 872, 877

Daniels, M. J. and Kass, R. E. (1999). "Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models." *Journal of the American Statistical Association*, 94(448): pp. 1254–1263. 883

— (2001). "Shrinkage Estimators for Covariance Matrices." *Biometrics*, 57(4): 1173–1184. 883

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *Annals of Statistics*, 21: 1272–1317. 871

Fan, J., Feng, Y., and Wu, Y. (2009). "Network exploration via the adaptive LASSO and SCAD penalties." *Annals of Applied Statistics*, 3(2): 521–541. 879, 882

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9(3): 432–441. 867, 871, 875, 876, 877

Griffin, J. E. and Brown, P. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. 877

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). "Joint estimation of multiple graphical models." *Biometrika*, 98(1): 1–15. 867

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 869, 871, 875

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, 20: 388–400. 872

Khondker, Z., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. (2012). "Bayesian covariance lasso." *Statistics and Its Interface (forthcoming)*. 868

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5(2): 369–412. 875

Li, Q. and Lin, N. (2010). "The Bayesian Elastic Net." *Bayesian Analysis*, 5(1): 151–170. 877

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). "Bayesian correlation estimation." *Biometrika*, 91(1): 1–14. 870, 883

Liechty, M. W., Liechty, J. C., and Müller, P. (2009). "The Shadow Prior." *Journal of Computational and Graphical Statistics*, 18(2): 368–383. 870

Marlin, B. M. and Murphy, K. P. (2009). "Sparse Gaussian graphical models with unknown block structure." In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 705–712. New York, NY, USA: ACM. 868

Marlin, B. M., Schmidt, M., and Murphy, K. P. (2009). "Group Sparse Priors for Covariance Estimation." In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 868

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the lasso." *The Annals of Statistics*, 34(3): 1436–1462. 867

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 869, 875

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). "Sparse permutation invariant covariance estimation." *Electronic Journal of Statistics*, 2: 494–515. 867, 875

Roverato, A. (2000). "Cholesky decomposition of a hyper-inverse Wishart matrix." *Biometrika*, 87: 99–112. 871

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data." *Science*, 308(5721): 523–529. 876, 877

Wang, H. and Li, S. Z. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. 884

West, M. (1987). "On Scale Mixtures of Normal Distributions." *Biometrika*, 74(3): pp. 646–648. 869

Wong, F., Carter, C., and Kohn, R. (2003). "Efficient estimation of covariance selection models." *Biometrika*, 90: 809–30. 883

Yang, R. and Berger, J. O. (1994). "Estimation of a Covariance Matrix Using the Reference Prior." *The Annals of Statistics*, 22(3): pp. 1195–1211. 871

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94(1): 19–35. 867

Acknowledgments

The author is grateful for the constructive comments of the editor, associate editor, and two anonymous referees on the original version of this manuscript.