Bayesian Analysis (2019) **14**, Number 2, pp. 449–476

Efficient Bayesian Regularization for Graphical Model Selection

Suprateek Kundu*, Bani K. Mallick†, and Veera Baladandayuthapani‡

Abstract. There has been an intense development in the Bayesian graphical model literature over the past decade; however, most of the existing methods are restricted to moderate dimensions. We propose a novel graphical model selection approach for large dimensional settings where the dimension increases with the sample size, by decoupling model fitting and covariance selection. First, a full model based on a complete graph is fit under a novel class of mixtures of inverse–Wishart priors, which induce shrinkage on the precision matrix under an equivalence with Cholesky-based regularization, while enabling conjugate updates. Subsequently, a post-fitting model selection step uses penalized joint credible regions to perform model selection. This allows our methods to be computationally feasible for large dimensional settings using a combination of straightforward Gibbs samplers and efficient post-fitting inferences. Theoretical guarantees in terms of selection consistency are also established. Simulations show that the proposed approach compares favorably with competing methods, both in terms of accuracy metrics and computation times. We apply this approach to a cancer genomics data example.

Keywords: covariance selection, Cholesky-based regularization, joint penalized credible regions, shrinkage priors, selection consistency.

1 Introduction

Recent technological advances in many scientific disciplines, such as genomics, imaging and environmental sciences, have resulted in datasets with numerous variables. A convenient framework for analyzing and interpreting the relationships between the variables is provided by graphical models, which detect a network of dependencies among a group of p measurements obtained from n samples, denoted by $x_i = (x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n$. Here, we are concerned with Gaussian graphical models for continuous data, that are designed to detect conditional dependency relationships by discovering a pattern of zeros in the inverse covariance or precision matrix, a process typically referred to as covariance selection (Dempster, 1972). Our goal is to propose a novel, flexible and efficient Bayesian covariance selection strategy in large dimensional settings (we consider p in several hundreds) which has theoretical guarantees and encouraging numerical performance.

*Department of Biostatistics & Bioinformatics, Emory University, 1518 Clifton Road, Atlanta, Georgia 30322, U.S.A., supratEek.kundu@emory.edu

†Department of Statistics, Texas A&M University, 3143 TAMU, College Station, Texas 77843-3143, U.S.A., bmallick@stat.tamu.edu

‡Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A., veera@mdanderson.org

 https://doi.org/10.1214/17-BA1086

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = (\mathbf{x}_{c1}, \ldots, \mathbf{x}_{cp})$ be the $n \times p$ dimensional data matrix, with subscript c denoting the columns. The cornerstone of Bayesian approaches for Gaussian graphical models has been discrete mixture formulations that specify

$$\mathbf{x}_l \mid \Sigma_G \sim N(\theta, \Sigma_G), \quad \Sigma_G \sim \pi(\Sigma_G \mid G), \quad G \sim \pi(G), \quad l = 1, \ldots, n, \tag{1}$$

where the graph G is defined using a set of nodes or vertices $V^* = \{1, \ldots, p\}$ and an edge set $E = (e_{ij})$ with $e_{ij} = 1$ if and only if the (i, j)th entry of the precision matrix is non-zero. The term discrete mixture comes from the fact that the prior on the covariance in model (1) can be represented as the convex linear combination $\pi(\Sigma) = \sum_{G \in \mathcal{G}} \pi(G)\pi(\Sigma_G \mid G)$, with G having a discrete probability distribution. For a fixed graph G, the support of Σ_G^{-1} is the cone M_G^+, the space of all positive definite matrices having exact zeros for off-diagonals corresponding to absent edges. Here θ denotes the mean which is usually set to zero after centering the measurements.

Some prominent examples of discrete mixture priors include the hyper inverse–Wishart prior (Dawid and Lauritzen, 1993) for the covariance and the G-Wishart prior for the precision (Diaconis and Ylvisaker, 1979; Roverato, 2000; Atay-Kayis and Massam, 2005). The implementation of most of these approaches relies on reversible jump Markov chain Monte Carlo (MCMC) algorithms (Giudici and Green, 1999; Dellaportas et al., 2003; Wong et al., 2003; Wang and West, 2009; Green and Thomas, 2013). For a comparison of model fitting approaches for decomposable graphical models, refer Fitch et al. (2014). These algorithms explore the graph space and subsequently select graphs with high posterior probabilities $Pr(G|X)$ or estimate a graph by including edges having a posterior inclusion probability greater than some threshold. Jones et al. (2005) proposed the shotgun stochastic search algorithm designed to efficiently move toward regions of high posterior probability in the model space using a parallel computing approach, whereas Scott and Carvalho (2008) developed a greedy approach called the feature inclusion search algorithm for decomposable Gaussian graphical models. Recently, Mohammadi and Wit (2015) proposed a trans–dimensional Markov chain Monte Carlo approach based on a continuous–time birth-death process.

As p increases, the cardinality of the graph space increases exponentially, making it computationally intractable if not impossible for many discrete mixture-based approaches to efficiently explore the graph space. This problem is somewhat akin to known difficulties encountered by stochastic search variable selection approaches (George and McCulloch, 1993) in navigating the model space for large dimensional regression settings (Bondell and Reich, 2012; Kundu and Dunson, 2014). However, the problem is far more severe in the context of graphical model estimation, as the graph space (having cardinality $2^{p(p-1)/2}$) explodes far more quickly. As a result the usual discrete mixture-based approaches can fail to discover models with high posterior probabilities, while estimates of the edge-specific posterior inclusion probability are susceptible to instability under finite runs of the Markov chain Monte Carlo, as demonstrated in our simulations. Moreover in large dimensions, the results can be sensitive to the choice of the prior on the graph space. An additional constraint is that the optimal graph is often restricted to lie in the class of decomposable graphs, due to the computationally demanding heuristic approximations required for non-decomposable models (Atay-Kayis and Massam, 2005; Lenkoski and Dobra, 2011).

Motivated by the success of shrinkage methods in Bayesian variable selection, we propose a shrinkage approach for estimating graphical models which bypasses the limitations of discrete mixture approaches and has connections with global-local priors (Carvalho et al., 2009, 2010; Polson and Scott, 2011) in regression settings. The proposed approach decouples model fitting and covariance selection. We first fit the full model based on a complete graph under a class of conjugate shrinkage priors, which is denoted as regularized inverse–Wishart priors hereafter. Our approach is novel in assigning suitable priors on the scale matrix of the inverse–Wishart, which can be marginalized to induce adaptive shrinkage on the elements of the Cholesky factor of the precision matrix, thus leading to a Cholesky-based regularization (Pourahmadi, 1999; Smith and Kohn, 2002; Wu and Pourahmadi, 2003; Frühwirth-Schnatter and Tüchler, 2008). However, unlike the usual Cholesky-based regularization approaches, the proposed prior is order invariant and allows for conjugate updates of the precision matrix, leading to efficient posterior computation.

Although shrinkage priors have elegant properties and are routinely used, they do not immediately provide an automated procedure for model selection. While some valid thresholding approaches are available for sparse covariance matrix estimation (Bickel and Levina, 2008; Cai and Liu, 2012), there is a lack of systematic thresholding approaches to obtain sparse precision matrices in the Bayesian paradigm. However, there has been an increasing interest in continuous shrinkage approaches for Bayesian variable selection, which apply decision theoretic methods to reduce unimportant effect sizes to zero for model selection purposes (Fouskakis et. al, 2009; Bondell and Reich, 2012; Hahn and Carvalho, 2015). This motivates us to propose a decision theoretic approach for graphical model estimation, that uses L_0 penalized joint credible regions to perform neighborhood selection for each node. Our work is distinct compared to the approach in Hahn and Carvalho (2015), who used expected loss functions involving a L_0 penalization as a post-processing decision theoretic step for model selection. We show that the resulting approach achieves neighborhood selection consistency for fixed and increasing dimensions, and also yields precision matrices which are positive definite and asymptotically consistent for fixed dimensions.

In summary, the proposed approach overcomes several difficulties associated with existing Bayesian alternatives: (i) it obviates having to specify prior graph probabilities, which can adversely affect final inferences under mis-specification; (ii) it does not require long runs of Markov chain Monte Carlo to search over the model space, with the computation involving a straightforward fully Gibbs sampler; (iii) it is computationally efficient and feasible for large dimensions; (iv) it is applicable to a broad class of models, including decomposable and non-decomposable graphs; and (v) it attains selection consistency in fixed p and $p_n = o(n)$ settings.

2 Shrinkage priors for precision matrices

2.1 The regularized inverse–Wishart prior

In this section, we propose shrinkage priors on the precision matrix characterized by mixtures of inverse–Wishart priors on the covariance. Without loss of generality we

assume a zero mean model, i.e. set $\theta = 0$ in (1), indicating the data matrix X is appropriately centered. The general construction of the prior can be written as,

$$\mathbf{x}_l \sim N(0, \Sigma),\ \Sigma|D \sim \text{Inverse Wishart}(b, D),\ l = 1, \ldots, n, \tag{2}$$

where $D = \text{diag}(d_1, \ldots, d_p)$, and $d_k \sim \mathcal{G}_k(\bullet)$, with $\mathcal{G}_k(\bullet), k = 1, \ldots, p$, denoting mixing distributions allowing for adaptive shrinkage across different scales. By setting $\mathcal{G}_k(\bullet)$ to different mixing distributions, various types of shrinkage can be obtained.

Model (2) relies on a conjugate inverse–Wishart prior on $\Sigma = \Omega^{-1}$, and varies from the traditional discrete mixture formulation (1), in having a continuous mixture representation for the covariance as $\pi(\Sigma) = \int N(\Sigma \mid D) d\pi(D)$. The traditional model (1) constrains the support of Ω_G to the cone M_G^+ which depends on $G \in \mathcal{G}$, while the continuous mixture prior in (2) has an unconstrained support M^+ (the space of all positive definite matrices). Our choice of the inverse–Wishart formulation (2) is based on both theoretic and computational considerations: (i) it induces a Gaussian distribution on the off-diagonals of Σ^{-1} (Lemma 1), which is a necessary condition in establishing model selection consistency; and (ii) the associated conjugacy allows us to draw posterior samples of Σ^{-1} in an efficient manner even for large dimensions.

Some notations we use hereafter are defined as follows. The covariance matrix is denoted as $\Sigma \equiv \Sigma_p = \left(\begin{smallmatrix} \Sigma_{p-1,\mathbf{11}} & \sigma_{p,\mathbf{21}} \\ \sigma_{p,\mathbf{12}} & \sigma_{p,pp} \end{smallmatrix}\right)$, with $\Sigma^{-1}_{k,\mathbf{11}} \equiv \Omega_k = \left(\begin{smallmatrix} \Omega^{k-1}_{\mathbf{11}} & \omega_{k,\mathbf{21}} \\ \omega_{k,\mathbf{12}} & \omega_{k,kk} \end{smallmatrix}\right)$, where $\Sigma_{p-1,\mathbf{11}}$ denotes the principal minor of dimension $p-1$ derived from the first $p-1$ rows and columns of Σ, and $\Omega^{k-1}_{\mathbf{11}}$ denotes the principal minor of dimension $k-1$ for Ω_k. Let $\omega_{k,ij}$ denote the j-th element in the i-th row of Ω_k.

We now state the following well-known result as a Lemma (Carvalho and Scott, 2009), which serves as a first step toward understanding the regularization properties of the prior in (2). The Lemma captures the distribution of elements in the last row of $\Sigma^{-1} \equiv \Omega_p$ conditional on D. The corresponding result for any row can be adapted in a straightforward manner.

Lemma 1. *For $\Sigma \sim$ Inverse Wishart$(b, diag(d_1, \ldots, d_p))$, we have $\pi(\omega_{p,pp}) = Ga(\omega_{p,pp} \mid b/2; d_p/2)$, and, $\pi(\omega_{p,\mathbf{12}} \mid \omega_{p,pp}) = \prod_{l=1}^{p-1} N(\omega_{p,pl} \mid 0; \omega_{p,pp}/d_l)$.*

Lemma 1 shows us that the precision off-diagonals have a scale mixture representation under a prior on $d_1, \ldots, d_p$, and a careful choice of $\mathcal{G}_k(\bullet)$ is likely to yield a prior on Σ^{-1} with desirable shrinkage properties. We propose the following priors on $d_1, \ldots, d_p$, which achieves joint regularization for all the elements in Σ^{-1} (see Theorem 1)

$$d_k \sim \text{Inverse Gamma}((b+1)/2, \lambda_k^2/2), \quad \lambda_k \sim \text{Ga}(a_{\lambda,k}, b_{\lambda,k}), \quad k = 1, \ldots, p, \tag{3}$$

where b is the degrees of freedom of the inverse–Wishart prior in (2). The hyperparameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ control the shrinkage under our approach, and one can propose hyperpriors on $\boldsymbol{\lambda}$ as in (3) to achieve a hierarchical specification that lets the data control the degree of shrinkage. We demonstrate the shrinkage properties induced by $\boldsymbol{\lambda}$ in Figure 1, which plots the density of the precision off-diagonals under formulation (2)–(3), for varying shrinkage parameters. From Figure 1, it is evident that higher values of λ/b encourage greater shrinkage. We provide an analytic justification for such a phenomenon in Remark 3 in the next section.

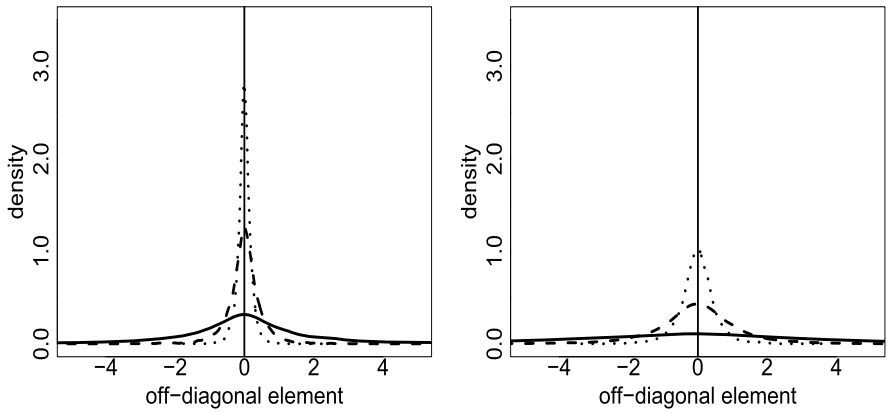

Figure 1: Prior realizations for precision off-diagonals under the regularized inverse–Wishart prior for $\lambda = 5$ (solid), $\lambda = 10$ (dashed), and $\lambda = 15$ (dotted). Left panel has $b, p = 10$, right panel has $b, p = 20$.

Due to its regularization properties as elaborated in the next section, we denote the prior formulation in (2)–(3) as the regularized inverse–Wishart prior on the covariance matrix. We note that our specification is related but different to Huang and Wand (2013) whose focus was on covariance matrix estimation. They had a similar generic specification as in (2), but used conjugate Gamma priors on $d_1, \ldots, d_p$, which resulted in the elements of the covariance matrix having marginally non-informative distributions for certain hyper-parameter choices. However, such an approach is not equipped to provide adequate regularization for precision off-diagonals, and may perform much worse compared to the regularized inverse Wishart method in terms of graph estimation, as demonstrated is simulations. We note that while it is also possible to use a fixed choice of D, our hierarchical specification is motivated by global-local priors resulting in adaptive shrinkage.

2.2 Connections to Cholesky-based regularization

Next, we explicitly establish how the regularized inverse–Wishart prior induces shrinkage in Σ^{-1} through equivalence with a Cholesky-based regularization approach. Note that model (2) allows for the following equivalent representation as a set of regressions

$$x_k = -\sum_{l=1}^{k-1} \omega_{k,kk}^{-1} \omega_{k,kl} x_l + \epsilon_k, \quad \epsilon_k \sim N(0, \omega_{k,kk}^{-1}), k = 1, \ldots, p, \tag{4}$$

with the order of equations being irrelevant to the subsequent discussion and theory. The series of equations (4) suggests that $\Sigma^{-1} = T'V^{-1}T$, with $V = \text{diag}(\omega_{p,pp}^{-1}, \ldots, \omega_{1,11}^{-1})$ and T being a lower triangular matrix that has $t_{kl} = \omega_{k,kk}^{-1}\omega_{k,kl}, l < k$ and $t_{kk} = 1, k = 1, \ldots, p$. Equation series (4) is equivalent to a Cholesky decomposition, with the upper

triangular matrix $T'V^{-1/2}$ being the Cholesky factor. Since $\Sigma^{-1} = (T'V^{-1/2})(V^{-1/2}T)$, the regularized inverse–Wishart prior on Σ will induce corresponding priors on the elements of the Cholesky factor consisting of the regression coefficients in (4). Theorem 1 shows that the induced prior after marginalizing out D in (2)–(3) regularizes the elements in the Cholesky factor $T'V^{-1/2}$, and hence the off-diagonal elements in Σ^{-1}.

Theorem 1. *Under the prior defined in (2) and (3), we can marginalize out D to obtain* $\pi(\Omega|\boldsymbol{\lambda}) \propto \prod_{k=1}^{p} \lambda_k^b (\omega_{k,kk})^{(b+p-1)/2} \exp(-\lambda_p(\omega_{p,pp})^{1/2} + \sum_{k=1}^{p-1} \lambda_k (|\sum_{l=0}^{p-k-1} \omega_{p-l,p-l,p-l}^{-1} \times \omega_{p-l,p-l,k}^2 + \omega_{k,kk}|)^{1/2})$.

The proof is presented in the Supplementary Materials (Kundu et al., 2018), and relies on the fact that $\text{trace}(D\Omega_{\mathcal{P}}) = d_p\omega_{p,pp} + \sum_{k=1}^{p-1} d_k(\sum_{l=0}^{p-k-1} \omega_{p-l,p-l,p-l}^{-1}\omega_{p-l,p-l,k}^2 + \omega_{k,kk})$, as well as using the identity $\det(\Omega_{\mathcal{P}}) = \prod_{k=1}^{p} \omega_{k,kk}$. It is straightforward to see that the resulting prior $\pi(\Omega|\boldsymbol{\lambda})$ is proper.

Remark 1. The prior in Theorem 1 is maximized with respect to λ_k when $\lambda_k = b(|\sum_{l=0}^{p-k-1} \omega_{p-l,p-l,p-l}^{-1}\omega_{p-l,p-l,k}^2 + \omega_{k,kk}|)^{-1/2}$, and similar conclusions hold for any element in $\boldsymbol{\lambda}$. The above can be seen by taking the logarithm of $\pi(\Omega \mid \boldsymbol{\lambda})$ as expressed in Theorem 1, and then differentiating both sides with respect to λ_k and subsequently equating to zero. Thus, a large value of λ_k/b implies shrinkage for the elements of the Cholesky factor $T'V^{-1/2}$ and supports our earlier observation in Figure 1 that larger values of $\lambda_k, k = 1, \ldots, p$, encourage greater shrinkage in Σ^{-1}.

Remark 2. Although Theorem 1 imposes shrinkage due to equivalence with Cholesky-based regularization, it has an important difference in terms of being order invariant.

2.3 Posterior computation steps

The Markov chain Monte Carlo sampler for the regularized inverse–Wishart approach proceeds by using a straightforward fully Gibbs approach, and employing conjugacy to sample the precision matrices as a whole. We consider Gamma hyperpriors on $\lambda_k \sim \text{Ga}(a_{\lambda,k}, b_{\lambda,k}), k = 1, \ldots, p$, and fix $b = 3$ in our computations as in Jones et al. (2005). The posterior computation steps are highlighted in the Supplementary Materials.

3 Model selection and consistency

3.1 Model selection

We develop a post-Markov chain Monte Carlo fitting strategy for graphical model estimation, which assigns exact zeros to precision off-diagonals that correspond to absent edges, by using a decision theoretic approach incorporating joint penalized credible regions. We note that the proposed decision theoretic approach is very general since (a) it can be applied to posterior samples of Σ^{-1} under any prior specification; and (b) it does not make any assumptions about the underlying graph structure, which allows for both decomposable and non-decomposable graphs. The decision theoretic approach performs neighborhood selection (Meinshausen and Bühlmann, 2006) for each node in the graph,

which are then combined to obtain estimates for the entire edge set. The neighborhood for node $i \in V^*$ is defined as $\text{ne}_i = \{j \in V^* \setminus \{i\} : (i,j) \in E\}$, and is estimated by using equivalent L_0 minimization based approaches in regression settings. In this paper, we adapt the approach proposed by Bondell and Reich (2012) to our context, which uses an approximation to solve the L_0 problem and is briefly summarized below.

For a $n \times 1$ vector of responses y and $n \times p$ covariate matrix Z, Bondell and Reich (2012) first fit the full regression model

$$y = Z\boldsymbol{\beta} + \epsilon,\ \epsilon_i \sim N(0,\sigma^2),\ \beta_j \sim N(0,\sigma^2/\tau),\ \sigma^2 \sim \pi(\sigma^2),\ i = 1,\ldots,n, j = 1,\ldots,p, \tag{5}$$

and subsequently perform variable selection under a post-MCMC decision theoretic approach. First, they estimate an ordered sequence of models corresponding to a sequence of credible regions $\mathcal{C}_\alpha$ that have probability content $1-\alpha$, with $\alpha \in (0,1)$ indexing the sequence. The model corresponding to a credible region $\mathcal{C}_\alpha$ is obtained via a sparse solution for $\boldsymbol{\beta}$ induced by minimizing the L_0 norm $||\boldsymbol{\beta}||_0$ which constrained to lie within $\mathcal{C}_\alpha$. In particular, they use the following criteria

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{\beta}||_0, \quad \text{subject to } \boldsymbol{\beta} \in \mathcal{C}_\alpha = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \le C_\alpha\}, \tag{6}$$

where $Pr(\boldsymbol{\beta} \in \mathcal{C}_\alpha) = 1-\alpha$, and $\hat{\boldsymbol{\beta}}, \hat{\Sigma}$ are the posterior mean and covariance of β, based on the full model (5). Since solving the exact L_0 minimization problem (8) involves a combinatorial search which is computationally infeasible for moderate to high dimensions, they approximate the L_0 criterion by a smooth homotopy between L_0 and L_1, which can be solved using existing algorithms such as least angle regression (Efron et al., 2004). Finally, an optimal value of α is chosen from the sequence, which yields a point estimate comprising exact zero effect sizes for unimportant predictors.

Coming back to our graphical model selection context, denote $\boldsymbol{\beta}_k = \{\beta_{kj} = -\omega_{p,kk}^{-1} \times \omega_{p,kj} : j \ne k\}, k = 1,\ldots,p$, which are nothing but the conditional regression coefficients in the following equivalent representation under our original formulation (2)

$$\begin{aligned} x_{ik} &= \sum_{j \ne k, j=1}^{p} \beta_{kj} x_{ij} + \epsilon_{ik},\ \epsilon_{ik} \sim N(0, \omega_{p,kk}^{-1}),\ \beta_{kj} \sim N(0, \omega_{p,kk}^{-1}/d_j), \\ \omega_{p,kk} &\sim \text{Ga}(b/2, d_k/2),\ D \sim \pi(D), \quad j \ne k, j = 1,\ldots,p, \end{aligned} \tag{7}$$

where the conditional normality of the regression coefficients in (7), which is necessary for establishing model selection consistency, is guaranteed by the inverse–Wishart prior on Σ. After convergence of the Markov chain Monte Carlo, the posterior samples of $(\omega_{p,kk}, \boldsymbol{\beta}_k)$ can be viewed as arising from the stationary distribution $\pi(\omega_{p,kk}, \boldsymbol{\beta}_k | X)$, implied by (7). It is worth pointing out here that although the reverse process of fitting the series of p regression models in (7) may potentially result in similar edge set estimates under suitable choices of hyper-parameters, it is not a valid joint distribution procedure, does not yield positive definite precision matrix estimates, and involves two times the number of parameters compared to the proposed approach. Hence we do not consider this alternate approach further in our work. Finally, we note that the

hierarchical formulation (2) specifies a global-local prior (Polson and Scott, 2011; Carvalho et al., 2009, 2010) on the conditional regression coefficients in (7), with $\omega_{p,kk}$ and $d_j^{-1}, j \neq k, j = 1, \ldots, p$, being the global and local scale parameters under the k-th regression. The global scale parameter controls the global shrinkage to the origin, and the local scales allow deviations in the degree of shrinkage, which enables a sharp spike at zero along with thick tails where necessary.

The posterior samples of Σ^{-1} can then be used directly to obtain posterior realizations of $\boldsymbol{\beta}_k, k = 1, \ldots, p$. Further, we note that (7) is very similar to regression model (5), which allows us to adapt the penalized joint credible region approach to obtain a sparse solution of $\boldsymbol{\beta}_k$ corresponding to level α as

$$\tilde{\boldsymbol{\beta}}_k^{\alpha} = \arg\min_{\boldsymbol{\beta}_k} ||\boldsymbol{\beta}_k||_0, \text{ subject to } \boldsymbol{\beta}_k \in \mathcal{C}_\alpha = \{\boldsymbol{\beta}_k : (\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k) \leq C_\alpha\}, \tag{8}$$

where $\hat{\boldsymbol{\beta}}_k$ and $\hat{\Sigma}_k$ are the posterior mean and covariance of $\boldsymbol{\beta}_k$ respectively, under the regularized inverse–Wishart approach. The solution $\tilde{\boldsymbol{\beta}}_k^{\alpha}$ corresponds to a distinct estimated neighborhood $\hat{\text{ne}}_{k,\alpha}$=$\{l \in V^* : \tilde{\beta}_{kl}^{\alpha} \neq 0, l \neq k\}$ for node $k \in V^*$, since $\tilde{\beta}_{kj}^{\alpha} = 0$ implies that the (k,j)-th precision matrix element is zero under the equivalence $\beta_{kj} = -\omega_{p,kk}^{-1}\omega_{p,kj}$, $j \neq k, k = 1, \ldots, p$. As in Bondell and Reich (2012), we solve an approximate version of the L_0 optimization in (8). The proposed solution is unique for both $p \leq n$ and $p > n$ settings for each regression in (7). The details of the computational procedure for the solution can be found in Section 3.2.

Under our approach, two estimates for the edge set are possible for a given α: $\hat{E}_{\alpha,\wedge} = \{(k,l) : k \in \hat{\text{ne}}_{l,\alpha} \wedge l \in \hat{\text{ne}}_{k,\alpha}\}$ and $\hat{E}_{\alpha,\vee} = \{(k,l) : k \in \hat{\text{ne}}_{l,\alpha} \vee l \in \hat{\text{ne}}_{k,\alpha}\}$. Although the two edge sets are not guaranteed to be equal for finite samples, both estimates are equal asymptotically due the neighborhood selection consistency, as elaborated in Theorem 2. Hereafter, we suppress the second subscript and denote the estimated edge set for level α as $\hat{E}_\alpha$. The precision matrix corresponding to level α can be computed as $\hat{\Omega}_{\hat{E}_\alpha} = \hat{\Omega} \otimes ADJ_\alpha$, where $\hat{\Omega}$ is the posterior mean of the Markov chain Monte Carlo samples under the regularized inverse–Wishart approach, ADJ_α is the adjacency matrix corresponding to the edge set $\hat{E}_\alpha$ and $\otimes$ denotes the element-wise product. As demonstrated in the next section, a careful choice of α depending on n leads to asymptotic consistency for $\hat{\Omega}_{\hat{E}_\alpha}$ for fixed p settings, which also implies asymptotic positive definiteness.

Noting that the above estimate for the precision matrix is not guaranteed to be positive definite for finite samples, we propose an alternate estimator obtained by fixing the off-diagonals corresponding to absent edges to be zero, while rescaling the elements in the Cholesky factor of $\hat{\Omega}$ in order to ensure positive definiteness. The estimator and the algorithm needed to obtain it, are described in Section 3.5.

3.2 Computation steps for neighborhood selection

Following Bondell and Reich (2012), who noted that solving an exact L_0 minimization problem involves a combinatorial search which is computationally infeasible for moderate to high dimensions, we replace the L_0 criterion in (8), by a criterion proposed by Lv

and Fan (2009) which is a smooth homotopy between L_0 and L_1. Instead of optimizing (8) under the L_0 penalty, we use the modified penalty $\sum_{j=1,j\neq k}^{p} \tau_a(|\beta_{kj}|)$ where

$$\tau_a(|t|) = (\frac{|t|}{a+|t|})I(|t| \neq 0) + (\frac{a}{a+|t|})|t|, \ a > 0. \tag{9}$$

The above homotopy approximates the L_0 criteria in the limiting sense as $\tau_0(|t|) = \lim_{a\to 0+} \tau_a(|t|) = I(|t| \neq 0)$. Thus, as an approximation to (8) in the article, the k-th neighborhood is now obtained by minimizing

$$\sum_{j=1,j\neq k}^{p} \lim_{a\to 0+} \tau_a(|\beta_{kj}|) \text{ subject to } \boldsymbol{\beta}_k \in \mathcal{C}_{\alpha,k}, k = 1,\ldots,p. \tag{10}$$

As in Bondell and Reich (2012), the above can be solved using a local linear approximation (Zou and Li, 2008) which reduces the modified optimization problem (10) to

$$\begin{aligned} \tilde{\beta}_k &= \arg\min_{\beta_k}(\beta_k - \hat{\beta}_k)^T\hat{\Sigma}_k^{-1}(\beta_k - \hat{\beta}_k) + \Delta_\alpha \sum_{j=1,j\neq k}^{p} \lim_{a\to 0+}(a + |\hat{\beta}_{kj}|)^{-2}|\beta_{kj}|, \\ &= \arg\min_{\beta_k}(\beta_k - \hat{\beta}_k)^T\hat{\Sigma}_k^{-1}(\beta_k - \hat{\beta}_k) + \Delta_\alpha \sum_{j=1,j\neq k}^{p} \hat{\beta}_{kj}^{-2}|\beta_{kj}|, \end{aligned} \tag{11}$$

where Δ_α corresponds to a penalty parameter having a one-to-one correspondence with α. The above is a Lagrangian optimization problem and can be equivalently written as

$$\tilde{\beta}_k^* = \arg\min_{\beta_k^*}(Y_k^* - X_k^*\beta_k^*)^T(Y_k^* - X_k^*\beta_k^*) + \Delta_\alpha \sum_{j=1,j\neq k}^{p} |\beta_{kj}^*|, \tag{12}$$

where $Y_k^* = \hat{\Sigma}_k^{-1/2}\hat{\beta}_k$, $X_k^* = \hat{\Sigma}_k^{-1/2}B_k$, and B_k is a diagonal matrix having elements $\hat{\beta}_{kj}^2, j \neq k, j = 1,\ldots,p$. Equation (12) is just the usual LASSO problem and can be solved using the efficient LARS algorithm, and solution for the original quantity of interest can simply be obtained using $\tilde{\beta}_k = B_k\tilde{\beta}_k^*$. For $p < n$ settings, the above is a strictly convex problem with having full rank (i.e. rank(X_k^*)=$p-1$), and hence has a unique solution. For $p > n$, results in Tibshirani (2013) suggest that our solution is unique with probability one, as the responses are continuous under our set-up.

We note that the proposed neighborhood selection method via penalized credible regions can be interpreted as a decision theoretic approach. In particular, the solution to the neighborhood estimation equation (11) can be approximated as the minimizer to an expected loss function as in the following Lemma. Let $\tilde{D}_{-k}$ denote the posterior mean for diagonal matrix D without the k-th diagonal entry, and let $||\boldsymbol{\beta}||^2_{2,A^{1/2}}$ denote the scaled L_2 norm $\boldsymbol{\beta}^T A\boldsymbol{\beta}$.

Lemma 2. *The solution to equation (11) corresponding to level α can be approximated as $\tilde{\boldsymbol{\beta}}_k^\alpha = arg\ min_{\boldsymbol{\beta}_k}\ \Delta_\alpha||\boldsymbol{\beta}_k||_0 + ||X_{-k}\boldsymbol{\beta}_k - X_{-k}\hat{\boldsymbol{\beta}}_k||^2_{2,(1/\sqrt{\tilde{s}_k})I} + ||\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k||^2_{2,(1/\sqrt{\tilde{s}_k})\tilde{D}_{-k}^{1/2}}$, where $||\boldsymbol{\beta}||^2_{2,A^{1/2}} = \boldsymbol{\beta}^T A\boldsymbol{\beta}$ and $\tilde{s}_k \to \omega^{-1}_{0,kk}$.*

The proof is presented in the Supplementary Materials. In the above decision theoretic criteria, the first term determines the sparsity of the solution, the second term is based on an expected scaled squared loss function under a posterior predictive distribution similar to the loss function for regression in Hahn and Carvalho (2015), and the third term imposes an additional scaled squared loss penalty which encourages the solution for the regression coefficients to concentrate around the posterior mean of the regression coefficients. The scaling matrices $(1/\tilde{s}_k)I$ and $(1/\tilde{s}_k)\tilde{D}_{-k}$ account for the variance corresponding to the prior $\boldsymbol{\beta}_k \sim N(0, \omega_{p,kk}^{-1} D_{-k}^{-1})$ under the regularized inverse-Wishart prior.

3.3 Selection consistency

In this section, we establish that the proposed model selection approach leads to consistent neighborhood selection under some suitable assumptions. Suppose that for a given sample of size n, we estimate the neighborhood corresponding to level α_n in the ordered sequence, and denote the corresponding estimated neighborhood for the k-th node as $\hat{ne}_{k,n}$. By choosing α_n such that $1-\alpha_n \to 1$ as $n \to \infty$, we attain neighborhood selection consistency which is mathematically defined as $P(\hat{ne}_i = ne_i^*) \to 1$, as $n \to \infty$, for all $i \in V^*$, where $\hat{ne}$ and ne^* refer to the estimated and true neighborhoods respectively. In other words, the probability of the estimated neighborhood for each node being equal to the true neighborhood asymptotically goes to 1 as the coverage increases with n.

For a sample size n, denote the credible region for $\boldsymbol{\beta}_k$ with content $1-\alpha_n$ as $\mathcal{C}_{n,k} = \{\boldsymbol{\beta}_k : (\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k) \leq C_n\}$. Let E_0 be the true edge set corresponding to an undirected graph and having true neighborhood ne_{k0} for node k, $k = 1, \ldots, p$. Consider the following assumptions:

(A1) The true model is $\mathbf{x} \sim N(0, \Omega_{E_0}^{-1})$, where $\Omega_{E_0} = (\omega_{0,ij})_{i,j=1}^p$ has exact zeros for the off-diagonals corresponding to absent edges in E_0.

(A2) Ω_{E_0} is positive definite with $c_1 n^{-1/2} < \omega_{0,ii}^{-1} |\omega_{0,ij}| < c_2$ for finite and positive constants c_1, c_2, for all $\{\omega_{0,ij} : (i,j) \in E_0\}$.

(A3) When $p_n = o(n)$, each node has a finite number of neighbors.

Assumption *(A1)* states that the true model is a Gaussian graphical model with edge set E_0. Assumption *(A2)* implies that the true precision off-diagonal elements corresponding to edges in E_0 are sufficiently large but bounded above. Assumption *(A3)* ensures that the true model is sparse under increasing dimensions. The following Theorem establishes selection consistency.

Theorem 2. *For fixed p, suppose assumptions* (A1)–(A2) *hold, and choose a sequence of credible regions $\mathcal{C}_{n,k}$ such that $C_n \to \infty$ and $n^{-1}C_n \to 0$. For $p_n = o(n)$, suppose assumptions* (A1)–(A3) *hold, and choose $\mathcal{C}_{n,k}$ such that $C_n \to \infty$, $p_n^{-1}C_n \to \infty$ and $n^{-1}C_n \to 0$. Then neighborhood selection consistency is attained under the regularized inverse–Wishart approach for both cases.*

The proof of Theorem 2 is provided in the Supplementary Materials. Let the estimated generic edge set for level α_n be denoted as $\hat{E}_n$, with the corresponding estimated precision matrix denoted as $\hat{\Omega}_{\hat{E}_n}$. The following result holds for both $\hat{E}_{n,\wedge}$ and $\hat{E}_{n,\vee}$ for the case of fixed p.

Corollary 1. *If neighborhood selection consistency holds, then* $Pr(\hat{E}_n = E_0) \to 1$ *and* $Pr(\hat{\Omega}_{\hat{E}_n} = \Omega_{E_0}) \to 1$ *as* $n \to \infty$*, for fixed* p*.*

The first part of Corollary 1 follows since the conditional independence structure of a multivariate normal can be consistently estimated by combining the neighborhood estimates of all variables, for fixed p. The edge set estimates, $\hat{E}_{n,\wedge}$ and $\hat{E}_{n,\vee}$, converge asymptotically when the truth is a Gaussian graphical model, due to consistent selection of the neighborhoods in Theorem 2. The proof for the second part of Corollary 1 is in the Supplementary Materials. Since Ω_{E_0} is positive definite by assumption, $\hat{\Omega}_{\hat{E}_n}$ is asymptotically positive definite for fixed p.

The above results are based on the assumption that the scale matrix D is diagonal. However, the next Corollary shows that consistency still holds under a non-diagonal $D \in M^+$ under the following additional assumption. The proof is presented in the Supplementary Materials. Let D_{-k} denote the D matrix excluding the k-th row and column, $\mathbf{d}^T_{k,-j}$ denote the k-th row of D excluding the j-th element, and $\Omega_{0,k}$ denote the true precision matrix of $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)$.

(A4) All elements of D are $o(n)$ and $\lim_{n\to\infty}(1/n)\Omega_{0,k}D^{-1}_{-k}\mathbf{d}^T_{k,-j} \to \mathbf{0}_{p_n}$ for $p_n = o(n)$, with $\mathbf{0}_{p_n}$ denoting a zero vector with p_n elements.

Corollary 2. *Suppose* $D \in M^+$ *is fixed and non-diagonal, and that assumptions* (A1)–(A4) *hold. Then the neighborhood selection consistency holds as in Theorem 2.*

3.4 Edge selection based on false discovery rates

Bayesian methods usually obtain point estimates of the graph by including edges that have a posterior probability > 0.5, or reporting the graph that has the highest log marginal likelihood. In lieu of these approaches, we propose an approach based on controlling the false discovery rate, which includes a natural multiplicity correction and can directly control the level of sparsity in edge selection. This approach is based on the strategy that the edges which are strongly supported by the data will likely appear often in the ordered sequence of neighborhoods computed via penalized credible regions, whereas other edges with weaker evidence will likely appear less often. This approach has similarities to the one used to compute brain network using sliding window correlations (Monti et al., 2014).

Instead of specifying a confidence level α and then computing the corresponding neighborhoods, the proposed strategy for edge selection instead relies on fitting a series of neighborhoods for varying sparsity levels corresponding to different values of the penalty parameter Δ_α which has a one-to-one but highly non-linear relationship with α (Bondell and Reich, 2012). After fitting a series of neighborhoods corresponding to varying sparsity levels, we select those edges as important which have pseudo-probabilities

greater than some threshold, where these pseudo-probabilities are computed based on the frequency of occurrence of an edge in the ordering. We adapt the approaches in Morris et al. (2008) and Baladandayuthapani et al. (2010) to determine such a threshold, which is designed to control false discovery rates and is described below.

We first compute a pseudo posterior inclusion probability matrix $P = (P_{ij})$ by computing the proportion of times each edge is included in the ordering of graphs based on the sequence of credible regions having probability content $\{(1-\alpha_m)\}_{m=1}^{R}$, where R is the chosen number of credible regions in the sequence. Then $1 - P_{ij}$ can be considered akin to Bayesian q-values, or estimates of the "local false discovery rate" (Newton et al., 2004), as they measure the probability of a false positive if the (i,j)-th edge is called a 'discovery' or is significant. Given a desired global false discovery rate bound $\eta \in (0,1)$, one can determine a threshold c_η that flags the set of important edges as $\hat{E}_\eta = \{(i,j) : P_{ij} \geq c_\eta\}$. This yields a point estimate of the graph.

The threshold c_η can be chosen in the following manner. Let vec(P) be the vectorized upper triangular matrix of P excluding diagonals, containing the pseudo posterior inclusion probabilities of the edges stacked column–wise. We first sort vec(P) in descending order to yield the sorted vector vec$(\tilde{P})$=$\{\tilde{P}_k, k = 1, \ldots, p(p-1)/2\}$. Then we can estimate c_η as the ζ-th entry of vec$(\tilde{P})$, where $\zeta = \max\{j^* : \frac{1}{j^*}\sum_{k=1}^{j^*} \tilde{P}_k \leq \eta\}$, and a lower value of c_η leads to sparser graphs, while simultaneously controlling the FDR at a pre-specified level. However, we would note that this FDR is based on pseudo-probabilities which may not reflect the true posterior edge inclusion probabilities, and hence these FDR values and associated pseudo-probabilities need to be interpreted with care. Finally, if the upper limit of Δ_α is increased drastically, or alternately, if all the Δ_α values lie within a very small neighborhood of zero, then the resulting estimated graph may be somewhat sensitive to such choices. However, by choosing the threshold c_η appropriately, one can mitigate the sensitivity to the grid values to a large extent and obtain similar graphs with comparable false discovery rates.

3.5 Algorithm for estimating positive definite precision matrix

We propose an algorithm for estimating the precision matrix which rescales the elements of the Cholesky factor of the posterior mean of Ω given a set of zero entries under the estimated graph, in a manner that ensures positive definiteness. This algorithm is derived based on the results in Chan and Jeliazkov (2009), who propose a Markov chain Monte Carlo sampling approach for restricted covariance matrices. Our algorithm is different from their approach, and focuses on the inverse covariance matrix while requiring only a series of post-Markov chain Monte Carlo deterministic steps, instead of having to sample restricted precision matrices at each Markov chain Monte Carlo iteration. The resulting approach is thus computationally efficient even for large p. The algorithm takes in the positive definite posterior mean $\hat{\Omega}$ and the estimated edge set $\hat{E}$ as inputs, and outputs another positive definite estimate Ω^* which has exact zeros corresponding to absent edges in $\hat{E}$.

Using the form of the Cholesky factor in the proof of Theorem 1 and similar to eqn (18) in Chan and Jeliazkov (2009), it is straightforward to see that setting $\omega_{kl} = 0$

leads to an adjustment in the Cholesky factor matrix as $\tilde{\omega}_{k,kl} = -\omega_{l,ll}^{-1} \sum_{h=1}^{l-1} \omega_{k,kh}\omega_{l,lh}$, where $\tilde{\omega}_{k,kl}$ is the modified value for the (k,l)-th element in the Cholesky factor which maintains positive definiteness of the precision matrix under the restriction $\omega_{kl} = 0$. It is understood that the right hand side of the equation is zero when summing over the empty set (i.e. $\tilde{\omega}_{k,k1} = 0$). We propose the following post-Markov chain Monte Carlo algorithm which takes as inputs the estimated graph $\hat{G}$ having edge set $\hat{E}$, and the posterior mean of the precision matrix $\hat{\Omega}$, and outputs a positive definite matrix with exact zeros for absent edges.

Algorithm 1. Step 1: Accept $\hat{\Omega} = (T^*)F(T^*)^T$, the posterior mean of Ω as the input, along with the estimated edge set $\hat{E}$, where $T^* = (t_{ij})$ is a lower triangular matrix and $F = \text{diag}(f_1, \ldots, f_p)$.

Step 2: For increasing $k = 2, \ldots, p$, and $l = 2, \ldots, k-1$, if edge (k, l) is absent in $\hat{E}$, modify the (k, l)-th element in T^* as $-(1/f_l)\sum_{h=1}^{l-1} t_{kh}t_{lh}f_h$. It is understood that the (k, l)-th element is updated before the (k, l')-th element in T^*, for $1 \leq l < l' \leq k-1$. Note that this step implies that we fix the $(k, 1)$-th element in the Cholesky factor T^* as zero if edge $(k, 1)$ is missing from $\hat{E}$.

Step 3: The final estimated precision matrix is given as $\Omega^* = LF(L)^T$, where L now denotes the modified lower triangular matrix obtained by rescaling all elements in T^* corresponding to absent edges in $\hat{E}$, and F is unchanged.

This estimate Ω^* is guaranteed to be positive definite, and the number of steps required to compute Ω^* is equal to the number of absent edges. For greater clarity, we present a toy numerical example of the proposed positive definite estimator in the Supplementary Materials.

3.6 A related approach

We note that the idea of fitting a posterior distribution and then performing a post-MCMC decision theoretic step was also proposed recently in Hahn and Carvalho (2015). They primarily focused on variable selection in the context of a Bayesian linear regression model and extend the approach to graphical models, using a post-processing decision theoretic step which involved minimizing the expected loss $E(L(\tilde{Y}, \boldsymbol{\gamma}))$. For graphical model selection, the Hahn and Carvalho (2015) approach reduces to minimizing the expected loss function

$$\begin{aligned} &\arg\min{}_{\Omega} E[\lambda||\Omega||_0 - \log(det(\Omega)) - n^{-1}tr(\tilde{X}\tilde{X}^T\Omega)] \\ &= \arg\min{}_{\Omega}\{\lambda||\Omega||_0 - \log(det(\Omega)) - n^{-1}tr(\bar{\Sigma}\Omega)\}, \end{aligned} \tag{13}$$

where Ω represents the inverse covariance matrix, $\bar{\Sigma}$ denotes the posterior mean of the covariance matrix, and the expectation is taken with respect to the posterior predictive distribution of $\tilde{X}$. The authors propose to solve a surrogate problem by replacing the L_0 penalty in (13) with the L_1 penalty, which can be solved using a graphical lasso (Friedman et al., 2008) algorithm.

Although both approaches operate by fitting a Bayesian model, and then performing a post-processing step involving a L_0 penalization to estimate the model, there are important differences. Firstly, the proposed approach is based on neighborhood selection via penalized joint credible regions, whereas the Hahn and Carvalho (2015) relies on sparse precision matrix estimation. Second, our article develops a novel prior on the covariance matrix, which induces shrinkage on the precision matrix off-diagonals, whereas the Hahn and Carvalho (2015) article did not focus on developing novel priors. Third, the two approaches choose the tuning parameters (Δ_α and λ) in a completely different manner – the Hahn and Carvalho (2015) method uses the full posterior distribution, while the proposed approach fits a series of models under a sequence of tuning parameter values, and then selects the optimal graph using a criteria designed to control for false positive rates. Moreover, the proposed approach has attractive theoretical properties in terms of graphical model selection consistency, even when the number of nodes grows with the sample size, whereas the theoretical properties of the graphical modeling approach in Hahn and Carvalho (2015) have not been vetted, to our knowledge. The two approaches also perform differently in numerical studies, as elaborated in the simulation section.

4 Simulation studies

4.1 Description

We present several simulation scenarios comparing our approach to (a) frequentist graphical lasso (Friedman et al., 2008); (b) neighborhood selection approach by Meinshausen and Bühlmann (2006); (c) the hyper inverse–Wishart approach employing reversible jump Markov chain Monte Carlo (Giudici and Green, 1999); (d) shrinkage approaches such as the Bayesian graphical lasso and Bayesian adaptive graphical lasso (Wang, 2012); and (e) the unregularized inverse–Wishart prior which has the same formulation as (2), but with $D = dI_p$ and $d \sim Ga(1, 1)$, which resembles the prior in Huang and Wand (2013). The Matlab code for the Bayesian lasso and adaptive lasso were obtained from the supplementary materials of Wang (2012), while the frequentist graphical lasso and the method by Meinshausen and Bühlmann (2006) were implemented using the 'glasso' package in R. We wrote the code for implementing the reversible jump Markov chain Monte Carlo under the model $\mathbf{x}_i \sim N(0, \Sigma_G), \Sigma_G \mid G \sim HIW(b, D),\ i = 1, \ldots, n$, with $G \sim \pi(G)$ is restricted to the class of decomposable graphs, and $\pi(G)$ defined by independent Bernoulli(p^*) priors on the edge inclusion indicators. Here $p^* \sim U(0, 1)$ and HIW refers to the hyper inverse–Wishart prior. For the shrinkage procedures, 15000 MCMC iterations with a burn in of 5000 was used, while 100000 iterations with burn in of 10000 was used for the discrete mixture approach (c), with the initial adjacency matrix corresponding to a null graph.

We considered several cases for data-generation, with each case having 50 replicates. We fit a slightly modified model (2) for $p > n$ settings, by specifying $D \sim$ Wishart(b_D, I_p), where $D \in M^+$ is no longer constrained to be diagonal.

Case I. Data is generate from a Gaussian distribution with the covariance matrix being a fractional Gaussian noise process having elements

$$\sigma_{ij} = \frac{1}{2}\left[||i-j|+1|^{2H} - 2|i-j|^{2H} + ||i-j|-1|^{2H}\right],$$

where $H \in [0.5, 1]$ is the Hurst parameter, and chosen to be $H = 0.7$.

Case II. We generate data emulating a real data application, using mRNA expression levels for 49 genes that are available from The Cancer Genome Atlas portal. The full details and the corresponding results are available in the Supplementary materials.

Case III. We generate data from a two component mixture of Gaussians, i.e. $\mathbf{x}_i \sim \pi N(-1_p, \Sigma) + (1-\pi)N(1_p, \Sigma)$, where 1_p denotes a p-vector of ones, Σ is defined as in Case I, and $\pi \in (0, 1)$ is the mixing proportion. This case corresponds to a non-Gaussian truth, violating the Gaussian assumption inherent in our formulation.

Case IV. We generate data from a Gaussian graphical model as in Peng et al. (2009), where the edges were generated randomly with probability 0.002, and the precision matrix off-diagonals were set to zero for absent edges and were generated from a $U(-1, 1)$ distribution otherwise, with the k-th diagonal then being computed as $\omega_{kk} = 1 + \sum_{l \neq k} |\omega_{kl}|$, to maintain diagonal dominance. This case corresponds to a sparse graphical model.

In addition to the above cases, we also look at another simulation scenario based on a genomics example, the results for which are presented in the Supplementary section. Since Cases I and III considered here correspond to non-sparse precision matrices having few exact zero entries, we adopt a slightly different notion to define the true edge set. In particular, the true edge set includes all edges corresponding to absolute partial correlations greater than a certain threshold c_m. We examine point estimates corresponding to true edge sets ES025 and ES005, obtained by choosing $c_m = 0.025$, and $c_m = 0.005$ respectively. The edge set ES005/ES025 essentially treats all edges corresponding to an absolute partial correlation $> 0.005/0.025$ as important and other edges as unimportant.

For our approach, the point estimate is obtained under a false discovery rate of 0.2, whereas the estimates for the frequentist approaches were obtained by minimizing a Bayesian Information Criteria or BIC, as in Yuan and Lin (2007). The point estimate of the graph under the discrete mixture approach corresponds to all edges having posterior inclusion probability ≥ 0.5, while the heuristic thresholding method of Wang (2012) is adopted for Bayesian graphical lasso.

4.2 Results

For comparing results, we look at the area under the curve as well as the sensitivity and specificity levels under the point estimate for the graph. We also report the sensitivity corresponding to a specificity of 90% for $p > n$ cases. The results are presented in Tables 1–2, and the ROC curves for Case I are illustrated in Figure 2. In addition, we examined if the reported metrics are significantly better under one particular method compared to others using a permutation test. We note that it was not feasible to obtain results for competing Bayesian approaches under $p > n$ settings due to an unrealistic computational burden.

Case I ($p < n$)								
(n, p)	300,100		400,200		500,100		500,200	
	ROC005	ROC025	ROC005	ROC025	ROC005	ROC025	ROC005	ROC025
RIW	66	90	66	90	66	90	68	91
GL	61	84	64	85	64	87	63	85
MB	58	75	60	78	60	79	60	79
IW	58	82	61	86	58	85	60	83
BGLA	55	69	57	72	52	72	59	71
BGAD	58	71	56	75	56	76	60	75
HIW	53	67	57	60	51	65	58	62
HC	67	90	68	91	67	89	68	92
(n, p)	300,100				400,200			
	SP005	SE005	SP025	SE025	SP005	SE005	SP025	SE025
RIW	95	25	91	61	98	18	98	55
GL	99	07	99	23	99	07	99	28
MB	100	0	100	0	100	0	100	0
IW	97	14	95	51	94	19	97	49
BGLA	80	34	80	61	87	28	88	47
BGAD	94	19	94	43	95	19	95	41
HIW	99	07	99	19	99	06	99	08
HC	100	11	100	35	100	12	99	37
(n, p)	500,100				500,200			
	SP005	SE005	SP025	SE025	SP005	SE005	SP025	SE025
RIW	92	33	90	75	98	24	91	78
GL	99	07	99	33	99	07	99	28
MB	100	0	100	0	100	0	100	0
IW	98	14	93	71	98	11	98	69
BGLA	74	47	72	72	85	33	84	60
BGAD	95	23	93	47	95	21	95	45
HIW	99	08	99	22	99	07	99	11
HC	100	12	99	38	100	12	100	37

Table 1: Area under the curve for true edge sets ES005 (ROC005) and ES025 (ROC025) under Case I for $p < n$, along with the sensitivity and specificity under the point estimate. The largest squared standard errors across rows corresponding to the area under the curve are 0.01, 0.008, 0.01, 0.01, 0.01, 0.008, 0.07 and 0.03 for $p < n$. The reported estimates are inflated by 100. RIW, IW, BGLA, BGAD, HIW, GL, MB, and HC, refer to regularized inverse-Wishart, inverse-Wishart, Bayesian graphical lasso, Bayesian adaptive graphical lasso, hyper inverse-Wishart, frequentist graphical lasso, Meinshausen and Bühlmann (2006) method, and the approach by Hahn and Carvalho (2015), respectively. Results based on 50 replicates.

Case I								
(n,p)	200,300		200,400		200,300		200,400	
Method	ROC005	ROC025	ROC005	ROC025	SE005*	SE025*	SE005*	SE025*
RIW	66	83	64	83	32	63	33	64
GLASSO	61	77	60	77	31	63	31	64
MB	56	71	56	70	25	54	24	53
HC	64	81	64	81	33	63	34	65
(n,p)	200,300		200,300		200,400		200,400	
Method	SP005	SE005	SP025	SE025	SP005	SE005	SP025	SE025
RIW	99	10	99	27	99	10	99	25
GLASSO	99	04	99	17	84	18	91	18
MB	99	10	99	30	99	10	99	31
HC	100	05	100	22	99	03	99	21
Case III								
(n,p)	200,300		200,400		200,300		200,400	
Method	ROC005	ROC025	ROC005	ROC025	SE005*	SE025*	SE005*	SE025*
RIW	65	83	65	83	33	65	32	64
GLASSO	61	77	61	76	33	66	32	64
MB	57	72	56	72	24	54	24	54
HC	64	81	64	81	33	63	34	65
(n,p)	200,300		200,300		200,400		200,400	
Method	SP005	SE005	SP025	SE025	SP005	SE005	SP025	SE025
RIW	99	09	99	27	99	08	99	25
GLASSO	99	05	99	15	86	15	90	16
MB	99	09	99	30	99	09	99	30
HC	100	07	100	21	100	05	99	23
Case IV								
(n,p)	200,300		200,400		200,300		200,400	
Method	ROC005	ROC025	ROC005	ROC025	SE005*	SE025*	SE005*	SE025*
RIW	65	83	65	83	33	65	32	64
GLASSO	61	77	61	76	33	66	32	64
MB	57	72	56	72	24	54	24	54
HC	53	56	52	57	20	33	18	32
(n,p)	200,300		200,300		200,400		200,400	
Method	SP005	SE005	SP025	SE025	SP005	SE005	SP025	SE025
RIW	99	09	99	27	99	08	99	25
GLASSO	99	05	99	15	86	15	90	16
MB	99	09	99	30	99	09	99	30
HC	99	06	99	06	99	07	99	06

Table 2: Results for $p > n$ cases under Cases I, III, and IV. ROC005 and ROC025 correspond to the area under the curve for true edge sets ES005 and ES025, obtained by including all edges corresponding to $|\rho| > 0.005$ and $|\rho| > 0.025$ respectively. SE005* and SE025* refer to the sensitivity controlling for a specificity of 0.9, for true edge sets ES005 and ES025. SP005, SP025, refer to specificity, while SE005, SE025, refer to sensitivity corresponding to true edge sets ES005 and ES025. The reported estimates are inflated by a factor of 100.

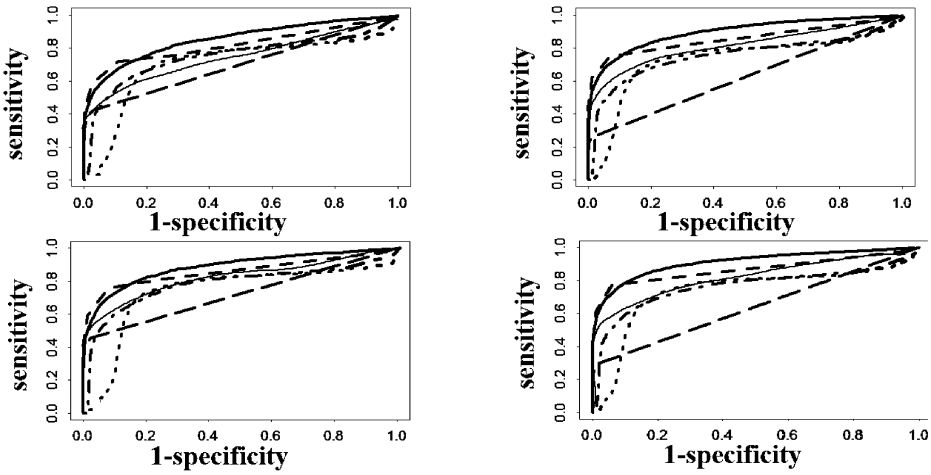

Figure 2: ROC curves for true edge set corresponding to absolute partial correlation threshold of 0.025, for $p < n$ settings under Case I. Thick solid line is the regularized inverse–Wishart, dashes is frequentist graphical Lasso, dots is Bayesian graphical lasso, dots & dashes are Bayesian adaptive graphical lasso, long dashes is hyper inverse–Wishart, thin black line corresponds to Meinshausen and Bühlmann (2006).

Under Case I, we find that both the proposed approach and the HC2015 method have a significantly higher area under the curve compared to other approaches. Moreover the proposed approach has a significantly higher sensitivity compared to all approaches except the Bayesian graphical lasso in some cases, with the latter reporting denser graphs having low specificity levels. On the other hand, the proposed method has slightly lower specificity levels compared to some of the other approaches which report much sparser graphs. For $p > n$ settings, the sensitivity corresponding to a specificity of 90% is very similar under the proposed method, HC2015, and graphical lasso, but it is significantly lower under the penalized neighborhood selection approach (MB). Similar conclusions hold when the data is generated under Case III, from a bimodal distribution.

For Case IV involving sparse true graphs, the proposed approach performs significantly better compared to all other approaches in terms of area under the curve, whereas both GL and the proposed approach have equivalent sensitivity levels for a specificity of 90%, which is significantly higher than the MB method as well as the Hahn and Carvalho (2015) approach. Under Case IV, it is evident that the Hahn and Carvalho (2015) approach has the least favorable performance in terms of the area under the curve and sensitivity levels corresponding to 90% specificity.

In addition, the limitations of the discrete mixture approaches are evident from significantly lower areas under the curve and low sensitivity levels in Case I, in spite of the true graph being decomposable. Moreover, the results are highly sensitive to different choices of the initial adjacency matrices, and somewhat unstable for higher dimensions under finite runs of the Markov chain Monte Carlo. We note that a similar unstable

behavior was reported by Scott and Carvalho (2008) for Metropolis-based approaches under discrete mixture priors.

Finally, we note that the proposed approach is several orders of magnitude faster compared to other Bayesian graphical modeling approaches. As noted above, we have difficulty implementing the competing Bayesian approaches for large dimensions due to an unrealistically large net computation time. More details, including the computation times for Bayesian approaches under Case I are presented in the Supplementary Materials.

4.3 Comments

From the simulation results, it is clear that the proposed approach (i) compares favorably to, and often dominates, competing Bayesian approaches in terms of true graph recovery in large dimensions; (ii) has a higher area under the curve compared to penalized approaches in several cases, and a similar performance in other scenarios; and (iii) has demonstrably better computational efficiency compared to other Bayesian approaches, with the latter approaches quickly becoming computationally infeasible for increasing p.

5 Application to cancer genomics

We consider the problem of inferring the association networks between microRNAs, or miRNAs, and the corresponding target genes or messenger RNAs, or mRNAs. MiRNAs are small non-coding RNA molecules, which regulate gene expression levels by silencing the target mRNAs. Our motivating dataset is derived from The Cancer Genome Atlas based study of glioblastoma multiforme, a rapidly growing malignant brain tumor that is the most common in adults. mRNAs and miRNAs play complementary roles in the development and disease progression of this tumor (Tang et al., 2013). The main scientific question of interest is to find major miRNA regulators of mRNA expression in individuals with tumors, by jointly analyzing mRNA and miRNA data. For our analyses, we focus on a set of 49 genes mapped to core pathways implicated in glioblastoma multiforme such as the receptor tyrosine kinase, phosphatidylinositol-3-OH kinase and etinoblastoma pathways (Cancer Genome Atlas Research Network, 2008).

For inferring the gene regulatory network, we chose the top 200 prognostic miRNAs from a list of 538 candidate miRNAs, in addition to the 49 mRNAs, with the measurements being obtained from $n = 280$ samples. We fit the regularized inverse–Wishart model to the combined mRNA and miRNA measurements having dimension $p = 249$. We use $\eta = 0.2$ in our false discovery rate based approach for edge selection. In Table 6, we provide a ranked list of important miRNAs having negative partial correlations with their target mRNAs, along with the magnitude and 99% credible intervals of the partial correlation. We omit those mRNAs not having negative associations with any miRNA. In biological terms, the partial correlation between a miRNA and its target mRNA measures the association between the two, after accounting for the remaining mRNAs and miRNAs, with negative associations implying a down-regulation of target

mRNAs. The credible intervals for the partial correlation provide a measure of uncertainty for the ranking of important miRNAs, with overlapping credible intervals for two miRNAs implying a potential change in rankings under different experimental and biological conditions for a particular target mRNA.

Figure 3 plots the negative partial correlations between important miRNAs and target mRNAs, along with 99% credible intervals. It is evident that several important miRNAs have overlapping credible intervals, thus pointing to uncertainty in the rankings. Several miRNAs identified by our approach are known to downregulate miRNAs, as suggested by the algorithm TargetScan of Lewis et al. (2005)- in particular, HSA-MIR-143,HSA-MIR-28, for KRAS, HSA-MIR-125A for CDKN2B, HSA-LET-7G,HSA-MIR-142-5P,HSA-MIR-144,HSA-MIR-15B,HSA-MIR-187,HSA-MIR-24,HSA-MIR-492, for CDK6. In addition, we find that KSHV-K12-9 down-regulates several mRNA expressions, and this miRNA is known to be associated with glioblastoma (Delfino et al., 2011).

We also identify several hub genes based on mRNA expressions, EFGR, RAF1, NF1, SPRY2, CDKN2A, and PIK3C2G, with such nodes having greater than 8 neighbors. Some of the highly connected genes such as PI3KC2G, EGFR and CDKN2A, with 14, 9 and 9 connections respectively, have been previously shown to be associated with glioblastoma (Dong et al., 2010; Wong et al., 1992). We further explored the biological implications of our results using Ingenuity Pathway Analysis (IPA version 16542223) which identified a number of enriched pathways including; GLIOMA, GBM, PTEN signaling and other cancer related molecular mechanisms. Most of these genes encode proteins critical to cellular functions such as DNA recombination, and repair, cellular development, cell cycle and connective tissue development which may be attributed to their highly connected nature. The estimated graph is shown in Figure 4.

The estimated miRNA graph had 9 hub nodes each having greater than 10 neighbors, while 107 nodes did not have any neighbors. These hub nodes were ebv-mir-bart7, hsa-mir-106a, hsa-mir-142-3p, hsa-mir-17-5p, hsa-mir-let7, hsa-mir-181c, hsa-mir-184, hsa-mir-19a, and hsa-mir-20a. Analogous to gene expression a similar analysis of the miRNA with at least 4 neighbors (based on partial correlations) using IPA suggest they are critical for various cellular processes in cancer progression. The selected molecules modulate important transcription factors and signaling molecules including genes such as MYC, CCLE1 and CLDND1 which have been shown to be associated with cancer, inflammatory response and connective tissue disorders. These findings concur with studies that have indicated associations between GBM and significant modulation of listed miRs such as miR-106 (26 connections), miR-184 (15 connections) and miR-let-7a (26 connections), as in Lee et al. (2011).

6 Discussion

In summary, we propose a novel graphical model estimation approach which fits a conjugate model to the inverse covariance matrix that shrinks the off-diagonal elements corresponding to absent edges to zero, and subsequently uses a post-processing decision theoretic step involving neighborhood selection which infers absent edges based on penalized credible regions. By decoupling model fitting and selection, the proposed method

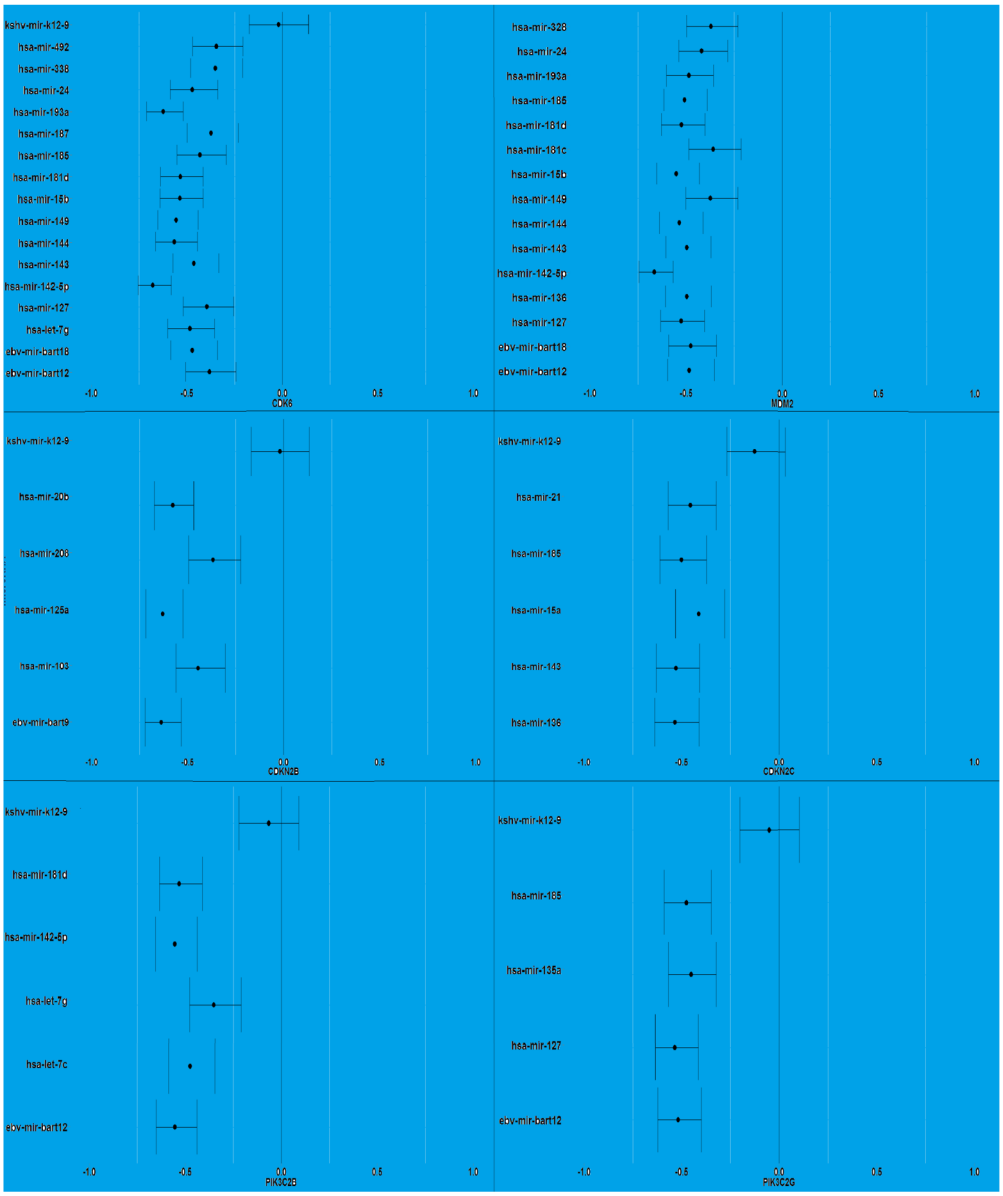

Figure 3: Parent miRNAs depicted with the strength of negative partial correlations along with 99% credible intervals for important mRNA–miRNA associations. The mRNAs are labeled on the horizontal axis, while the miRNAs are labeled on the vertical. Refer Table 5 in main article.

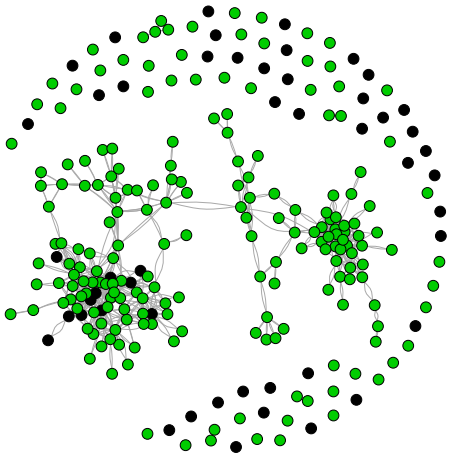

Figure 4: Estimated graph for mRNA–miRNA. Black nodes represent mRNA, and green nodes are miRNA. The edges reflect connections between pairs of miRNAs and pairs of mRNAs, and also miRNA-mRNA connections corresponding to negative partial correlations. In biological terms, the partial correlation between a miRNA and its target mRNA measures the association between the two, after accounting for the remaining mRNAs and miRNAs, with negative associations implying a down-regulation of target mRNAs. Our analysis identified several hub genes based on mRNA expressions, EFGR, RAF1, NF1, SPRY2, CDKN2A, and PIK3C2G, with such nodes having greater than 8 neighbors. In addition, there were 9 hub miRNAs each having greater than 10 neighbors, while 107 miRNAs did not have any neighbors. These hub nodes were ebv-mir-bart7, hsa-mir-106a, hsa-mir-142-3p, hsa-mir-17-5p, hsa-mir-let7, hsa-mir-181c, hsa-mir-184, hsa-mir-19a, and hsa-mir-20a. Such hub nodes are known to modulate important transcription factors and signaling molecules including genes such as MYC, CCLE1 and CLDND1 which have been shown to be associated with cancer, inflammatory response and connective tissue disorders.

overcomes several difficulties for existing Bayesian graphical modeling approaches such as sensitivity to the prior on the model space, assumptions on the underlying graphical structure, and mixing issues associated with discrete mixture approaches. In addition, it has attractive theoretical properties and is scalable to high dimensional settings. Our work makes a timely and important contribution to the sparse but appealing body of work involving systematic decision theoretic Bayesian approaches for efficient graphical model computation.

We demonstrate the numerical advantages of the proposed approach over commonly used Bayesian and penalized approaches using extensive simulation studies. The method is applied to an important problem of target prediction using genomic data, where main scientific question of interest is to find major miRNA regulators of mRNA expression in individuals with tumors and characterize uncertainty of the resulting estimators, by jointly analyzing mRNA and miRNA data. Our inferential methodology gives a

mRNA	miRNA	ρ	ρ^L	ρ^U	mRNA	miRNA	ρ	ρ^L	ρ^U
CDKN2C	hsa-mir-136	-0.53	-0.64	-0.41	CDKN2B	ebv-mir-bart9	-0.63	-0.72	-0.53
CDKN2C	hsa-mir-143	-0.53	-0.63	-0.40	CDKN2B	hsa-mir-125a	-0.63	-0.71	-0.52
CDKN2C	hsa-mir-185	-0.50	-0.61	-0.37	CDKN2B	hsa-mir-20b	-0.57	-0.67	-0.47
CDKN2C	hsa-mir-21	-0.45	-0.57	-0.32	CDKN2B	hsa-mir-103	-0.44	-0.56	-0.30
CDKN2C	hsa-mir-15a	-0.41	-0.53	-0.27	CDKN2B	hsa-mir-208	-0.37	-0.49	-0.22
CDKN2C	kshv-mir-k12-9	-0.12	-0.27	0.032	CDKN2B	kshv-mir-k12-9	-0.02	-0.17	0.13
RB1	hsa-mir-10a	-0.65	-0.74	0.60	KRAS	hsa-mir-144	-0.64	-0.71	-0.53
RB1	hsa-mir-137	-0.59	-0.69	-0.48	KRAS	hsa-mir-10a	-0.60	-0.69	0.57
RB1	hsa-mir-135a	-0.56	-0.66	-0.44	KRAS	hsa-mir-149	-0.55	-0.64	-0.43
RB1	hsa-mir-148b	-0.51	-0.62	-0.39	KRAS	hsa-mir-28	-0.40	-0.52	-0.27
RB1	ebv-mir-bart18	-0.51	-0.61	-0.38	PIK3C2G	hsa-mir-127	-0.53	-0.63	-0.41
RB1	hsa-mir-193a	-0.51	-0.62	-0.38	PIK3C2G	ebv-mir-bart12	-0.52	-0.62	-0.40
RB1	hsa-mir-149	-0.49	-0.61	-0.37	PIK3C2G	hsa-mir-185	-0.47	-0.58	-0.34
CDK6	hsa-mir-142-5p	-0.67	-0.75	-0.58	PIK3C2G	hsa-mir-135a	-0.45	-0.56	-0.32
CDK6	hsa-mir-193a	-0.62	-0.71	-0.52	PIK3C2G	kshv-mir-k12-9	-0.05	-0.20	0.10
CDK6	hsa-mir-144	-0.56	-0.66	-0.44	MDM2	hsa-mir-142-5p	-0.66	-0.74	-0.57
CDK6	hsa-mir-149	-0.55	-0.65	-0.44	MDM2	hsa-mir-15b	-0.55	-0.65	-0.43
CDK6	hsa-mir-15b	-0.53	-0.64	-0.41	MDM2	hsa-mir-127	-0.53	-0.63	-0.40
CDK6	hsa-mir-181d	-0.53	-0.63	-0.41	MDM2	hsa-mir-144	-0.53	-0.64	-0.41
CDK6	ebv-mir-bart18	-0.47	-0.58	-0.34	MDM2	hsa-mir-181d	-0.52	-0.62	-0.40
CDK6	hsa-mir-24	-0.47	-0.58	-0.33	MDM2	hsa-mir-185	-0.51	-0.61	-0.39
CDK6	hsa-mir-143	-0.47	-0.57	-0.33	MDM2	hsa-mir-136	-0.50	-0.60	-0.37
CDK6	hsa-let-7g	-0.47	-0.60	-0.35	MDM2	hsa-mir-143	-0.50	-0.60	-0.37
CDK6	hsa-mir-185	-0.43	-0.55	-0.29	MDM2	ebv-mir-bart12	-0.48	-0.59	-0.35
CDK6	ebv-mir-bart12	-0.38	-0.51	-0.24	MDM2	hsa-mir-193a	-0.48	-0.60	-0.36
CDK6	hsa-mir-127	-0.39	-0.52	-0.25	MDM2	ebv-mir-bart18	-0.47	-0.59	-0.34
CDK6	hsa-mir-187	-0.37	-0.49	-0.23	MDM2	hsa-mir-24	-0.42	-0.53	-0.28
CDK6	hsa-mir-338	-0.35	-0.47	-0.20	MDM2	hsa-mir-328	-0.37	-0.49	-0.23
CDK6	hsa-mir-492	-0.34	-0.47	-0.20	MDM2	hsa-mir-149	-0.37	-0.50	-0.23
CDK6	kshv-mir-k12-9	-0.02	-0.17	0.13	MDM2	hsa-mir-181c	-0.37	-0.48	-0.21
PIK3C2B	ebv-mir-bart12	-0.55	-0.65	-0.43	PIK3C2B	hsa-mir-142-5p	-0.55	-0.65	-0.43
PIK3C2B	hsa-let-7c	-0.47	-0.58	-0.34	PIK3C2B	hsa-mir-181d	-0.53	-0.63	-0.41
PIK3C2B	hsa-let-7g	-0.35	-0.47	-0.20	PIK3C2B	kshv-mir-k12-9	-0.06	-0.22	0.09

Table 3: Analysis of gene regulatory network showing important miRNAs and target mRNAs having negative partial correlations (ρ), along with point-wise 99% credible intervals (ρ^L, ρ^U). For a particular mRNA, the miRNAs are ranked in order of the magnitude of its regulatory effects.

holistic systems-level view of the mRNA-miRNA regulatory mechanisms using graphical modeling by detecting miRNA targets conditional on other miRNAs and mRNAs in the relevant biological pathways. In addition our framework provides a probabilistic quantification of potential targets that can be used for miRNA target ranking and subsequent biological and experimental validation.

While the approach provides an efficient computational method for graphical model selection, there are some potential limitations. For example, it does not immediately yield a positive definite precision matrix with exact zeros corresponding to absent edges, and it is only possible to obtain such an estimate after implementing an additional post-processing algorithm. Moreover the current method is limited to high dimensions involving several hundred nodes, but may not be scalable to extremely high dimensions involving tens of thousands of nodes which is encountered in genome-wide association studies. In future work, our goal is to extend the methodology, computation, and theory to $p >> n$ settings. In addition we would like to generalize the regularized inverse shrinkage prior to more general class of priors which induce different types of shrinkage marginally on the elements of inverse covariance matrix.

Supplementary Material

Supplementary Materials for "Efficient Bayesian Regularization for Graphical Model Selection" (DOI: 10.1214/17-BA1086SUPP; .pdf).

References

Ambros, V. (2004). "The functions of animal microRNAs." *Nature*, 431, 350–355.

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. MR2201362. doi: https://doi.org/10.1093/biomet/92.2.317. 450

Baladandayuthapani, V., Ji. Y., Talluri, R., Nieto-Barajas, L. E. and Morris, J. S. (2010). "Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data." *Journal of American Statistical Association*, 105, 1358–1375. MR2796556. doi: https://doi.org/10.1198/jasa.2010.ap09250. 460

Bondell, H. D. and Reich, B. J. (2012). "Consistent high-dimensional Bayesian variable selection via penalized credible regions." *Journal of the American Statistical Association*, 107, 1610–1624. MR3036420. doi: https://doi.org/10.1080/01621459.2012.716344. 450, 451, 455, 456, 457, 459

Bickel, P. J. and Levina, E. (2008). "Covariance regularization by thresholding". *Annals of Statistics*, 36(6), 2577–2604. doi: https://doi.org/10.1214/08-AOS600. 451

Cai, T. and Liu, W. (2012). "Adaptive Thresholding for Sparse Covariance Matrix Estimation". *Journal of the American Statistical Association*, 106, 672–684. doi: https://doi.org/10.1198/jasa.2011.tm10560. 451

Cancer Genome Atlas Research Network (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature*, 455, 1061-8. 467

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling sparsity via the horseshoe." *Journal of Machine Learning Research W&CP*, 5, 73–80. 451, 456

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97, 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 451, 456

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96(3), 497–512. MR2538753. doi: https://doi.org/10.1093/biomet/asp017. 452

Chan, J. C. and Jeliazkov, I. (2009). "Estimation of Restricted Covariance Matrices." *Journal of the Computational and Graphical Statistics*, 18(2), 457–480. 460

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. MR1241267. doi: https://doi.org/10.1214/aos/1176349260. 450

Dellaportas, P., Giudici, P. and Roberts, G. (2003). "Bayesian inference for non-decomposable graphical Gaussian models." *Sankhyā*, 65, 43–55. MR2016776. 450

Delfino K. R., Serão, N. V., Southey, B. R., Rodriguez-Zas, S. L. (2011). "Therapy-, gender- and race-specific microRNA markers, target genes and networks related to glioblastoma recurrence and survival." *Cancer Genomics Proteonomics*, 8, 173–183. 468

Dempster, A. P. (1972). "Covariance Selection." *Biometrics*. 28, 157–175. 449

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate Priors for Exponential Families." *Annals of Statistics*, 7, 269–281. MR0520238. 450

Dong, H., Luo, L., Hong, S., Siu, H., Xiao, Y., Jin, L., Chen, R., and Xiong, M. (2010). "Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma." *BMC Systems Biology*, 4, 163. 468

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *The Annals of Statistics*, 32, 407–499. MR2060166. doi: https://doi.org/10.1214/009053604000000067. 455

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9, 432–441. 461, 462

Fitch, M. A., Jones, M. B., and Massam, H. (2014). "The Performance of Covariance Selection Methods That Consider Decomposable Models Only." *Bayesian Analysis*, 9, 659–684. MR3256059. doi: https://doi.org/10.1214/14-BA874. 450

Fouskakis, D.; Ntzoufras, I.; Draper, D. (2009). "Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care." *Annals Applied Statistics* 3, 663–690. doi: https://doi.org/10.1214/08-AOAS207. 451

Frühwirth-Schnatter, S. and Tüchler, R. (2008). "Bayesian parsimonious covariance estimation for hierarchical linear mixed models." *Statistics and Computing*, 18, 1–13. MR2416434. doi: https://doi.org/10.1007/s11222-007-9030-2. 451

George, E. I. and McCulloch, R. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88, 881–889. 450

Giudici, P. and Green, P. J. (1999). "Decomposable Graphical Gaussian Model Determination." *Biometrika*, 86, 785–801. MR1741977. doi: https://doi.org/10.1093/biomet/86.4.785. 450, 462

Green, P. J. and Thomas, A. (2013). "Sampling decomposable graphs using a Markov chain on junction trees." *Biometrika*, 100, 91–110. MR3034326. doi: https://doi.org/10.1093/biomet/ass052. 450

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association*, 110, 435–448. MR3338514. doi: https://doi.org/10.1080/01621459.2014.993077. 451, 458, 461, 462, 464, 466

Herranz, H. and Cohen, S. M. (2010). "MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems." *Genes and Development*, 24, 1339–44.

Huang, A. and Wand, M. P. (2013). "Simple marginally non-informative prior distributions for covariance matrices." *Bayesian Analysis*, 8, 439–452. MR3066948. doi: https://doi.org/10.1214/13-BA815. 453, 462

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. MR2210226. doi: https://doi.org/10.1214/088342305000000304. 450, 454

Kundu, S. and Dunson, D. B. (2014). "Bayes variable selection in semi-parametric linear models." *Journal of the American Statistical Association*, 109, 437–447. MR3180575. doi: https://doi.org/10.1080/01621459.2014.881153. 450

Kundu, S., Mallick, B. K., and Baladandayuthapani, V. (2018). "Supplementary Materials for "Efficient Bayesian Regularization for Graphical Model Selection"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1086SUPP. 454

Lee, S. T., Chu, K., Oh, H. J., Im, W. S., Lim, J. Y., Kim, S. K., Park, C. K., Jung, K. H., Lee, S. K., Kim, M., and Roh, J. K. (2011). "Let-7 microRNA inhibits the proliferation of human glioblastoma cells", *Journal of Neuro-Oncology*, 102, 19–24. doi: https://doi.org/10.1214/17-BA1086SUPP. 468

Lenkoski, A. and Dobra, A. (2011). "Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior." *Journal of Computational and Graphical Statistics*, 20, 140–157. MR2816542. doi: https://doi.org/10.1198/jcgs.2010.08181. 450

Lewis, B. P., Burge, C. B., Bartel, D. P. (2005). "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets." *Cell*, 120, 15–20. 468

Lv, J. and Fan, Y. (2009). "A unified approach to model selection and sparse recovery using regularized least squares." *Annals of Statistics*, 37, 3498–3528. 456

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional Graphs and Variable Selection with the Lasso", *Annals of Statistics*, 34, 1436–1462. MR2278363. doi: https://doi.org/10.1214/009053606000000281. 454, 462, 464, 466

Mohammadi, A. and Wit, E. C. (2015). "Bayesian Structure Learning in Sparse Gaussian Graphical Models." *Bayesian Analysis*, 10, 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 450

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., Montana, G. (2014). "Estimating time-varying brain connectivity networks from functional MRI time series." *NeuroImage*, 103, 427–443. 459

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A. and Coombes, K. R. (2008). "Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models." *Biometrics*, 64, 479–489. MR2432418. doi: https://doi.org/10.1111/j.1541-0420.2007.00895.x. 460

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5, 155–176. 460

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). "Partial Correlation Estimation by Joint Sparse Regression Models." *Journal of the American Statistical Association*, 104, 735–746. 463

Polson, N. G. and Scott, J. (2011). "Shrink globally, act locally: sparse Bayesian regularization and prediction." In *Bayesian Statistics 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.)*, pages 501–538. Oxford University Press, New York. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 451, 456

Pourahmadi, M. (1999). "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation." *Biometrika*, 86, 677–690. MR1723786. doi: https://doi.org/10.1093/biomet/86.3.677. 451

Roverato, A. (2000). "Cholesky decomposition of a hyper inverse Wishart matrix." *Biometrika*, 87, 99–112. MR1766831. doi: https://doi.org/10.1093/biomet/87.1.99. 450

Scott, J. G. and Carvalho, C. M. (2008). "Feature-Inclusion Stochastic Search for Gaussian Graphical Models." *Journal of Computational and Graphical Statistics*, 17, 790–808. MR2649067. doi: https://doi.org/10.1198/106186008X382683. 450, 467

Smith, M. and Kohn, R. (2002). "Parsimonious covariance matrix estimation for longitudinal data." *Journal of the American Statistical Association*, 97, 1141–1153. MR1951266. doi: https://doi.org/10.1198/016214502388618942. 451

Tang, W., Duan, J., Zhang, J. G., and Wang, Y. P. (2013). "Subtyping glioblastoma by combining miRNA and mRNA expression data using compressed sensing-based approach." *EURASIP Journal on Bioinformatics and Systems Biology*, 2. 467

Tibshirani, R. J. (2013). "The lasso problem and uniqueness." *Electronic Journal of Statistics*, 7, 1456–1490. MR3066375. doi: https://doi.org/10.1214/13-EJS815. 457

Wang, H. (2012). "Bayesian Graphical Lasso Models and Efficient Posterior Computation." *Bayesian Analysis*, 7, 771–790. MR3000017. doi: https://doi.org/10.1214/12-BA729. 462, 463

Wang, H. and West, M. (2009). "Bayesian analysis of matrix normal graphical models." *Biometrika*, 96, 821–834. MR2564493. doi: https://doi.org/10.1093/biomet/asp049. 450

Wong, A. J., Ruppert, J. M., Bigner, S. H., Grzeschik, C. H., Humphrey, P. A., Bigner, D. S., and Vogelstein, B. (1992). "Structural alterations of the epidermal growth factor receptor gene in human gliomas." *Proceedings of the National Academy of Sciences of the United States of America*, 89, 2965–2969. 468

Wong, F., Carter, C., and Kohn, R. (2003). "Efficient Estimation of Covariance Selection Models." *Biometrika*, 90, 809–830. MR2024759. doi: https://doi.org/10.1093/biomet/90.4.809. 450

Wu, W. B. and Pourahmadi, M. (2003). "Nonparametric estimation of large covariance matrices of longitudinal data." *Biometrika*, 90, 831–44. MR2024760. doi: https://doi.org/10.1093/biomet/90.4.831. 451

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94, 19–35. MR2367824. doi: https://doi.org/10.1093/biomet/asm018. 463

Zou, H. and Li, R. (2008). "One-step sparse estimates in nonconcave penalized likelihood models (with discussion)." *Annals of Statistics*, 36, 1509–1566. 457

Acknowledgments

Research reported in this publication was supported by National Cancer Institute of the National Institutes of Health under award number R01CA194391, and by award number NIH R01 CA160736. Support was also provided by National Science Foundation under award number NSF DMS-1463233.