Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Bayesian Regularization for Graphical Models With Unequal Shrinkage

Lingrui Gan, Naveen N. Narisetty & Feng Liang

To cite this article: Lingrui Gan, Naveen N. Narisetty & Feng Liang (2019) Bayesian Regularization for Graphical Models With Unequal Shrinkage, Journal of the American Statistical Association, 114:527, 1218-1231, DOI: 10.1080/01621459.2018.1482755

To link to this article: https://doi.org/10.1080/01621459.2018.1482755

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 527, 1218–1231: Theory and Methods
https://doi.org/10.1080/01621459.2018.1482755

Bayesian Regularization for Graphical Models With Unequal Shrinkage

Lingrui Gan, Naveen N. Narisetty, and Feng Liang

Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL

ABSTRACT

We consider a Bayesian framework for estimating a high-dimensional sparse precision matrix, in which adaptive shrinkage and sparsity are induced by a mixture of Laplace priors. Besides discussing our formulation from the Bayesian standpoint, we investigate the MAP (maximum a posteriori) estimator from a penalized likelihood perspective that gives rise to a new nonconvex penalty approximating the ℓ_0 penalty. Optimal error rates for estimation consistency in terms of various matrix norms along with selection consistency for sparse structure recovery are shown for the unique MAP estimator under mild conditions. For fast and efficient computation, an EM algorithm is proposed to compute the MAP estimator of the precision matrix and (approximate) posterior probabilities on the edges of the underlying sparse structure. Through extensive simulation studies and a real application to a call center data, we have demonstrated the fine performance of our method compared with existing alternatives. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2017
Revised April 2018

KEYWORDS
Bayesian regularization; Precision matrix estimation; Sparse Gaussian graphical model; Spike-and-slab priors

1. Introduction

Covariance matrix and precision matrix (inverse of the covariance matrix) are among the most fundamental quantities in Statistics as they describe the dependence between different variables (components) of a multivariate observation. Not surprisingly, they play pivotal roles in many statistical problems including graphical models, classification, clustering, and regression, which are used extensively in many application areas including biological, engineering, and finance. Take the Gaussian graphical model (GGM) as an example. The precision matrix provides great insight into the conditional dependence structure in a graph, since the conditional independence of ith and jth variables of an undirected Gaussian Markov random field is equivalent to the (i, j)th entry of the precision matrix being zero, see a recent review by Pourahmadi (2013). Such results have helped researchers to identify complex network structures in applications such as high-throughput biological data, for example, in Wille et al. (2004).

Estimating the precision matrix, especially under the high-dimensional setting where the variable dimension p can possibly be larger than the sample size n, is a particularly challenging problem. Given the current prevalence of high-dimensional data and the wide utility of precision matrix, this problem has received significant attention in recent literature. When the sample covariance matrix is positive definite, its inverse is a natural estimator for the precision matrix. However, the inverse of sample covariance matrix as an estimator is demonstrated to have poor performance in numerous studies (Johnstone 2001; Paul 2007; Pourahmadi 2013). Moreover, when $p > n$, the precision matrix estimation problem is ill-posed without further restricting assumptions. One of the most commonly used assumptions to remedy this issue is to assume that the precision matrix is sparse, that is, a large majority of its entries are zero (Dempster 1972), which turns out to be quite useful in practice in the aforementioned GGM owing to its interpretability. Another possibility is to assume a sparse structure on the covariance matrix through, for example, a sparse factor model (Carvalho et al. 2008; Fan, Fan, and Lv 2008; Fan, Liao, and Mincheva 2011; Bühlmann and Van De Geer 2011; Pourahmadi 2013; Ročková and George 2016a), to obtain a sparse covariance matrix estimator, and invert it to estimate the precision matrix. However, the precision matrix estimator obtained from this strategy is not guaranteed to be sparse, which is important for interpretability in our context.

Regularization provides a general framework for dealing with high-dimensional problems. There are two major approaches that use regularization to estimate the precision matrix and its sparse structure.

The first one is *regression-based approach* where a sparse regression model is estimated separately for each column to identify and estimate the nonzero elements of that column in the precision matrix Θ (Meinshausen and Bühlmann 2006; Peng et al. 2009; Zhou, van de Geer, and Bühlmann 2009; Khare, Oh, and Rajaratnam 2015). This approach focuses more on the sparse selection of the entries, and the estimated precision matrix is generally not positive definite.

The other is *likelihood-based approach* which aims to optimize the negative log-likelihood function (1) together with an element-wise penalty term on Θ (Yuan and Lin 2007; Banerjee, El Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008; Fan, Feng, and Wu 2009). Among these methods, graphical Lasso (GLasso; Friedman, Hastie, and

CONTACT Naveen N. Narisetty naveen@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Tibshirani 2008) is the most commonly used owing to its scalability. GLasso estimator for the precision matrix is also not guaranteed to be positive definite. Mazumder and Hastie (2012) proposed algorithms that modify GLasso and ensure positive definiteness of the estimated precision matrix. Apart from these two general approaches, regularization can be applied with other forms of loss functions, an example of which is the CLIME estimator proposed by Cai, Liu, and Luo (2011).

Theoretical properties of the likelihood-based methods for Gaussian graphical models have been studied in the literature. In Rothman et al. (2008), Lam and Fan (2009), and Loh and Wainwright (2015), estimation error rates in Frobenius norm have been established for likelihood-based estimators with Lasso and SCAD penalties. For GLasso, stronger results in entrywise maximum norm are obtained by Ravikumar et al. (2011) under a restrictive assumption on Θ, called the irrepresentable assumption, when the multivariate distribution of the observations has an exponential tail (such as sub-Gaussian distributions). A slower rate is shown when the distribution has a polynomial tail (such as t-distributions with sufficiently large degrees of freedom). Similar results on estimation error rate in maximum norm are shown by Loh and Wainwright (2017) for nonconvex penalized estimators under sub-Gaussian distributions but their results require beta-min conditions. Cai, Liu, and Luo (2011) provided such results for CLIME estimator both under exponential and polynomial tails with the assumption that all the absolute column sums of Θ are bounded.

The precision matrix estimation problem is less studied under the Bayesian framework possibly due to the high computational cost associated with MCMC when p is large. Marlin and Murphy (2009) proposed a Bayesian model and a variational Bayes algorithm for GGMs with a block structure. Wang (2012) proposed a Bayesian version of GLasso and the associated posterior computation algorithms. Carvalho and Scott (2009), Dobra, Lenkoski, and Rodriguez (2011), Wang and Li (2012), and Mohammadi et al. (2015) used G-Wishart priors and proposed stochastic search methods for the computation. Banerjee and Ghosal (2015) studied a Bayesian approach with mixture prior distributions that have a point-mass and a Laplace distribution. They provided posterior consistency results and a computational approach using Laplace approximation. With the exception of Banerjee and Ghosal (2015), theoretical properties of Bayesian methods for sparse precision matrix estimation have not been studied. The results of Banerjee and Ghosal (2015) are on estimation error rate in Frobenius norm similar to those of Rothman et al. (2008), but assume the underlying distribution to be Gaussian.

In this article, we propose a new Bayesian approach for estimation and structure recovery for GGMs. Specifically, to achieve adaptive shrinkage, we model the off-diagonal elements of Θ using a continuous spike-and-slab prior with a mixture of two Laplace distributions, which is known as the spike-and-slab Lasso prior in Ročková (2018), Ročková and George (2016a), and Ročková and George (2016b). Continuous spike-and-slab priors are commonly used for high-dimensional regression (George and McCulloch 1993; Ishwaran and Rao 2005; Narisetty and He 2014) and a Gibbs sampling algorithm is often used for posterior computation. However, such a Gibbs sampler for our problem has an extremely high computational burden and instead we propose a novel EM algorithm for computation, which is motivated by the EM algorithm for linear regression from Ročková and George (2014) and the one for factor models from Ročková and George (2016a). Our novel computational and theoretical contributions in the article are summarized as follows:

- We propose a new approach for precision matrix estimation, named BAGUS, short for "Bayesian regularization for Graphical models with Unequal Shrinkage." The adaptive (unequal) shrinkage is due to the nonconvex penalization by our Bayesian formulation.
- Although the Gaussian likelihood is used in our Bayesian formulation, our theoretical results hold beyond GGMs. We have shown that our procedure enjoys the optimal estimation error rate of $O_p(\sqrt{\frac{\log p}{n}})$ in the entrywise maximum norm and selection consistency under both exponential and polynomial tail distributions with very mild conditions. Our theoretical result is stronger than the best existing result by Cai, Liu, and Luo (2011), as we assume boundedness of Θ in operator norm which is weaker than the assumption of bounded absolute column sum of Θ.
- We propose a fast EM algorithm which produces a maximum a posteriori (MAP) estimate of the precision matrix and (approximate) posterior probabilities on all edges that can be used to learn the graph structure. The EM algorithm has computational complexity comparable to the state-of-the-art GLasso algorithm (Mazumder and Hastie 2012).
- Our algorithm is guaranteed to produce a symmetric and positive definite estimator unlike many existing estimators including CLIME.

The remaining part of the article is organized as follows. In Section 2, we present our model and prior set-up in the Bayesian framework along with a discussion on its penalized likelihood perspective. In Section 3, we provide our theoretical consistency results followed by the details of the EM algorithm in Section 4. Section 5 presents numerical results in extensive simulation studies and a real application for predicting telephone center call arrivals. Proofs, technical details, and R code used for empirical results can be found in online supplementary material.

Notation

For a $p \times q$ matrix $A = [a_{ij}]$, we denote its Frobenius norm by $\|A\|_F = \sqrt{\sum_{(i,j)} a_{ij}^2}$, the entrywise ℓ_∞ norm (i.e., maximum norm) $\|A\|_\infty = \max_{(i,j)} |a_{ij}|$, and its spectral norm by $\|A\|_2 = \sup\{\|A\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^q, \|\mathbf{x}\| \le 1\}$ where $\|\mathbf{x}\|$ denotes the l_2 norm of vector $\mathbf{x}$. For a $p \times p$ square matrix A, let A^- denote the off-diagonal elements of A, A^+ the diagonal elements of A, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the smallest and the largest eigenvalues, respectively. For a square symmetric matrix A, its spectral norm is equal to its maximum eigenvalue, that is, $\|A\|_2 = \lambda_{\max}(A)$, and its maximum absolute column sum (i.e., the ℓ_1/ℓ_1 operator norm) is the same as its maximum absolute row sum (i.e., the ℓ_∞/ℓ_∞ operator norm), denoted by $|||A|||_\infty = \max_{1 \le j \le q} \sum_{i=1}^p |a_{ij}|$.

Let $\Theta^0 = [\theta_{ij}^0]$ and $\Sigma^0 = [\sigma_{ij}^0]$ denote the true precision matrix and covariance matrix, respectively. Let $S^0 = \{(i, j) :$

$\theta_{ij}^0 \neq 0\}$ denote the index set of all nonzero entries in Θ^0 and S^{0c} is its complement. Define $\theta_{\max}^0 = \max_{ij} |\theta_{ij}^0|$ and $M_{\Sigma^0} = \left|\left|\left|\Sigma^0\right|\right|\right|_\infty$. Define $\Gamma = \Theta^{-1} \otimes \Theta^{-1}$ as the Hessian matrix of $g := -\log\det(\Theta)$. $\Gamma_{(j,k),(l,m)}$ corresponds to the second partial derivative $\frac{\partial^2 g}{\partial\theta_{jk}\partial\theta_{lm}}$, and for any two subsets T_1 and T_2 of $\{(i, j)\colon 1 \leq i, j \leq p\}$, we use $\Gamma_{T_1T_2}$ to denote the matrix with rows and columns of Γ indexed by T_1 and T_2, respectively. We further denote $M_{\Gamma^0} = \left|\left|\left|\Gamma^{0\,-1}_{S^0S^0}\right|\right|\right|_\infty = \left|\left|\left|(\Theta^0 \otimes \Theta^0)_{S^0S^0}\right|\right|\right|_\infty$. Define the column sparsity $d = \max_{i=1,2,\ldots,p} \text{card}\{j : \theta_{ij}^0 \neq 0\}$ and the off-diagonal sparsity $s = \text{card}(S^0) - p$, where *card* denotes the cardinality of the set in its argument.

2. Bayesian Regularization for Graphical Models

Our data consist of a random sample of n observations $Y_1, \ldots, Y_n$ which are assumed to be iid p-variate random vectors following a multivariate distribution with mean zero and precision matrix Θ. In short, we use the following notation:

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(0, \Theta^{-1}).$$

Our primary goal is to estimate Θ and identify the sparse structure in the elements of Θ. For the Bayesian framework, we work with the Gaussian log -likelihood given by

$$\ell(\Theta) = \log f(Y_1, \ldots, Y_n|\Theta) = \frac{n}{2}\Big(\log\det(\Theta) - \text{tr}(S\Theta)\Big), \tag{1}$$

where $S = [s_{ij}] = \frac{1}{n}\sum Y_i Y_i^t$ denotes the sample covariance matrix of the data. We note that in spite of working with the Gaussian likelihood, we allow the observations to have non-Gaussian distributions including those with polynomial tails.

2.1. Bayesian Formulation

Next we describe our prior specification on the following two groups of parameters: the diagonal entries $\{\theta_{ii}\}$ and the off diagonal entries, where the latter is reduced to the upper triangular entries $\{\theta_{ij}\colon i < j\}$ due to symmetry.

On the upper triangular entries θ_{ij} $(i < j)$, we place the following spike-and-slab prior, known as the spike-and-slab Lasso prior developed in a series of work by Ročková (2018), Ročková and George (2016a), and Ročková and George (2016b):

$$\pi(\theta_{ij}) = \frac{\eta}{2v_1}\exp\Big\{-\frac{|\theta_{ij}|}{v_1}\Big\} + \frac{1-\eta}{2v_0}\exp\Big\{-\frac{|\theta_{ij}|}{v_0}\Big\}, \tag{2}$$

which is a mixture of two Laplace distributions of different scales v_0 and v_1 with $v_1 > v_0 > 0$. The mixture distribution (2) represents our prior on θ_{ij} which could take values of relatively large magnitude modeled by the Laplace distribution with scale parameter v_1 (i.e., the "slab" component), or which could take values of very small magnitude modeled by the Laplace distribution with scale parameter v_0 (i.e., the "spike" component). In the traditional spike-and-slab prior, the "spike" component is set to be a point mass at zero, which corresponds to our setting with $v_0 = 0$. Here, we use a continuous version of the spike-and-slab prior, in which v_0 is set be nonzero but relatively small compared with v_1. Continuous spike-and-slab priors with normal components were proposed by George and McCulloch (1993) in the linear regression context and their high-dimensional shrinkage properties were studied by Ishwaran and Rao (2005) and Narisetty and He (2014). Ročková (2016) and Ročková and George (2016b) considered the spike-and-slab Lasso prior given by (2) for linear regression and studied the adaptive shrinkage property of such priors as well as various asymptotic properties concerning the posterior mode. An advantage of continuous spike-and-slab priors is that the continuous prior distributions on θ_{ij} allow the use of efficient algorithms that do not require switching the active dimension of the parameter.

For the diagonal entries θ_{ii} of the precision matrix, a weakly informative Exponential prior is specified since θ_{ii} do not need to be shrunk to zero:

$$\pi(\theta_{ii}) = \tau\exp(-\tau\theta_{ii})\mathbb{1}(\theta_{ii} > 0).$$

Although Θ can be fully parameterized by these two groups of parameters, they are not independent as the determinant of Θ needs to be positive. Therefore, the support for the joint prior distribution on elements of Θ is restricted such that Θ is positive definite, that is, $\Theta \succ 0$. In addition, we constrain the spectral norm of Θ to be upper bounded: $\|\Theta\|_2 \leq B$. Such a constraint is not very restrictive since it often appears in the assumptions for theoretical studies of precision matrix estimation anyway: a large spectral norm of Θ implies high correlation among variables, a setup in which most methods fail. An important consequence of this constraint will be discussed in Section 2.3.

So our prior distribution on Θ is given by

$$\pi(\Theta) = \prod_{i<j}\pi(\theta_{ij})\prod_i \pi(\theta_{ii})\mathbb{1}(\Theta \succ 0)\mathbb{1}(\|\Theta\|_2 \leq B). \tag{3}$$

2.2. The Penalized Likelihood Perspective

If estimation of Θ is of main interest, then a natural choice is the MAP estimator $\tilde{\Theta}$ that maximizes the posterior distribution $\pi(\Theta|Y_1, \ldots, Y_n)$. This is equivalent to minimizing the following objective function under the constraint $\|\Theta\|_2 \leq B$ and $\Theta \succ 0$:

$$\begin{aligned} L(\Theta) &= -\log\pi(\Theta|Y_1, \ldots, Y_n) \\ &= -\ell(\Theta) - \sum_{i<j}\log\pi(\theta_{ij}|\eta) - \sum_i \log\pi(\theta_{ii}|\tau) + \text{Const.} \\ &= \frac{n}{2}\Big(\text{tr}(S\Theta) - \log\det(\Theta)\Big) + \sum_{i<j}\text{pen}_{SS}(\theta_{ij}) \\ &\quad + \sum_i \text{pen}_1(\theta_{ii}) + \text{Const.}, \end{aligned} \tag{4}$$

where

$$\text{pen}_{SS}(\theta) = -\log\left[\Big(\frac{\eta}{2v_1}\Big)e^{-\frac{|\theta|}{v_1}} + \Big(\frac{1-\eta}{2v_0}\Big)e^{-\frac{|\theta|}{v_0}}\right] \tag{5}$$

and $\text{pen}_1(\theta) = \tau|\theta|$.

If viewed from the penalized likelihood perspective, the objective function $L(\Theta)$ employs two penalty functions, induced by our Bayesian formulation. The penalty function on the diagonal entries, $\text{pen}_1(\theta)$, is the same as the Lasso penalty. The hyperparameter τ is suggested to be small, so the Lasso penalty mainly shrinks the estimates of θ_{ii} instead of truncating them to be zero.

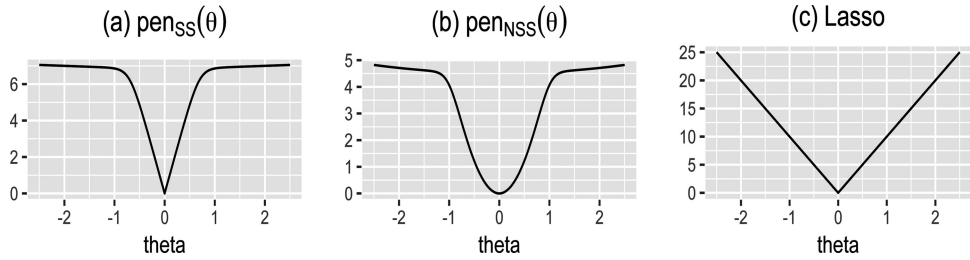

Figure 1. Plot of different penalty functions. (a) Penalty induced from the spike-and-slab prior with a mixture of Laplace distributions; (b) penalty induced from the spike-and-slab prior with a mixture of normal distributions; and (c) Lasso penalty.

More importantly, the penalty function on the off-diagonal entries, $\text{pen}_{\text{SS}}(\theta)$, coming from the spike-and-slab prior has an interesting shrinkage property. To highlight the difference between this penalty and the Lasso penalty, we plotted them in Figure 1. We also compare our spike-and-slab penalty with the spike-and-slab penalty that arises by using a mixture of two normal distributions (George and McCulloch 1997) instead of Laplace distributions:

$$\text{pen}_{\text{NSS}}(\theta) = -\log\left[\left(\frac{\eta}{\sqrt{2\pi}v_1}\right)e^{-\frac{\theta^2}{2v_1}} + \left(\frac{1-\eta}{\sqrt{2\pi}v_0}\right)e^{-\frac{\theta^2}{2v_0}}\right],$$

where "NSS" in the subscript stands for normal spike-and-slab prior. In Figure 1, we set $v_0 = 0.1$ and $v_1 = 10$ for both $\text{pen}_{\text{SS}}(\theta)$ and $\text{pen}_{\text{NSS}}(\theta)$. Also, we subtract their values at 0 so the corresponding penalty at $\theta = 0$ is zero. We can see that the penalty function we use, $\text{pen}_{\text{SS}}(\theta)$, provides the best continuous approximation of the L_0 penalty among the three.

To gain more insight about the penalty functions, we plot the derivatives/subgradient of the spike-and-slab penalty $\text{pen}_{\text{SS}}(\theta)$ in Figure 2. A simple calculation reveals that

$$\begin{aligned}\frac{\partial}{\partial|\theta|}\text{pen}_{\text{SS}}(\theta) &= \frac{1}{v_1}\frac{\frac{\eta}{2v_1}e^{-\frac{|\theta|}{v_1}}}{\pi(\theta)} + \frac{1}{v_0}\frac{\frac{1-\eta}{2v_1}e^{-\frac{|\theta|}{v_0}}}{\pi(\theta)} \\ &= \frac{w(\theta)}{v_1} + \frac{1-w(\theta)}{v_0}, \end{aligned} \tag{6}$$

which is a weighted average of $1/v_1$ and $1/v_0$ with the weight $w(\theta)$ being the conditional probability of θ belonging to the "slab" component (Ročková and George 2016b). Recall that the derivative of a penalty function should ideally have its maximum at zero and then decay gradually to 0 (asymptotically), because a nondecreasing derivative with respect to $|\theta|$ leads to a bias and affects the performance in finite sample settings (Fan and Li 2001; Loh and Wainwright 2017). This is the case with $\text{pen}_{\text{SS}}(\theta)$: As $|\theta|$ becomes larger, the mixing weight w gets larger, which leads to a smooth transition from a large penalty $1/v_0$ produced from the "spike" component, to a smaller penalty $1/v_1$ from the "slab" component. From Figure 2, we can see that $\text{pen}_{\text{NSS}}(\theta)$ does not have this desired property, and neither does the Lasso penalty.

2.3. Posterior Maximization and Local Convexity

The nonconvexity of our spike-and-slab penalty $\text{pen}_{\text{SS}}(\theta)$ leads to desired shrinkage and selection behavior, but it could bring additional computation challenges as the posterior objective function $L(\Theta)$ is no longer convex and may have multiple local optima. However, this is not a problem in our case with the upper bound on the spectral norm of Θ (3). More specifically, the following theorem ensures that the optimization of $L(\Theta)$ with the spectral norm constraint is a convex optimization problem, that is, locally within the spectral norm ball, we are dealing with convex optimization resulting in a unique MAP estimate. This result is motivated by Lemma 6 from Loh and Wainwright (2017).

Theorem 1. If $B < (2nv_0)^{\frac{1}{2}}$, then $\min_{\Theta\succ 0, \|\Theta\|_2 \le B} L(\Theta)$ is a strictly convex problem.

Proof. Decompose $L(\Theta)$ as the sum of the following two terms: $-\ell(\Theta) - \frac{1}{8v_0}\|\Theta\|_F^2$ and $\sum_{i<j}\text{pen}_{\text{SS}}(\theta_{ij}) + \sum_i \text{pen}_1(\theta_{ii}) + \frac{1}{8v_0}\|\Theta\|_F^2$. We prove the theorem by checking that the second-order subgradient of each term in the decomposition of $L(\Theta)$ is positive which would imply that both the terms are strictly convex.

The second-order subgradient of the first term is given by $-\nabla^2\ell(\Theta) - \frac{1}{4v_0}$, where $-\nabla^2\ell(\Theta) = \frac{n}{2}(\Theta\otimes\Theta)^{-1}$.

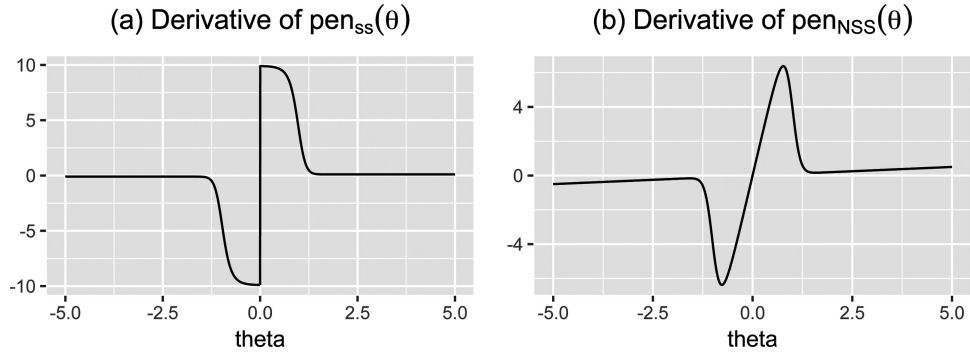

Figure 2. Plot of the derivative/subgradient of the penalty functions.

The smallest eigenvalue of $-\nabla^2\ell(\Theta)$ can be bounded as

$$\lambda_{\min}\left(-\nabla^2\ell(\Theta)\right) - = \frac{n}{2}\lambda^{-1}_{\max}(\Theta\otimes\Theta) = \frac{n}{2}\lambda^{-2}_{\max}(\Theta) > \frac{1}{4v_0},$$

where the last inequality is because $\|\Theta\|_2 \le B \le (2nv_0)^{\frac{1}{2}}$ implies that $\lambda^2_{\max}(\Theta) \le 2nv_0$ and leads to $\frac{n}{2}\lambda^{-2}_{\max}(\Theta) \ge \frac{1}{4v_0}$. Therefore, $-\nabla^2\ell(\Theta) - \frac{1}{4v_0}$ is strictly convex.

We now consider the second-order subgradient of $\text{pen}_{\text{SS}}(\theta_{ij})$:

$$|\text{pen}_{\text{SS}}{}''(\theta_{ij})| = \frac{(\frac{1}{v_0}-\frac{1}{v_1})\frac{\eta v_0}{(1-\eta)v_1}e^{\frac{\theta_{ij}}{v_0}-\frac{\theta_{ij}}{v_1}}}{(\frac{\eta v_0}{(1-\eta)v_1}e^{\frac{\theta_{ij}}{v_0}-\frac{\theta_{ij}}{v_1}}+1)^2} \le \frac{1}{4}\left(\frac{1}{v_0}-\frac{1}{v_1}\right) < \frac{1}{4v_0},$$

where the first inequality is because for any x, $\frac{|x|}{(1+|x|)^2} \le \frac{1}{4}$. This implies that the second term in the decomposition of $L(\Theta)$ is also strictly convex and the theorem is proved. □

2.4. Uncovering the Sparse Structure

In many applications, identifying the zero entries in Θ (referred to as structure estimation or graph selection) is also of major interest along with the estimation of Θ. Inference on the latent sparse structure of Θ or equivalently the sparse structure of a graph can be directly induced from our spike-and-slab prior. We can reexpress the spike-and-slab prior (2) as the following two-level hierarchical prior:

$$\begin{cases} \theta_{ij} \mid r_{ij} = 0 \sim \text{DE}(0, v_0) \\ \theta_{ij} \mid r_{ij} = 1 \sim \text{DE}(0, v_1), \end{cases} \tag{7}$$

where r_{ij} follows:

$$r_{ij} \mid \eta \sim \text{Bern}(\eta). \tag{8}$$

Here $\text{DE}(0, v)$ denotes the double exponential (Laplace) distribution with scale v and $\text{Bern}(\eta)$ denotes the Bernoulli distribution with probability η .

We can view the binary variable r_{ij} as the indicator for the sparsity pattern: $r_{ij} = 1$ implies θ_{ij} being the "signal" (i.e., from the slab component), and $r_{ij} = 0$ implies θ_{ij} being the "noise" (i.e., from the spike component). In the fully Bayesian approach, the posterior inclusion probability for an edge connecting i and j is given by

$$\mathbb{P}(r_{ij} = 1|Y_1, \ldots, Y_n) = \int \mathbb{P}(r_{ij} = 1|\theta_{ij})\pi(\theta_{ij}|Y_1, \ldots, Y_n)d\theta_{ij},$$

which is the integrated probability of θ_{ij} being from the slab component (corresponding to $\gamma_{ij} = 1$) with respect to the posterior distribution of θ_{ij}. In our analysis, we approximate this probability by using the MAP estimator $\tilde{\Theta}$ as follows:

$$p_{ij} = \mathbb{P}(r_{ij} = 1|\tilde{\theta}_{ij}) = \frac{\left(\frac{\eta}{2v_1}\right)e^{-\frac{|\tilde{\theta}_{ij}|}{v_1}}}{\left(\frac{\eta}{2v_1}\right)e^{-\frac{|\tilde{\theta}_{ij}|}{v_1}} + \left(\frac{1-\eta}{2v_0}\right)e^{-\frac{|\tilde{\theta}_{ij}|}{v_0}}}. \tag{9}$$

We can then threshold p_{ij} to identify the edges: if p_{ij} is greater than a prespecified threshold such as 0.5, then the (i, j) pair is identified as an edge.

Denote $\mathbb{P}(r_{ij} = 1|\tilde{\theta}_{ij} = 0)$ by $p^\star(0)$. The quantity $\frac{1}{p^\star(0)} - 1 = v_1(1-\eta)/(v_0\eta)$ represents the interplay of all the parameters (v_0, v_1, η) and it plays an important role both in our asymptotic analysis for precision matrix estimation that will be presented in the next section, and also in the analysis of Ročková and George (2016b) and Ročková (2018) for high-dimensional linear regression.

3. Theoretical Results

Let $\tilde{\Theta}$ denote the MAP estimator, the unique minimizer of the loss function (4). In this section, we provide theoretical results on the estimation accuracy of $\tilde{\Theta}$. We also show that the structure selected based on thresholding the posterior probabilities p_{ij} matches the true sparse structure with probability going to one.

3.1. Conditions

3.1.1. Tail Conditions on the Distribution of Y

In our analysis, we do not restrict to the situation where the true distribution of Y is Gaussian. Instead, we provide analysis for two cases according to the tail conditions on the true distribution of a p-variate random vector $Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(p)})$.

(C1) Exponential tail condition: Suppose that there exists some $0 < \eta_1 < 1/4$ such that $\frac{\log p}{n} < \eta_1$ and

$$Ee^{tY^{(j)^2}} \le K \text{ for all } |t| \le \eta_1, \text{ for all } j = 1, \ldots, p, \tag{10}$$

where K is a bounded constant.

(C2) Polynomial tail condition: Suppose that for some $\gamma, c_1 > 0$, $p \le c_1 n^\gamma$, and for some $\delta_0 > 0$,

$$E|Y^{(j)}|^{4\gamma+4+\delta_0} \le K, \quad \text{for all } j = 1, \ldots, p. \tag{11}$$

Note that when Y follows a Gaussian or a sub-Gaussian distribution, condition (C1) is satisfied. When $p = n$, condition (C2) is satisfied for t-distributions with degrees of freedom greater than 8. When $p = n^2$, condition (C2) is satisfied for t-distributions with degrees of freedom greater than 12. The same tail conditions are also considered by Cai, Liu, and Luo (2011) and Ravikumar et al. (2011).

3.1.2. Conditions on Θ^0

We make the following assumption on the true precision matrix Θ^0 for studying estimation accuracy.

(A1) $\lambda_{\max}(\Theta^0) \le 1/k_1 < \infty$ or equivalently $0 < k_1 \le \lambda_{\min}(\Sigma^0)$, where k_1 is some constant greater than 0.

Note that because the largest eigenvalue of Θ^0 is bounded, all the elements of Θ^0 are bounded, and cannot grow with p and n.

In addition, we make the minimum signal assumption below for studying sparse structure recovery.

(A2) The minimal "signal" entry satisfies $\min_{(i,j)\in S^0}|\theta^0_{ij}| \ge K_0\sqrt{\frac{\log p}{n}}$, where $K_0 > 0$ is a sufficiently large constant not depending on n.

Similar and in some cases stronger assumptions are imposed in other theoretical analysis of precision matrix estimation and sparse structure recovery (Rothman et al. 2008; Lam and Fan 2009; Ravikumar et al. 2011; Cai, Liu, and Luo 2011; Loh and Wainwright 2017). For a comparison of various theoretical results, see the discussion in Section 3.3.

3.2. Theoretical Results

The following theorem gives estimation accuracy under the entrywise ℓ_∞ norm. In particular, the following theorem implies that with an appropriate choice of $(\nu_0, \nu_1, \eta, \tau)$ and B, we could achieve the $O_p(\sqrt{\frac{\log p}{n}})$ error rate for distributions with an exponential or a polynomial tail.

Theorem 2 (Estimation accuracy in entrywise ℓ_∞ norm).

Assume condition (A1) holds. For any predefined constants $C_3 > 0$, $\tau_0 > 0$, define $C_1 = \eta_1^{-1}(2 + \tau_0 + \eta_1^{-1}K^2)$ when the exponential tail condition (C1) holds, and $C_1 = \sqrt{(\theta^0_{\max} + 1)(4 + \tau_0)}$ when the polynomial tail condition (C2) holds. Assume that

(i) the prior hyperparameters ν_0, ν_1, η, and τ satisfy

$$\begin{cases} \frac{1}{n\nu_1} = C_3\sqrt{\frac{\log p}{n}}(1-\varepsilon_1), \quad \frac{1}{n\nu_0} > C_4\sqrt{\frac{\log p}{n}} \\ \frac{\nu_1^2(1-\eta)}{\nu_0^2\eta} \le p^{\varepsilon}, \quad \text{and} \quad \tau \le C_3\frac{n}{2}\sqrt{\frac{\log p}{n}} \end{cases} \tag{12}$$

for some constants $\varepsilon_1 > 0$, $C_4 > C_3$ and some sufficiently small ε,

(ii) the spectral norm B satisfies $\frac{1}{k_1} + 2d(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}} < B < (2n\nu_0)^{\frac{1}{2}}$, and

(iii) the sample size n satisfies $\sqrt{n} \ge M\sqrt{\log p}$,

where $M = \max\{2d(C_1 + C_3)M_{\Gamma^0}\max(3M_{\Sigma^0}, 3M_{\Gamma^0}M_{\Sigma^0}{}^3, \frac{2}{k_1^2}), \frac{2C_3\varepsilon_1}{k_1^2}\}$.

Then, the MAP estimator $\tilde{\Theta}$ satisfies

$$\|\tilde{\Theta} - \Theta^0\|_\infty \le 2(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}} \tag{13}$$

with probability greater than $1 - \delta_1$, where $\delta_1 = 2p^{-\tau_0}$ when condition (C1) holds, and $\delta_1 = O(n^{-\delta_0/8} + p^{-\tau_0/2})$ when condition (C2) holds.

Theorem 2 shows that the estimation error of our MAP estimator $\tilde{\Theta}$ can be controlled through an interplay between the parameters $(\nu_0, \nu_1, \eta, \tau)$ and B. To help readers understand this result, we provide an explanation of the required conditions.

In our proof, the term $\frac{1}{n}\text{pen}'_{SS}(\theta)$, which decreases from $1/(n\nu_0)$ to $1/(n\nu_1)$ when $|\theta|$ increases from zero to infinity, serves as an adaptive thresholding value. The conditions in (12) ensure the following properties of this adaptive thresholding rule: (1) to eliminate noise, $1/(n\nu_0)$ is set to be bigger than $\sqrt{(\log p)/n}$, the typical noise level in high-dimensional analysis; (2) to reduce bias due to thresholding, $1/(n\nu_1)$ is set to be of a constant order of $\sqrt{(\log p)/n}$, or much smaller by varying ε_1; (3) the thresholding level should be close to $1/(n\nu_1)$ when θ is of a certain order bigger than the noise level $\sqrt{(\log p)/n}$, which is ensured by the upper bound on $\frac{\nu_1^2(1-\eta)}{\nu_0^2\eta}$.

The upper bound on B in condition (ii) is to ensure that our objective function $L(\Theta)$ is strictly convex. However, B cannot be too small, otherwise, even if $L(\Theta)$ is convex, the constrained local mode cannot achieve the desired estimation accuracy $\|\tilde{\Theta} - \Theta^0\|_\infty = O_p(\sqrt{\log p/n})$.

When $M_{\Gamma^0}, M_{\Sigma^0}$ remain constant as a function of (n, p, d), Theorem 2 guarantees that with proper tuning, an estimation error bound of $O(\sqrt{\log p/n})$ in ℓ_∞ norm can be achieved for the MAP estimator $\tilde{\Theta}$ with high probability. Similar results can be found in Ravikumar et al. (2011) and Loh and Wainwright (2017) when $M_{\Gamma^0}, M_{\Sigma^0}$ are constants. If $M_{\Gamma^0}, M_{\Sigma^0}$ are of the order $O(p)$, then we require the sample size n to grow faster than the order $O(p)$.

Theorem 2 follows from a more general result stated as Theorem A in Appendix A from the online supporting material. The specific definition for C_4 and the one for ε are also provided in Theorem A in Appendix A in the online supporting material.

We now present the following result on estimation accuracy of $\tilde{\Theta}$ in terms of Frobenius norm, spectral norm, and ℓ_∞/ℓ_∞ operator norm. This result is based on Theorem 2 and Lemma from Appendix A.

Theorem 3 (Estimation accuracy in other norms).

Under the same conditions of Theorem 2,

(i) if the exponential tail condition (C1) holds, then

$$\begin{aligned} &\|\tilde{\Theta} - \Theta^0\|_F < 2\Big(\eta_1^{-1}(2 + \tau_0 + \eta_1^{-1}K^2) + C_3\Big) \\ &\quad \times M_{\Gamma^0}\sqrt{\frac{(p+s)\log p}{n}}, \\ &\Big|\Big|\Big|\tilde{\Theta} - \Theta^0\Big|\Big|\Big|_\infty, \|\tilde{\Theta} - \Theta^0\|_2 < 2\Big(\eta_1^{-1}(2 + \tau_0 + \eta_1^{-1}K^2) + C_3\Big) \\ &\quad \times M_{\Gamma^0}\min\{d, \sqrt{p+s}\}\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{14}$$

with probability greater than $1 - 2p^{-\tau_0}$;

(ii) if the polynomial tail condition (C2) holds, then

$$\begin{aligned} &\|\tilde{\Theta} - \Theta^0\|_F < 2\left(\sqrt{(\theta^0_{\max} + 1)(4 + \tau_0)} + C_3\right) \\ &\quad \times M_{\Gamma^0}\sqrt{\frac{(p+s)\log p}{n}}, \\ &\Big|\Big|\Big|\tilde{\Theta} - \Theta^0\Big|\Big|\Big|_\infty, \|\tilde{\Theta} - \Theta^0\|_2 < 2\left(\sqrt{(\theta^0_{\max} + 1)(4 + \tau_0)} + C_3\right) \\ &\quad \times M_{\Gamma^0}\min\{d, \sqrt{p+s}\}\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{15}$$

with probability greater than $1 - O(n^{-\delta_0/8} + p^{-\tau_0/2})$.

Next, we discuss selection consistency for the sparse structure before providing a comparison of our results with the existing results in Section 3.3.

As discussed in Section 2.4, we propose to estimate S^0, the set of nonzero elements of Θ, by thresholding the inclusions probability p_{ij} that is defined at (9). The following theorem shows that $\hat{S}^0 = \{(i, j) : p_{ij} \ge T\}$, the set of edges with posterior probability greater than T, is a consistent estimator of S^0 for any $0 < T < 1$.

Theorem 4 (Selection consistency). Assume the same conditions in Theorem 2 and condition (A2) with the following restriction:

$$\epsilon_0 < \frac{1}{\log p}\log\left(\frac{v_1(1-\eta)}{v_0\eta}\right) < (C_4 - C_3)\big(K_0 - 2(C_1+C_3)M_{\Gamma^0}\big) \tag{16}$$

for some arbitrary small constant $\epsilon_0 > 0$. Then, for any T such that $0 < T < 1$, we have

$$\mathbb{P}(\hat{S^0} = S^0) \to 1.$$

A proof of Theorem 4 is provided in Appendix B.

In our model, sparsity is induced by an interplay between the parameters v_0, v_1, and η through $\log(v_1(1-\eta)/(v_0\eta))$. When $\log(v_1(1-\eta)/(v_0\eta))$ falls in the gap mentioned in Equation (16), the selection consistency can be achieved.

3.3. Comparison with Existing Results

We compare our results with those of GLasso (Ravikumar et al. 2011), CLIME (Cai, Liu, and Luo 2011) and the nonconvex regularization based method by Loh and Wainwright (2017).

In Ravikumar et al. (2011), the irrepresentable condition, $\left|\left|\left|\Gamma_{S^{0c}S^0}\Gamma^{-1}_{S^0S^0}\right|\right|\right|_\infty \le 1-\alpha$, is needed to establish the rate of convergence in entrywise ℓ_∞ norm. Such an assumption is quite restrictive, and is not needed for our results. In addition, under the polynomial tail condition, the rate of convergence established in Ravikumar et al. (2011) is $O_p(\sqrt{\frac{p^c}{n}})$, slower than our rate $O_p(\sqrt{\frac{\log p}{n}})$.

The theoretical results for CLIME (Cai, Liu, and Luo 2011) are similar to ours in terms of estimation accuracy. However, the main difference is the assumption on Θ^0. We assume boundedness of the largest eigenvalue of Θ^0, which is strictly weaker than the boundedness of $\left|\left|\left|\Theta^0\right|\right|\right|_\infty$ (the ℓ_∞/ℓ_∞ operator norm), the assumption imposed for CLIME. The weakness of our assumption follows from Hölder's inequality. To illustrate the strict difference between these assumptions, we consider the following precision matrix as an example:

$$\theta^0_{ii} = 1, \forall i; \quad \theta^0_{1,i} = \theta_{i,1} = \frac{1}{\sqrt{p}}, \text{ if } i \neq 1;$$
$$\theta^0_{ij} = 0 \qquad \text{if } i \neq j \text{ and } i \neq 1. \tag{17}$$

The precision matrix above has the so-called star structure, which is frequently observed in networks with a hub. In Figure 3, we plot the maximum eigenvalue and the maximum of the absolute row sum of this matrix with varying dimension p. We can see that it is easy to satisfy the upper bound on maximum eigenvalue, but not the upper bound on the ℓ_∞/ℓ_∞ operator norm, since the latter is diverging with p.

The major difference between our results and those from Loh and Wainwright (2017) is also in the weakness of the assumptions. The beta-min condition (minimal signal strength) is needed for the rate of estimation accuracy established in Loh and Wainwright (2017), while we do not require this assumption for estimation consistency. In addition, their results are only available for sub-Gaussian distributions, while we consider a much broader class of distributions, that is, distributions with exponential or polynomial tails.

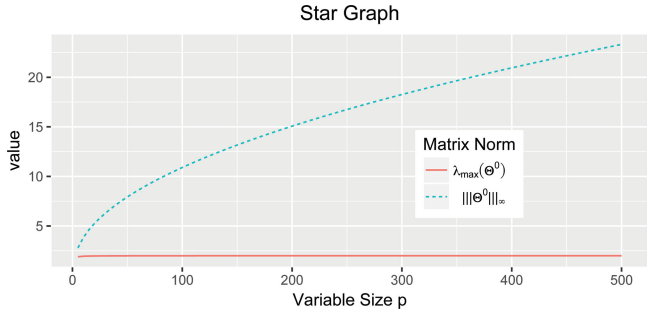

Figure 3. Plots of the maximum eigenvalue (solid line) and the ℓ_∞/ℓ_∞ operator norm (dashed line) for precision matrices with the star structure (17). Our model assumption corresponds to an upper bound on the solid line, while the one for CLIME corresponds to an upper bound on the dashed line.

4. Computation With EM Algorithm

We now describe how to compute the MAP estimate $\tilde{\Theta}$. Directly optimizing the negative log of the posterior distribution (4) is not easy. One numerical complication comes from the penalty term (5): it has a summation inside the logarithm due to the mixture prior distribution on θ_{ij}. The expectation-maximization (EM) algorithm is a popular tool in handling such a complication.

Recall the two-level hierarchical representation of the prior on θ_{ij} introduced in Section 2.4. Define R as the $p \times p$ matrix with binary entries r_{ij}. Then the full posterior distribution $\pi(\Theta, R|Y_1, \ldots, Y_n)$ is proportional to

$$f(Y_1, \ldots, Y_n|\Theta) \cdot \left[\prod_{i<j} \pi(\theta_{ij}|r_{ij})\pi(r_{ij}|\eta)\right] \cdot \left[\prod_i \pi(\theta_{ii}|\tau)\right] \mathbb{1}(\Theta \succ 0)\mathbb{1}(\|\Theta\|_2 \le B). \tag{18}$$

We treat R as latent and derive an EM algorithm to obtain the MAP estimate of Θ from the M-step and the posterior distribution of R from the E-step upon convergence. The E-step of our algorithm is inspired by the EM algorithm for linear regression from Ročková and George (2014) and the one for factor models from Ročková and George (2016a), and the M-step of our algorithm is inspired by the optimization procedure used by GLasso (Banerjee, El Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008; Mazumder and Hastie 2012).

4.1. The E-Step

At the E-step, we first compute the distribution of R given the parameter value from the previous iteration $\Theta^{(t)}$. Note that the binary indicator r_{ij} does not appear in the likelihood function, and only appears in (7) and (8) in the prior specification. It is easy to show that $r_{ij}|\Theta^{(t)}, Y_1, \ldots, Y_n$ follows $\mathsf{Bern}(p_{ij})$ with

$$\log\frac{p_{ij}}{1-p_{ij}} = \log\frac{v_0}{v_1} + \log\frac{\eta}{1-\eta} - \frac{|\theta^{(t)}_{ij}|}{v_1} + \frac{|\theta^{(t)}_{ij}|}{v_0}. \tag{19}$$

Next we evaluate the expectation of $\log \pi(\Theta, R|Y_1, \ldots, Y_n)$ with respect to $\pi(R|\Theta^{(t)}, Y_1, \ldots, Y_n)$, which gives rise to the so-called Q function:

$$Q(\Theta|\Theta^{(t)}) = \Big\{ \frac{n}{2}\log\det(\Theta) - \frac{n}{2}\mathrm{tr}(S\Theta) + \sum_i (\log\tau - \tau\theta_{ii}) + \sum_{i<j} p_{ij}\Big[-\log(2v_1) - \frac{|\theta_{ij}|}{v_1} + \log\eta \Big] + \sum_{i<j}(1-p_{ij})\Big[-\log(2v_0) - \frac{|\theta_{ij}|}{v_0} + \log(1-\eta)\Big]\Big\} \times \mathbb{1}(\Theta \succ 0)\mathbb{1}(\|\Theta\|_2 \le B). \quad (20)$$

4.2. The M-Step

At the M-step of the $(t+1)$th iteration, we sequentially update Θ in a column by column fashion to maximize $Q(\Theta|\Theta^{(t)})$. Without loss of generality, we describe the updating rule for the last column of Θ while fixing the others.

For convenience, partition the covariance matrix W and the precision matrix Θ as follows:

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix} \qquad \Theta = \begin{bmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{bmatrix},$$

where W_{11} is the $(p-1)\times(p-1)$ submatrix, w_{12} is the $(p-1)\times 1$ vector at the last column of W, and w_{22} is the diagonal entry at the bottom-right corner. The sample covariance matrix S, the binary indicator matrix $R = [r_{ij}]$, and the conditional probability matrix $P = [p_{ij}]$ where p_{ij} is defined in (19) are also partitioned similarly. We list the following equalities from $W\Theta = \mathbf{I}_p$ which will be used in our algorithm:

$$\begin{bmatrix} W_{11} & w_{12} \\ \cdot & w_{22} \end{bmatrix} = \begin{bmatrix} \Theta_{11}^{-1} + \frac{\Theta_{11}^{-1}\theta_{12}\theta_{12}^T\Theta_{11}^{-1}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} & -\frac{\Theta_{11}^{-1}\theta_{12}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \\ \cdot & \frac{1}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \end{bmatrix}. \quad (21)$$

Given Θ_{11}, to update the last column $(\theta_{12}, \theta_{22})$, we set the subgradient of Q with respect to $(\theta_{12}, \theta_{22})$ to zero. First, take the subgradient of Q with respect to θ_{22}:

$$\frac{\partial Q}{\partial \theta_{22}} = \frac{n}{2}\frac{1}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} - \frac{n}{2}(s_{22}+\tau) = 0. \quad (22)$$

Due to Equations (21) and (22), we have

$$w_{22} = \frac{1}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} = s_{22} + \frac{2}{n}\tau,$$

which leads to the following update for θ_{22}:

$$\theta_{22} \leftarrow \frac{1}{w_{22}} + \theta_{12}^T\Theta_{11}^{-1}\theta_{12}. \quad (23)$$

Next take the subgradient of Q with respect to θ_{12}:

$$\frac{\partial Q}{\partial \theta_{12}} = \frac{n}{2}\left(\frac{-2\Theta_{11}^{-1}\theta_{12}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} - 2s_{12}\right) - \left(\frac{1}{v_1}p_{12} + \frac{1}{v_0}(1-p_{12})\right) \odot \mathrm{sign}(\theta_{12}) = n\left(-\Theta_{11}^{-1}\theta_{12}w_{22} - s_{12}\right) - \left(\frac{1}{v_1}p_{12} + \frac{1}{v_0}(1-p_{12})\right) \odot \mathrm{sign}(\theta_{12}) = 0, \quad (24)$$

where $A \odot B$ denotes the element-wise multiplication of two matrices. Here the second line of (24) is due to the identities in (21). To update θ_{12}, we then solve the following stationary equation for θ_{12} with coordinate descent, under the constraint $\|\Theta\|_2 \le B$:

$$ns_{12} + nw_{22}\Theta_{11}^{-1}\theta_{12} + \left(\frac{1}{v_1}P_{12} + \frac{1}{v_0}(1-P_{12})\right) \odot \mathrm{sign}(\theta_{12}) = 0. \quad (25)$$

The coordinate descent algorithm for updating θ_{12} is summarized in Algorithm 1. Since only one column is changed, checking the bound $\|\Theta\|_2 \le B$ is computationally feasible (see Appendix C in the supplementary material for more details). In practice, we could also proxy the constraint on $\|\Theta\|_2$ with a constraint on the largest absolute value of the elements in Θ. In our empirical studies, this relaxation performs quite well.

Algorithm 1 Coordinate Descent for θ_{12}

Initialize θ_{12} from the previous iteration as the starting point.
repeat
  for j in $1:(p-1)$ **do**
    Solve the following equation for θ_{12j}:

$$ns_{12j} + nw_{22}\Theta_{11}^{-1}{}_{j,\setminus j}\theta_{12\setminus j} + nw_{22}\Theta_{11}^{-1}{}_{j,j}\theta_{12j} + \Big[\Big(\frac{1}{v_1}P_{12} + \frac{1}{v_0}(1-P_{12})\Big) \odot \mathrm{sign}(\theta_{12})\Big]_j = 0.$$

  end for
until Converge or Max Iterations Reached.
If $\|\Theta\|_2 > B$:**Return** θ_{12} from the previous iteration
Else: **Return** θ_{12}

When updating $(\theta_{12}, \theta_{22})$, we need Θ_{11}^{-1}. Instead of directly computing the inverse of Θ_{11}, we compute it from

$$\Theta_{11}^{-1} = W_{11} - w_{12}w_{21}/w_{22},$$

which is derived from (21). After the update of $(\theta_{12}, \theta_{22})$ is completed, we ensure that $W\Theta = \mathbf{I}_p$ holds by updating W_{11} and w_{12} via identities from (21). Therefore, we always keep a copy of the most updated covariance matrix W in our algorithm. Note we do not update w_{22} here, only because the relationship related to w_{22} within $W\Theta = \mathbf{I}_p$ is already ensured. That is, if w_{22} is updated using (21), it remains unchanged.

4.3. The Output

The entire algorithm, BAGUS, is summarized and displayed as Algorithm 2. After convergence, we extract the following output from our algorithm: the P matrix, the posterior probability on the sparse structure, from the E-step and the MAP estimator $\tilde{\Theta}$ from the M-step.

To obtain an estimate of the sparse structure in R, we threshold the entries of P, namely:

$$\hat{r}_{ij} = 1, \text{ if } P_{ij} \ge 0.5; \quad \hat{r}_{ij} = 0, \text{ otherwise.}$$

As shown in Theorem 4, thresholding entries of P with any number T such that $0 < T < 1$ could recover the true sparse structure with probability converging to 1.

Algorithm 2 BAGUS

Initialize $W = \Theta$=**I**

repeat

Update P with each entry p_{ij} updated as $\log \frac{p_{ij}}{1-p_{ij}} \leftarrow \left(\log \frac{v_0}{v_1} + \log \frac{\eta}{1-\eta} - \frac{|\theta_{ij}^{(t)}|}{v_1} + \frac{|\theta_{ij}^{(t)}|}{v_0}\right)$.

for j in $1 : p$ **do**

Move the j-th column and j-th row to the end (implicitly), namely $\Theta_{11} := \Theta_{\backslash j \backslash j}, \theta_{12} := \theta_{\backslash jj}, \theta_{22} := \theta_{jj}$

Update w_{22} using $w_{22} \leftarrow s_{22} + \frac{2}{n}\tau$

Update θ_{12} by solving (25) with Coordinate Descent for θ_{12}.

Update θ_{22} using $\theta_{22} \leftarrow \frac{1}{w_{22}} + \theta_{12}^T \Theta_{11}^{-1} \theta_{12}$.

Update W_{11}, w_{12} using (21)

end for

until Converge

Return Θ, P

For many existing algorithms, the positive definiteness of the estimate of Θ is not guaranteed. For example, GLasso (Friedman, Hastie, and Tibshirani 2008) can only ensure the positive definiteness of the estimate of the covariance matrix W, but not of the estimate of the precision matrix Θ, as shown in Mazumder and Hastie (2012). The following theorem shows that MAP estimate $\tilde{\Theta}$ returned by our algorithm is ensured to be symmetric and positive definite.

Theorem 5 (Symmetry and positive definite). The estimate of Θ returned by BAGUS is always symmetric, and it is also positive definite if the initial value $\Theta^{(0)}$ is positive definite.

A proof is given in the supplementary material.

4.4. Remarks

- *Computation cost.* In BAGUS, the computation cost is $O(p^2)$ for updating one column. There are p columns in Θ to update, so the overall computational complexity of our algorithm is $O(p^3)$, which matches the computation cost for GLasso.
- *Parameter tuning.* BAGUS involves the following hyperparameters: η, τ, v_0, and v_1. We always set $\eta = 0.5$ and $\tau = v_0$ so that there are only two parameters v_0 and v_1 to be tuned. Parameter tuning has an empirical Bayes flavor. In our simulations, we use the theoretical results to set the rough range of the hyperparameters, and then use a BIC-like criterion to tune the hyperparameters:

$$\text{BIC} = n(\text{tr}(S\hat{\Theta}) - \log\det(\hat{\Theta})) + \log(n) \times \#\{(i,j) : 1 \le i < j \le p, \hat{\theta}_{ij} \neq 0\}. \quad (26)$$

The same BIC criterion is used by Yuan and Lin (2007) while a similar BIC criterion with a regression-based working likelihood is used by Peng et al. (2009).

5. Empirical Results

In this section, we compare our method with the competitive alternatives in both simulated and real datasets and study the performance of our approach.

5.1. Twelve Simulation Settings

Following the simulation studies from related work (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Peng et al. 2009; Cai, Liu, and Luo 2011), we generate data Y from a multivariate Gaussian distribution with mean 0 and precision matrix $\Theta^0 = (\theta_{ij}^0)$.

We consider four different models, that is, four different forms of Θ^0. The first three have been considered in Yuan and Lin (2007) and the fourth one is similar to the set-up in Peng et al. (2009).

1. Model 1 (star model): $\theta_{ii}^0 = 1$, $\theta_{1i}^0 = \theta_{i1}^0 = \frac{1}{\sqrt{p}}$.
2. Model 2 ($AR(2)$ model): $\theta_{ii}^0 = 1$, $\theta_{i,i-1}^0 = \theta_{i-1,i}^0 = 0.5$, and $\theta_{i,i-2}^0 = \theta_{i-2,i}^0 = 0.25$.
3. Model 3 (circle model): $\theta_{ii}^0 = 2$, $\theta_{i,i-1}^0 = \theta_{i-1,i}^0 = 1$, and $\theta_{1p}^0 = \theta_{p1}^0$=0.9.
4. Model 4 (random graph): The true precision matrix Θ^0 is set as follows.
   (a) Set $\theta_{ii}^0 = 1$.
   (b) Randomly select $1.5 \times p$ of the off-diagonal entries θ_{ij}^0 $(i \neq j)$ and set their values to be uniform from $[0.4, 1] \cup [-1, -0.4]$; set the remaining off-diagonal entries to be zero.
   (c) Calculate the sum of absolute values of the off-diagonal entries for each column, and then divide each off-diagonal entry by 1.1 fold of the corresponding column sum. Average this rescaled matrix with its transpose to obtain a symmetric and positive definite matrix.
   (d) Multiple each entry by σ^2, which is set to be 3.

For each model, we consider three cases with different values for p:

$$(1)\ p = 50; \quad (2)\ p = 100; \quad (3)\ p = 200.$$

So, we consider a total of 12 simulation settings. In each setting, $n = 100$ observations are generated, and results are aggregated based on 50 replications.

For estimation accuracy of Θ^0, we use Frobenius norm (denoted as Fnorm). For selection accuracy, we consider three criteria: sensitivity, specificity, and MCC (Matthews correlation coefficient):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}, \qquad \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}, \quad \text{and}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN-FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

where TP (true positive), FP (false positive), TN (true negative), and FN (false negative) are based on detection of edges in the graph corresponding to the true precision matrix Θ^0. MCC returns a value between -1 and $+1$, and the higher the MCC, the better the structure recovery is. A coefficient of $+1$ in MCC represents a perfect structure recovery, and we note that recovering all the edges simultaneously is very challenging and none of the existing methods are able to ensure that. In addition, we note that it may not be meaningful to compare the results across graphs with different values of p because the level of sparsity changes with p which makes it difficult to assess the difficulty of

Table 1. Model 1 star.

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	2.301(0.126)	0.687(0.015)	0.998(0.004)	0.339(0.011)
CLIME	3.387(0.401)	0.452(0.051)	0.971(0.023)	0.168(0.021)
SPACE	2.978(0.244)	0.972(0.039)	1.000(0.003)	0.824(0.163)
BAGUS	**1.053(0.107)**	1.000(0.000)	1.000(0.000)	**1.000(0.000)**
$n = 100, p = 100$				
GLasso	4.219(0.118)	0.715(0.007)	0.989(0.008)	0.260(0.005)
CLIME	4.818(0.449)	0.998(0.004)	0.336(0.000)	0.131(0.067)
SPACE	3.207(0.311)	0.987(0.022)	0.996(0.024)	0.842(0.162)
BAGUS	**1.499(0.138)**	1.000(0.000)	1.000(0.000)	**1.000(0.000)**
$n = 100, p = 200$				
GLasso	3.028(0.068)	0.947(0.003)	0.999(0.002)	0.389(0.009)
CLIME	5.595(0.528)	0.978(0.018)	0.000(0.000)	−0.014(0.006)
SPACE	3.735(0.294)	0.985(0.007)	1.000(0.000)	0.656(0.138)
BAGUS	**2.006(0.100)**	1.000(0.000)	1.000(0.001)	**1.000(0.001)**

NOTE: The best results in terms of MCC and Fnorm are highlighted in boldface.

Table 3. Model 3: Circle.

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	**4.319(0.174)**	0.492(0.064)	1.000(0.000)	0.196(0.024)
CLIME	5.785(0.440)	0.555(0.026)	1.000(0.000)	0.221(0.010)
SPACE	19.402(0.232)	0.930(0.006)	1.000(0.000)	0.595(0.019)
BAGUS	**4.253(0.578)**	0.993(0.004)	0.964(0.029)	**0.903(0.049)**
$n = 100, p = 100$				
GLasso	6.981(0.192)	0.647(0.005)	1.000(0.000)	0.189(0.002)
CLIME	19.282(2.802)	0.224(0.226)	0.995(0.015)	0.069(0.058)
SPACE	27.737(0.345)	0.975(0.010)	0.994(0.008)	0.674(0.062)
BAGUS	**6.012(0.513)**	0.996(0.002)	0.957(0.032)	**0.895(0.055)**
$n = 100, p = 200$				
GLasso	**7.664(0.209)**	0.752(0.003)	1.000(0.000)	0.172(0.001)
CLIME	33.009(0.535)	0.857(0.154)	0.769(0.167)	0.209(0.052)
SPACE	32.142(0.832)	0.981(0.012)	0.783(0.212)	0.485(0.129)
BAGUS	10.378(1.001)	0.995(0.001)	0.886(0.033)	**0.752(0.028)**

NOTE: The best results in terms of MCC and Fnorm are highlighted in boldface.

the setting based on p alone. For instance, for most models considered in our simulation study, the level of sparsity increases along with p, because of which all the methods have their specificity increasing when p gets larger (see Tables 1–4). So we recommend against comparing the results as p changes and instead to compare the results across different methods within the same setting.

In the simulation study, we compare our method, denoted as BAGUS, with the following alternatives: GLasso from Friedman, Hastie, and Tibshirani (2008), SPACE from Peng et al. (2009), and CLIME from Cai, Liu, and Luo (2011). They are all shown to have estimation consistency under various conditions as discussed in Section 3.3. We also considered the regression-based method from Meinshausen and Bühlmann (2006), but the results are not presented here because tuning the parameters as suggested in Meinshausen and Bühlmann (2006) gave us "NA" for MCC in multiple scenarios considered here.

For each simulated dataset, tuning for our model uses the aforementioned BIC criterion with a parameter set of $\eta = 0.5$, $v_0 = \tau = (0.4, 2, 4, 20) \times \sqrt{\frac{1}{n \log p}}$, and v_1 ranges from $v_0 \times (1.5, 3, 5, 10)$. The tuning parameters for GLasso are chosen with 10-fold CV, the tuning parameters for SPACE are chosen from the BIC-like criterion proposed in Peng et al. (2009) and the tuning and estimation for CLIME estimator is done using the R package `flare` (Li et al. 2015) as suggested on the homepage (*http://www-stat.wharton.upenn.edu/tcai/article/html/Precision-Matrix.html*) of Cai, Liu, and Luo (2011). For cross-validation, the number of λ values is set to be 40. Results for all the simulated cases are summarized in Tables 1–4.

In almost all the settings considered, our method BAGUS performs the best in terms of both selection accuracy, that is, MCC, and estimation accuracy, that is, Fnorm. We believe that

Table 2. Model 2: AR(2).

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	**3.361(0.240)**	0.479(0.056)	0.981(0.015)	0.251(0.028)
CLIME	3.758(0.381)	0.822(0.054)	0.906(0.039)	0.472(0.053)
SPACE	5.903(0.070)	0.982(0.004)	0.608(0.038)	0.656(0.029)
BAGUS	3.671(0.291)	0.997(0.002)	0.551(0.032)	**0.707(0.025)**
$n = 100, p = 100$				
GLasso	8.130(0.035)	0.901(0.007)	0.745(0.028)	0.382(0.017)
CLIME	5.595(1.578)	0.837(0.075)	0.821(0.191)	0.371(0.085)
SPACE	9.819(0.083)	0.991(0.002)	0.566(0.025)	0.625(0.021)
BAGUS	**5.330(0.369)**	0.998(0.001)	0.549(0.018)	**0.707(0.022)**
$n = 100, p = 200$				
GLasso	11.728(0.045)	0.990(0.001)	0.478(0.017)	0.481(0.014)
CLIME	11.552(0.382)	0.989(0.004)	0.580(0.031)	0.539(0.028)
SPACE	13.696(0.079)	0.995(0.000)	0.518(0.018)	0.588(0.013)
BAGUS	**8.214(0.548)**	0.998(0.001)	0.543(0.015)	**0.677(0.027)**

NOTE: The best results in terms of MCC and Fnorm are highlighted in boldface.

Table 4. Model 4: Random graph.

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	7.017(0.256)	0.877(0.010)	0.766(0.039)	0.417(0.027)
CLIME	11.347(0.452)	0.971(0.012)	0.614(0.068)	0.572(0.042)
SPACE	12.278(0.183)	1.000(0.000)	0.073(0.031)	0.257(0.051)
BAGUS	**5.811(0.357)**	0.999(0.001)	0.443(0.032)	**0.637(0.027)**
$n = 100, p = 100$				
GLasso	11.851(0.900)	0.837(0.047)	0.720(0.049)	0.285(0.033)
CLIME	12.649(1.587)	0.735(0.153)	0.761(0.120)	0.243(0.123)
SPACE	17.706(0.203)	1.000(0.000)	0.068(0.015)	0.236(0.028)
BAGUS	**8.754(0.366)**	0.999(0.001)	0.400(0.022)	**0.598(0.022)**
$n = 100, p = 200$				
GLasso	15.054(0.356)	0.951(0.012)	0.633(0.029)	0.307(0.017)
CLIME	23.568(0.954)	0.993(0.004)	0.469(0.048)	0.492(0.038)
SPACE	24.997(0.213)	0.999(0.000)	0.090(0.014)	0.221(0.024)
BAGUS	**13.096(0.522)**	0.999(0.000)	0.382(0.050)	**0.565(0.032)**

NOTE: The best results in terms of MCC and Fnorm are highlighted in boldface.

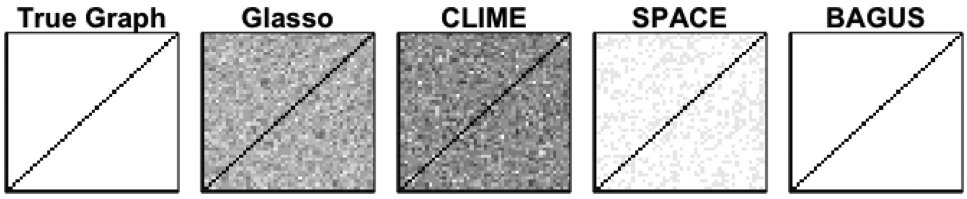

Figure 4. Average of the estimated precision matrices for the model with the star structure.

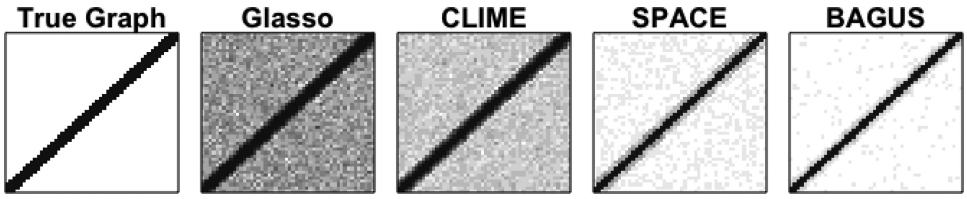

Figure 5. Average of the estimated precision matrices for the model with the AR(2) structure.

it is due to the adaptive nature of the Bayesian penalization and the weaker conditions under which the consistency results hold true for BAGUS. Other than BAGUS, SPACE usually performs well in terms of sparse selection and GLasso performs well in terms of estimation accuracy. However, SPACE has a large estimation error in most cases and GLasso tends to have smaller MCC. In our simulation study, CLIME estimator did not perform very well. It is particularly worth noting that for the star graph, where the assumption for CLIME fails (see discussion in Section 3.3), the performance of CLIME is particularly worse.

In Figure 4, we plot the receiver operating characteristic (ROC) curves for all the methods considered under different models by varying hyper (tuning) parameters for the case with $p = 50$. This is to see the performance of different methods by removing the effect of tuning. Our method BAGUS remains at the top in all the settings considered in terms of area under the ROC curve (AUC). This plot suggests that except for the star graph, performance of CLIME is not as poor as indicated by the selected graph, which suggests that the performance of CLIME could be improved by better tuning. However, for the star graph, CLIME is still observed to be particularly worse even in view of the ROC curve.

We also recorded the average of the estimated structures from the 50 replicates and compare it with the truth to get a visual understanding of the performance of different methods, shown in Figures 5–8. It is noticeable that GLasso and CLIME provide noisier estimates than BAGUS by including many zero entries in the selection; BAGUS and SPACE are sparser and appear closer to the true precision matrix. However, SPACE usually produces noisier estimates than BAGUS (for Models 1–3) and misses a lot of true signals for Model 4. In summary, BAGUS provides a highly competitive performance across the models considered.

5.2. *Real Application: Telephone Call Center Data*

We now apply our method to the analysis of data from a telephone call center in a major U.S. northeastern financial organization. The data consist of the arrival time of each phone call in 2002 every day from 7 am till midnight, except for 6 days when the data collecting machine is out of order. More details about this data can be found in Shen and Huang (2005).

Following the preprocessing as suggested by Huang et al. (2006) and Fan, Feng, and Wu (2009) for this dataset, we divide each day into 102 10-min intervals and count the number of call arrivals for each interval, denoted as N_{it} where $t = 1$: 102 and

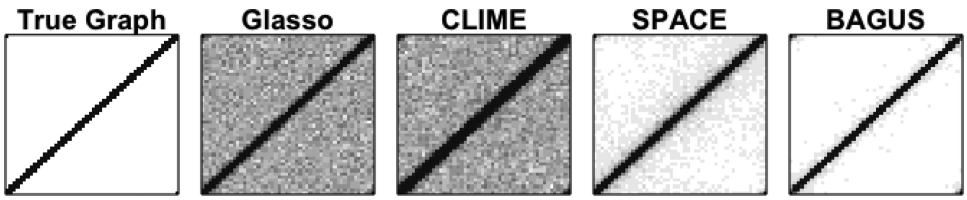

Figure 6. Average of the estimated precision matrices for the model with the circle structure.

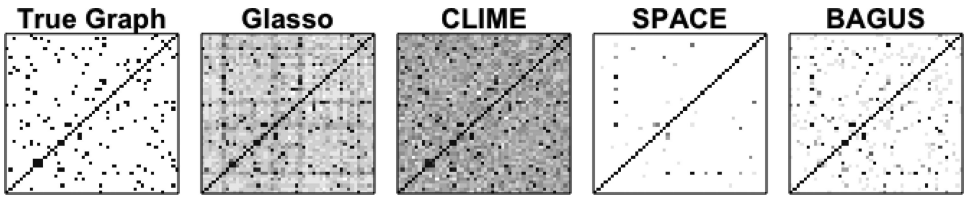

Figure 7. Average of the estimated precision matrices for the model with the random structure.

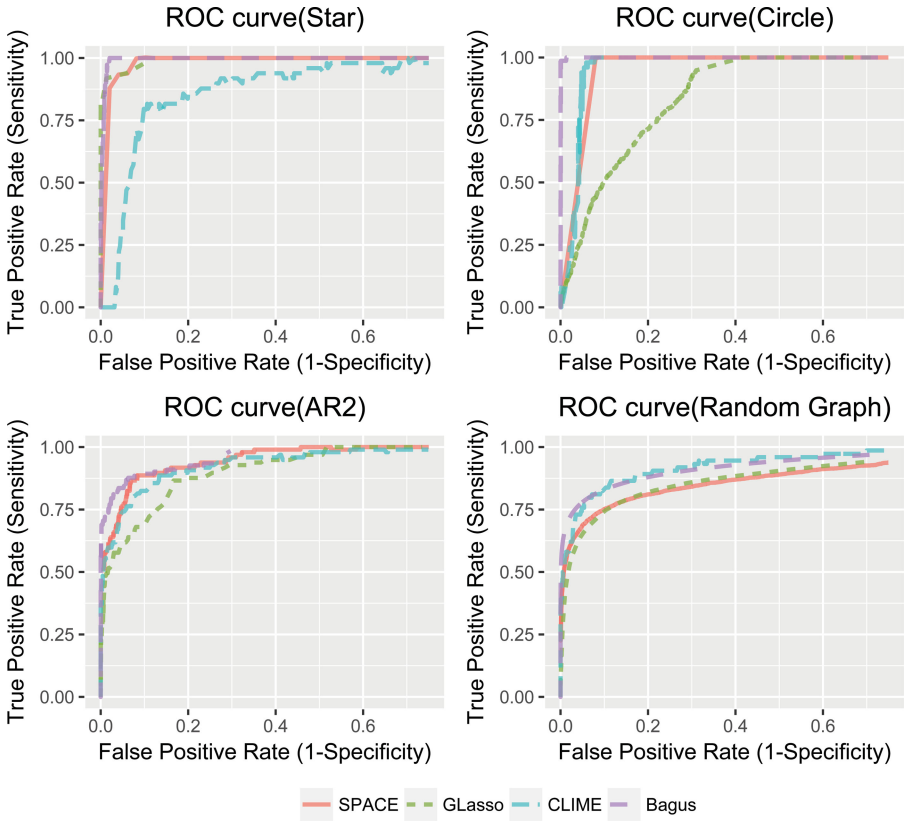

Figure 8. ROC curves for different methods and different data-generating models with $p = 50$.

$i = 1$: 239. Only 239 days of data are considered here, after we remove holidays and days when the data collecting machine was faulty. Represent the observations on the ith day as $Y_i = (Y_{i1}, Y_{i2}, \ldots)^T$, a 102×1 vector with $Y_{it} = \sqrt{N_{it} + \frac{1}{4}}$, a variance stabilizing transformation of the number of calls. Let μ and Θ denote the mean vector and precision matrix of the 102-dimensional vector Y.

We apply all the methods considered on the first 205 days of data to estimate Θ, as well as μ, and use the remaining 34 days of data to evaluate the performance. The performance evaluation is carried out as follows. First, divide the 102 observations for each day into two parts (Z_{i1} and Z_{i2}), where Z_{i1} is a 51×1 vector containing data from the first 51 intervals on the ith day and Z_{i2} is also a 51×1 vector containing the remaining 51 observations, then partition the mean vector μ and the precision matrix Θ accordingly. Under the multivariate Gaussian assumption, the best mean squared error forecast of Z_{i2} given Z_{i1} is given by

$$\mathbb{E}(Z_{i2}|Z_{i1}) = u_2 - \Theta_{22}^{-1}\Theta_{21}(Z_{i1} - u_1), \tag{27}$$

which is also the best linear unbiased predictor for non-Gaussian data. So plugging the estimates of μ and Θ based on the first 205 days into (27), we evaluate the prediction accuracy for Z_{i2} for the remaining 34 days. We adopt the same criterion used by Fan, Feng, and Wu (2009), the average absolute forecast error (AAFE), to measure the prediction performance:

$$\text{AAFE}_t = \frac{1}{34}\sum_{i=206}^{239} |\hat{Y}_{it} - Y_{it}|, \tag{28}$$

where $\hat{Y}_{it}$ and Y_{it} denote the predicted and observed values, respectively.

We compare the prediction performance based on estimates from our method BAGUS, the inverse of the sample covariance matrix (denoted as "Sample"), GLasso and CLIME. The prediction errors for these methods at all 51 time points are shown in Figure 9. Their average AAFE values are displayed in Table 5, along with the average AAFE values for Adaptive Lasso and SCAD taken from Fan, Feng, and Wu (2009).

From the results, we see that BAGUS and CLIME have a significantly improved performance in prediction accuracy when compared with the other methods. To look further into the estimates provided by these methods, we present the sparsity structures estimated from GLasso, CLIME, and BAGUS in Figure 10. In this figure, yellow points (appear in light tone when converted to grayscale) indicate signals and blue points (dark tone in grayscale) indicate noise. In the Gaussian graphical model context, a yellow point suggests that the call arrivals in the corresponding two time intervals are conditionally dependent. It is interesting to find that a strong autoregressive type of dependence structure is present in estimators from all methods. However, the methods differ in terms of the degree of autoregression suggested by their corresponding estimates. The estimated structure from BAGUS is the most sparse one and suggests a small degree of autoregression compared to those of GLasso and

Table 5. Average prediction error for different methods.

	Sample	GLasso	Adaptive Lasso	SCAD	CLIME	BAGUS
Average AAFE	1.46	1.38	1.34	1.31	1.14	**1.00**

NOTE: The best results in terms of prediction error are highlighted in boldface.

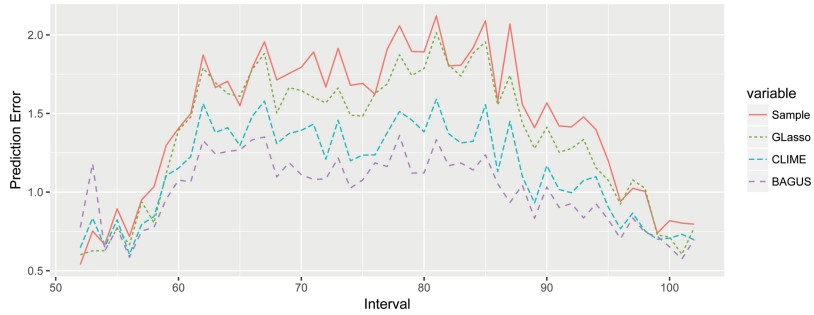

Figure 9. Prediction error for the call center data: AAFE_t on Y axis and t on X axis.

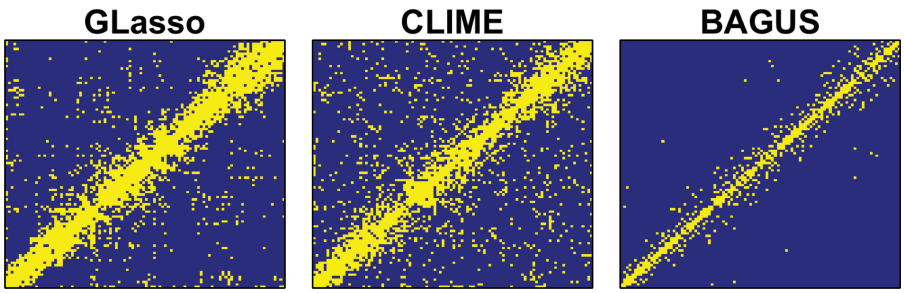

Figure 10. Sparsity structures estimated for different methods for the call center data.

CLIME. That is, BAGUS indicates that the telephone call arrivals majorly depend only on recent history, while others indicate dependence over a long history. Based on the prediction accuracies of different methods, the sparser dependence structure suggested by BAGUS seems sufficient to provide good prediction although it is difficult to know which structure, in reality, is closer to the underlying precision matrix. In terms of practical utility, this provides support in favor of storing and managing less amount of historical data that could potentially reduce cost of data management.

6. Conclusion

In high-dimensional data analysis, there is a large literature on penalization from a frequentist viewpoint majorly focusing on Lasso-based convex penalties and some nonconvex penalties such as SCAD. On the other hand, in the Bayesian framework, a variety of shrinkage and sparsity inducing prior distributions have been proposed. In the context of graphical models, our work demonstrates that spike-and-slab priors with Laplace distributions provide adaptive penalization that leads to better theoretical and empirical performance compared to state-of-the-art methods. Since some recent articles (Ročková and George 2016a; Deshpande, Ročková, and George 2017) have also found spike-and-slab Lasso priors to be useful in other high-dimensional contexts, we believe that our strategy of Bayesian regularization will be advantageous in a broad range of high-dimensional problems and that its success demonstrated in our work will motivate further interest in this direction.

Supplementary Materials

Supplementary material contains technical proofs for all the theorems from the main article.

Funding

The research is partially supported by the NSF Award DMS - 1811768.

References

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. [1218,1224]

Banerjee, S., and Ghosal, S. (2015), "Bayesian Structure Learning in Graphical Models," *Journal of Multivariate Analysis*, 136, 147–162. [1219]

Bühlmann, P., and Van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York: Springer Science & Business Media. [1218]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [1219,1222,1224,1226,1227]

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [1218]

Carvalho, C. M., and Scott, J. G. (2009), "Objective Bayesian Model Selection in Gaussian Graphical Models," *Biometrika*, 96, 497–512. [1219]

Dempster, A. P. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [1218]

Deshpande, S. K., Ročková, V., and George, E. I. (2017), "Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab Lasso," arXiv:1708.08911. [1230]

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011), "Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data," *Journal of the American Statistical Association*, 106, 1418–1433. [1219]

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197. [1218]

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration via the Adaptive LASSO and SCAD Penalties," *The Annals of Applied Statistics*, 3, 521–541. [1218,1228,1229]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1221]

Fan, J., Liao, Y., and Mincheva, M. (2011), "High Dimensional Covariance Matrix Estimation in Approximate Factor Models," *Annals of Statistics*, 39, 3320–3356. [1218]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1218,1224,1226,1227]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1219,1220]

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [1221]

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Matrix Selection and Estimation via Penalized Normal Likelihood," *Biometrika*, 93, 85–98. [1228]

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33, 730–773. [1219,1220]

Johnstone, I. M. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *Annals of Statistics*, 29, 295–327. [1218]

Khare, K., Oh, S.-Y., and Rajaratnam, B. (2015), "A Convex Pseudolikelihood Framework for High Dimensional Partial Correlation Estimation With Convergence Guarantees," *Journal of the Royal Statistical Society*, Series B, 77, 803–825. [1218]

Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *Annals of Statistics*, 37, 4254–4278. [1219,1223]

Li, X., Zhao, T., Wang, L., Yuan, X., and Liu, H. (2015), "The flare Package for High Dimensional Linear Regression and Precision Matrix Estimation in R," *The Journal of Machine Learning Research*, 16, 553–557. [1227]

Loh, P.-L., and Wainwright, M. J. (2017), "Support Recovery Without Incoherence: A Case for Nonconvex Regularization," *The Annals of Statistics*, 45, 2455–2482. [1219,1221,1223,1224]

——— (2015), "Regularized M-Estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *Journal of Machine Learning Research*, 16, 559–616. [1219]

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*, ACM, pp. 705–712. [1219]

Mazumder, R., and Hastie, T. (2012), "The Graphical Lasso: New Insights and Alternatives," *Electronic Journal of Statistics*, 6, 2125–2149. [1219,1224,1226]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1218,1227]

Mohammadi, A., and Wit, E. C. (2015), "Bayesian Structure Learning in Sparse Gaussian Graphical Models," *Bayesian Analysis*, 10, 109–138. [1219]

Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection With Shrinking and Diffusing Priors," *Annals of Statistics*, 42, 789–817. [1219,1220]

Paul, D. (2007), "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model," *Statistica Sinica*, 17, 1617–1642. [1219]

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. [1218,1227]

Pourahmadi, M. (2013), *High-Dimensional Covariance Estimation: With High-Dimensional Data* (*Wiley Series in Probability and Statistics*), New York: Wiley. [1218]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-Dimensional Covariance Estimation by Minimizing ℓ_1-Penalized Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [1219,1222,1223,1224]

Ročková, V. (2018), "Bayesian Estimation of Sparse Signals With a Continuous Spike-and-Slab Prior," *Annals of Statistics*, 46, 401–437. [1219,1220,1222]

Ročková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [1219,1224]

Ročková, V., and George, E. I. (2016a), "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity," *Journal of the American Statistical Association*, 111, 1608–1622. [1218,1219,1220,1224,1230]

——— (2016b), "The Spike-and-Slab Lasso," *Journal of the American Statistical Association*, accepted. [1219,1220,1221,1222]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [1219,1223]

Shen, H., and Huang, J. Z. (2005), "Analysis of Call centre Arrival Data Using Singular Value Decomposition," *Applied Stochastic Models in Business and Industry*, 21, 251–263. [1228]

Wang, H. (2012), "Bayesian Graphical Lasso Models and Efficient Posterior Computation," *Bayesian Analysis*, 7, 867–886. [1219]

Wang, H., and Li, S. (2012), "Efficient Gaussian Graphical Model Determination Under G-Wishart Prior Distributions," *Electronic Journal of Statistics*, 6, 168–198. [1219]

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., and Thiele, L. (2004), "Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis Thaliana," *Genome Biology*, 5, R92. [1218]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1218]

Zhou, S., van de Geer, S., and Bühlmann, P. (2009), "Adaptive Lasso for High Dimensional Regression and Gaussian Graphical Modeling," arXiv:0903.2515. [1218]