Journal of Machine Learning Research 16 (2015) 1035-1062 Submitted 10/13; Revised 9/14; Published

Joint Estimation of Multiple Precision Matrices with Common Structures

Wonyul Lee WONYULL@EMAIL.UNC.EDU
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260, USA

Yufeng Liu YFLIU@EMAIL.UNC.EDU
Department of Statistics and Operations Research
Department of Genetics
Department of Biostatistics
Carolina Center for Genome Sciences
University of North Carolina
Chapel Hill, NC 27599-3260, USA

Editor: Francis Bach

Abstract

Estimation of inverse covariance matrices, known as precision matrices, is important in various areas of statistical analysis. In this article, we consider estimation of multiple precision matrices sharing some common structures. In this setting, estimating each precision matrix separately can be suboptimal as it ignores potential common structures. This article proposes a new approach to parameterize each precision matrix as a sum of common and unique components and estimate multiple precision matrices in a constrained l_1 minimization framework. We establish both estimation and selection consistency of the proposed estimator in the high dimensional setting. The proposed estimator achieves a faster convergence rate for the common structure in certain cases. Our numerical examples demonstrate that our new estimator can perform better than several existing methods in terms of the entropy loss and Frobenius loss. An application to a glioblastoma cancer data set reveals some interesting gene networks across multiple cancer subtypes.

Keywords: covariance matrix, graphical model, high dimension, joint estimation, precision matrix

1. Introduction

Estimation of a precision matrix, which is an inverse covariance matrix, has attracted a lot of attention recently. One reason is that the precision matrix plays an important role in various areas of statistical analysis. For example, some classification techniques such as linear discriminant analysis and quadratic discriminant analysis require good estimates of precision matrices. In addition, estimation of a precision matrix is essential to establish conditional dependence relationships in the context of Gaussian graphical models. Another reason is that the high-dimensional nature of many modern statistical applications makes the problem of estimating a precision matrix very challenging. In situations where the dimension p is comparable to or much larger than the sample size n, more feasible and stable techniques are required for accurate estimation of a precision matrix.

To tackle such problems, various penalized maximum likelihood methods have been considered by many researchers in recent years (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008; Rothman et al., 2008; Lam and Fan, 2009; Fan et al., 2009, and many more). These approaches produce a sparse estimator of the precision matrix by maximizing the penalized Gaussian likelihood with sparse penalties such as the l_1 penalty and the smoothly clipped absolute deviation penalty (Fan and Li, 2001). Ravikumar et al. (2011) studied the theoretical properties of the l_1 penalized likelihood estimator for a broad class of population distributions.

Instead of using likelihood approaches, several techniques take advantage of the connection between linear regression and the entries of the precision matrix. See for example Meinshausen and Bühlmann (2006); Peng et al. (2009); Yuan (2010). In particular, these approaches convert the estimation problem of the precision matrix into relevant regression problems and solve them with sparse regression techniques accordingly. One advantage of these approaches is that they can handle a wide range of distributions including the Gaussian case. Cai et al. (2011) recently proposed a very interesting method to directly estimate the precision matrix without the Gaussian distributional assumption. This approach solves a constrained l_1 minimization problem to obtain a sparse estimator of the precision matrix. They showed that the proposed estimator has a faster convergence rate than the l_1 penalized likelihood estimator for some non-Gaussian cases.

All aforementioned approaches focus on estimation of a single precision matrix. The fundamental assumption of these approaches is that all observations follow the same distribution. However, in some real applications, this assumption can be unreasonable. As a motivating example, consider the glioblastoma multiforme (GBM) cancer data set studied by The Cancer Genome Atlas Research Network (The Cancer Genome Atlas Research Network, 2008). It is shown in the literature that the GBM cancer can be classified into four subtypes (Verhaak et al., 2010). In this case, it would be more realistic to assume that the distribution of gene expression levels can vary from one subtype to another, which results in multiple precision matrices to estimate (Lee et al., 2012). A naive way to estimate them is to model each subtype separately. However, in this separate approach, modeling of one subtype completely ignores the information on other subtypes. This can be suboptimal if there exists some common structure across different subtypes.

To improve the estimation in presence of some common structure, several joint estimation methods have been proposed recently in a penalized likelihood framework. See for example Guo et al. (2011); Honorio and Samaras (2012); Danaher et al. (2014). These methods employ various group penalties in the Gaussian likelihood framework to link the estimation of separate precision matrices.

In this article, we propose a new method to jointly estimate multiple precision matrices. Our approach uses a novel representation of each precision matrix as a sum of common and unique matrices. Then we apply sparse constrained optimization on the common and unique components. The proposed method is applicable for a broad class of distributions including both the Gaussian and some non-Gaussian cases. The main strength of our method is that it uses all available information to jointly estimate the common and unique structures, which can be more preferable than separate modelings. The estimation can be improved if the precision matrices are similar to each other. Furthermore, our method is able to discover unique structures of each precision matrix, which enables us to identify differences among multiple precision matrices. The proposed estimator is shown to achieve a faster convergence rate for the common structures in certain cases.

The rest of this article is organized as follows. In Section 2, we introduce our proposed method after reviewing some existing separate approaches. We establish its theoretical properties in Section 3. Section 4 develops computational algorithms to obtain a solution for the proposed method. Simulated examples are presented in Section 5 to demonstrate performance of our estimator and analysis of a glioblastoma cancer data example is provided in Section 6. The proofs of theorems are provided in Appendix.

2. Methodology

In this section, we introduce a new method for estimating multiple precision matrices in an l_1 minimization framework. Consider a heterogeneous data set with G different groups. For the gth group ($g = 1, \ldots, G$), let $\{x_1^{(\mathrm{g})}, \ldots, x_{n_g}^{(\mathrm{g})}\}$ be an independent and identically distributed random sample of size n_g, where $x_k^{(\mathrm{g})} = (x_{ki}^{(\mathrm{g})}, \ldots, x_{kp}^{(\mathrm{g})})^{\mathrm{T}}$ is a p-dimensional random vector with the covariance matrix $\Sigma_0^{(\mathrm{g})}$ and precision matrix $\Omega_0^{(\mathrm{g})} := (\Sigma_0^{(\mathrm{g})})^{-1}$. For detailed illustration of our proposed method, we first define some notations similar to Cai et al. (2011). For a matrix $X = (x_{ij}) \in \mathcal{R}^{p\times q}$, we define the elementwise l_1 norm $||X||_1 = \sum_{i=1}^{p}\sum_{j=1}^{q}|x_{ij}|$, the elementwise l_∞ norm $|X|_\infty = \max_{1\le i\le p, 1\le j\le q}|x_{ij}|$ and the matrix l_1 norm $||X||_{L_1} = \max_{1\le j\le q}\sum_{i=1}^{p}|x_{ij}|$. For a vector $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathcal{R}^p$, $|x|_1$ and $|x|_\infty$ denote vector l_1 and l_∞ norms respectively. The notation $X \succ 0$ indicates that X is positive definite. Let I be a $p \times p$ identity matrix. For the gth group, $\hat{\Sigma}^{(\mathrm{g})}$ denotes the sample covariance matrix. Write $\Omega_0^{(\mathrm{g})} = (\omega_{ij,0}^{(\mathrm{g})}); g = 1, \ldots, G$.

Our aim is to estimate the precision matrices, $\Omega_0^{(1)}, \ldots, \Omega_0^{(\mathrm{G})}$. The most naive way to achieve this goal is to estimate each precision matrix separately by taking the inverses of the sample covariance matrices. However, in high dimensional cases, the sample covariance matrices are not only unstable for estimating the covariance matrices, but also not invertible. To estimate the precision matrix in high dimensions, various estimators have been introduced in the literature. For example, various l_1 penalized Gaussian likelihood estimators have been studied intensively in the literature (see for example, Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008; Rothman et al., 2008). In this framework, the precision matrices can be estimated by solving the following G optimization problems:

$$\min_{\Omega^{(\mathrm{g})}\succ 0} \mathrm{tr}(\hat{\Sigma}^{(\mathrm{g})}\Omega^{(\mathrm{g})}) - \log\{\det(\Omega^{(\mathrm{g})})\} + \lambda_g \sum_{i\neq j}|w_{ij}^{(\mathrm{g})}|,\ g = 1, \ldots, G, \tag{1}$$

where λ_g is a tuning parameter which controls the degree of the sparsity in the estimated precision matrices. Other sparse penalized Gaussian likelihood estimators have been proposed as well (Lam and Fan, 2009; Fan et al., 2009).

Recently, Cai et al. (2011) proposed an interesting method of constrained l_1 minimization for inverse matrix estimation (CLIME), which can be directly implemented using linear programming. In particular, the CLIME estimator of $\Omega_0^{(\mathrm{g})}$ is the solution of the following optimization problem:

$$\min ||\Omega^{(\mathrm{g})}||_1 \text{ subject to: } |\hat{\Sigma}^{(\mathrm{g})}\Omega^{(\mathrm{g})} - I|_\infty \leq \lambda_g, \tag{2}$$

where $\hat{\Sigma}^{(\mathrm{g})}$ is the sample covariance matrix and λ_g is a tuning parameter. As the optimization problem in (2) does not require symmetry of the solution, the final CLIME estimator is obtained by symmetrizing the solution of (2). The CLIME estimator does not need the Gaussian distributional assumption. Cai et al. (2011) showed that the convergence rate of the CLIME estimator is faster than that of the l_1 penalized Gaussian likelihood estimator if the underlying true distribution has polynomial-type tails.

To estimate multiple precision matrices, $\Omega_0^{(1)}, \ldots, \Omega_0^{(\mathrm{G})}$, we can build G individual models using the optimization problem (1) or (2). However, these separate approaches can be suboptimal when the precision matrices share some common structure. Several recent papers have proposed joint estimations of multiple precision matrices under the Gaussian distributional assumption to improve estimation. In particular, such an estimator is the solution of

$$\min_{\{\Omega\}} \sum_{g=1}^{G} n_g \left[\mathrm{tr}(\hat{\Sigma}^{(\mathrm{g})}\Omega^{(\mathrm{g})}) - \log\{\det(\Omega^{(\mathrm{g})})\} \right] + P(\{\Omega\}),$$

where n_g is the sample size of the g-th group, $\{\Omega\} = \{\Omega^{(1)}, \ldots, \Omega^{(\mathrm{G})}\}$, and $P(\{\Omega\})$ is a penalty function that encourages similarity across the G estimated precision matrices. For example, Guo et al. (2011) employs a non-convex penalty called *hierarchical group penalty* which has the form, $P(\{\Omega\}) = \lambda \sum_{i \neq j} \left(\sum_{g=1}^{G} |\omega_{ij}^{(\mathrm{g})}| \right)^{1/2}$. Honorio and Samaras (2012) adopts a convex penalty, $P(\{\Omega\}) = \lambda \sum_{i \neq j} |(\omega_{ij}^{(1)}, \ldots, \omega_{ij}^{(\mathrm{G})})|_p$ $(p > 1)$ where $|\cdot|_p$ is the vector l_p norm. To separately control the sparsity level and the extent of similarity, Danaher et al. (2014) considered a *fused lasso penalty*, $P(\{\Omega\}) = \lambda_1 \sum_{g=1}^{G} \sum_{i \neq j} |\omega_{ij}^{(\mathrm{g})}| + \lambda_2 \sum_{g<g'} \sum_{ij} |\omega_{ij}^{(\mathrm{g})} - \omega_{ij}^{(\mathrm{g}')}|$. In some simulation settings, they showed that the joint estimation can perform better than separate l_1 penalized normal likelihood estimation. As pointed by Ravikumar et al. (2011), these penalized Gaussian likelihood estimators are applicable even for some mild non-Gaussian data since maximizing a penalized likelihood can be interpreted as minimizing a penalized log-determinant Bregman divergence. However, these approaches were mainly designed for Gaussian data and can be less efficient when the underlying distribution becomes far from Gaussian. In this paper, we propose a new joint method for estimating multiple precision matrices, which is less dependent on the distributional assumption and applicable for both Gaussian and non-Gaussian cases.

In our joint estimation method, we take the multi-task learning perspective and first define the common structure M_0 and the unique structure $R_0^{(\mathrm{g})}$ as

$$M_0 := \frac{1}{G} \sum_{g=1}^{G} \Omega_0^{(\mathrm{g})}, R_0^{(\mathrm{g})} := \Omega_0^{(\mathrm{g})} - M_0; g = 1, \ldots, G.$$

It follows from the definition that $\sum_{g=1}^{G} R_0^{(\mathrm{g})} = 0$, and consequently our representation is identifiable. The idea of decomposing parameters into common and individual structures was previously considered in the context of supervised multi-tasking learning (Evgeniou and Pontil, 2004). Their aim was to improve prediction performance of supervised multi-tasking learning. Here we focus on better estimation of precision matrices with the common and individual structures. The unique structure is defined to capture different strength of the edges across all classes. In a special case that an element of M_0 is zero, then the corresponding nonzero element in $R_0^{(\mathrm{g})}$ can be interpreted as a unique edge. Thus, the unique structure can address differences in magnitude as well as unique edges. If all precision matrices are very similar, then the unique structures defined above would be close to zero. In this case, it can be natural and advantageous to encourage sparsity among $\{R_0^{(1)}, \dots, R_0^{(\mathrm{G})}\}$ in the estimation. To estimate the precision matrices consistently in high dimensions, it is also necessary to assume some special structure of M_0 as well. In our work, we also assume that M_0 is sparse. To estimate $\{M_0, R_0^{(1)}, \dots, R_0^{(\mathrm{G})}\}$, we propose the following constrained l_1 minimization criterion:

$$\min\{||M||_1 + \nu \sum_{g=1}^{G} ||R^{(\mathrm{g})}||_1\}$$

$$\text{s.t } |\frac{1}{G}\sum_{g=1}^{G}\{\hat{\Sigma}^{(\mathrm{g})}(M + R^{(\mathrm{g})}) - I\}|_\infty \le \lambda_1, |\hat{\Sigma}^{(\mathrm{g})}(M + R^{(\mathrm{g})}) - I|_\infty \le \lambda_2, \sum_{g=1}^{G} R^{(\mathrm{g})} = 0, \qquad (3)$$

where λ_1 and λ_2 are tuning parameters and ν is a prespecified weight. Note that if $\lambda_1 > \lambda_2$, then the second inequality constraints in (3) imply the first inequality constraint. Therefore, we only consider a pair of (λ_1, λ_2) satisfying $\lambda_1 \le \lambda_2$. The first inequality constraint in (3) reflects how close the final estimators are to the inverses of the sample covariance matrices in an average sense. On the other hand, the second inequality constraint controls an individual level of closeness between the estimators and the sample covariance matrices.

For illustration, consider an extreme case where all the precision matrices are the same. In this case, the unique structures may be negligible and the first inequality constraint in (3) approximately reduces to $|(G^{-1}\sum_{g=1}^{G}\hat{\Sigma}^{(\mathrm{g})})M - I|_\infty \le \lambda_1$. Therefore, we can pool all the sample covariance matrices to estimate the common structure which is the precision matrix in this case. This would be advantageous than building each model separately. The value of ν in (3) reflects how complex the unique structures of the resulting estimators are. If the resulting estimators are expected to be very similar from each other, then a large value of ν is preferred. In Section 3, ν is set to be G^{-1} or $G^{-1/2}$ for our theoretical results.

Similar to Cai et al. (2011), the solutions in (3) are not symmetric in general. Therefore, the final estimators are obtained after a symmetrization step. Let $\{\hat{M}, \hat{R}^{(1)}, \dots, \hat{R}^{(\mathrm{G})}\}$ be the solution of (3). Then we define $\hat{\Omega}_1^{(\mathrm{g})} := \hat{M} + \hat{R}^{(\mathrm{g})}; g = 1, \dots, G$. The final estimator of $\{\Omega_0^{(1)}, \dots, \Omega_0^{(\mathrm{G})}\}$ is obtained by symmetrizing $\{\hat{\Omega}_1^{(1)}, \dots, \hat{\Omega}_1^{(\mathrm{G})}\}$ as follows. Let $\hat{\Omega}_1^{(\mathrm{g})} = (\hat{\omega}_{ij,1}^{(\mathrm{g})})$. Our joint estimator of multiple precision matrices (JEMP), $\{\hat{\Omega}^{(1)}, \dots, \hat{\Omega}^{(\mathrm{G})}\}$, is defined as symmetric matrices, $\{\hat{\Omega}^{(\mathrm{g})} = (\hat{\omega}_{ij}^{(\mathrm{g})}); g = 1, \dots, G\}$ with

$$\hat{\omega}_{ij}^{(\mathrm{g})} = \hat{\omega}_{ij,1}^{(\mathrm{g})} I\{\sum_{g=1}^{G}|\hat{\omega}_{ij,1}^{(\mathrm{g})}| \le \sum_{g=1}^{G}|\hat{\omega}_{ji,1}^{(\mathrm{g})}|\} + \hat{\omega}_{ji,1}^{(\mathrm{g})} I\{\sum_{g=1}^{G}|\hat{\omega}_{ij,1}^{(\mathrm{g})}| > \sum_{g=1}^{G}|\hat{\omega}_{ji,1}^{(\mathrm{g})}|\}; g = 1, \dots, G.$$

Note that the solution $\hat{\Omega}^{(\mathrm{g})}$ is not necessarily positive definite. Although there is no guarantee for the solution to be positive definite, it can be positive definite with high probability.

In our simulation study, we observed that within a reasonable range of tuning parameters, almost all solutions are positive definite. Furthermore, one can perform projection of the estimator to the space of positive definite matrices to ensure positive definitiveness as discussed in Yuan (2010).

As a remark, although we focus on generalizing CLIME for multiple graph estimation in this paper, our proposed common and unique structure approach can also be applied to the graphical lasso estimator under the Gaussian assumption as pointed out by one reviewer. As a future research direction, it would be interesting to investigate how the common and unique structure framework works in the graphical lasso estimator.

3. Theoretical Properties

In this section, we investigate theoretical properties of our proposed joint estimator JEMP. In particular, we first construct the convergence rate of our estimator in the high dimensional setting. Then we show that the convergence rate can be improved for the common structure of the precision matrices in certain cases. Finally, the model selection consistency is shown with an additional thresholding step.

For theoretical properties, we follow the set-up of Cai et al. (2011) and the results therein are also used for our technical derivations. In this section, for simplicity, we assume that $n = n_1 = \cdots = n_G$. We consider the following class of matrices,

$$\mathcal{U} := \{\Omega : \Omega \succ 0, \|\Omega\|_{L_1} \leq C_M\},$$

and assume that $\Omega_0^{(\mathrm{g})} \in \mathcal{U}$ for all $g = 1, \ldots, G$. This assumption requires that the true precision matrices are sparse in terms of the l_1 norm while allowing them to have many small entries. Write $E(x^{(\mathrm{g})}) = (\mu_1^{(\mathrm{g})}, \ldots, \mu_p^{(\mathrm{g})})^{\mathrm{T}}$. We also make the following moment condition on $x^{(\mathrm{g})}$ for our theoretical results.

Condition 1 *There exists some $0 < \eta < 1/4$ such that $E[\exp\{t(x_i^{(\mathrm{g})} - \mu_i^{(\mathrm{g})})^2\}] \leq K < \infty$ for all $|t| \leq \eta$ and all i, g and $G \log p/n \leq \eta$, where K is a bounded constant.*

Condition 1 indicates that the components of $x^{(\mathrm{g})}$ are uniformly sub-Gaussian. This condition is satisfied if $x^{(\mathrm{g})}$ follows a multivariate Gaussian distribution or has uniformly bounded components.

Theorem 1 *Assume Condition 1 holds. Let $\lambda_1 = \lambda_2 = 3C_M C_0(\log p/n)^{1/2}$, where $C_0 = 2\eta^{-2}(2 + \tau + \eta^{-1}e^2K^2)^2$ and $\tau > 0$. Set $\nu = G^{-1}$. Then*

$$\max_{ij}\left(\frac{1}{G}\sum_{g=1}^{G}|\hat{\omega}_{ij}^{(\mathrm{g})} - \omega_{ij,0}^{(\mathrm{g})}|\right) \leq 6C_M^2 C_0\left(\frac{\log p}{n}\right)^{1/2},$$

with probability greater than $1 - 4Gp^{-\tau}$.

In an average sense, the convergence rate can be viewed the same as that of the CLIME estimator which is of order $(\log p/n)^{1/2}$. In this theorem, the first inequality constraint in (3) does not play any role in the estimation procedure as we set $\lambda_1 = \lambda_2$. In the next theorem, with properly chosen λ_1, we construct a faster convergence rate for the common part under certain conditions.

Theorem 2 *Assume Condition 1 holds. Suppose that there exists* $C_R > 0$ *such that* $\|R_0^{(\mathrm{g})}\|_{L_1} \leq C_R$ *for all* $g = 1, \ldots, G$ *and* $(\sum_{g=1}^{G} \|R_0^{(\mathrm{g})}\|_{L_1}) \leq C_R G^{1/2}$*. Set* $\nu = G^{-1/2}$ *and let* $\lambda_1 = (C_M + C_R)C_0\{\log p/(nG)\}^{1/2}$ *and* $\lambda_2 = C_M C_0 (\log p/n)^{1/2}$*. Then*

$$|\hat{M} - M_0|_\infty \leq C_0(2C_M^2 + 4C_M C_R + C_R^2)\left(\frac{\log p}{nG}\right)^{1/2},$$

with probability greater than $1 - 2(1 + 3G)p^{-\tau}$*.*

Theorem 2 states that our proposed method can estimate the common part more efficiently with the corresponding convergence rate of order $\{\log p/(nG)\}^{1/2}$, which is faster than the order $(\log p/n)^{1/2}$.

Note that our theorems show consistency of our estimator in terms of the elementwise l_∞ norm. On the other hand, Guo et al. (2011) showed consistency of their estimator under the Frobenious norm. Therefore, our theoretical results are not directly comparable to the theorems in Guo et al. (2011). However, it is worthwhile to note that our Theorem 2 reveals the effect of G on the consistency while the theorems in Guo et al. (2011) do not show explicitly how their estimator can have advantage over separate estimation in terms of consistency.

Besides its estimation consistency, we also prove the model selection consistency of our estimator which means that it reveals the exact set of nonzero components in the true precision matrices with high probability. For this result, a thresholding step is introduced. In particular, a threshold estimator $\tilde{\Omega}^{(\mathrm{g})} = (\tilde{\omega}_{ij}^{(\mathrm{g})})$ based on $\{\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(\mathrm{G})}\}$ is defined as,

$$\tilde{\omega}_{ij}^{(\mathrm{g})} = \hat{\omega}_{ij}^{(\mathrm{g})} I\{|\hat{\omega}_{ij}^{(\mathrm{g})}| \geq \delta_n\},$$

where $\delta_n \geq 2C_M G\lambda_2$ and λ_2 is given in Theorem 1. To state the model selection consistency precisely, we define

$$\mathcal{S}_0 := \{(i,j,g) : \omega_{ij,0}^{(\mathrm{g})} \neq 0\}, \hat{\mathcal{S}} := \{(i,j,g) : \tilde{\omega}_{ij}^{(\mathrm{g})} \neq 0\} \text{ and } \theta_{\min} := \min_{(i,j,g)\in\mathcal{S}_0} \sum_{g=1}^{G} |\omega_{ij,0}^{(\mathrm{g})}|.$$

Then the next theorem states the model selection consistency of our estimator.

Theorem 3 *Assume Condition 1 holds. If* $\theta_{\min} > 2\delta_n$*, then*

$$pr(\mathcal{S}_0 = \hat{\mathcal{S}}) \geq 1 - 4Gp^{-\tau}.$$

4. Numerical Algorithm

In this section, we describe how to obtain the numerical solutions of the optimization problem (3). In Section 4.1, the optimization problem (3) is decomposed into p individual subproblems and a linear programming approach is used to solve them. In Section 4.2, we describe another algorithm using the alternating directions method of multiplier (ADMM). Section 4.3 explains how the tuning parameters can be selected.

4.1 Decomposition of (3)

Similar to the Lemma 1 in Cai et al. (2011), one can show that the optimization problem (3) can be decomposed into p individual minimization problems. In particular, let e_i be the ith column of I. For $1 \leq i \leq p$, let $\{\hat{m}_i, \hat{r}_i^{(1)}, \ldots, \hat{r}_i^{(\mathrm{G})}\}$ be the solution of the following optimization problem:

$$\min\{|m|_1 + \nu \sum_{g=1}^{G} |r^{(\mathrm{g})}|_1\}$$

$$\text{s.t. } |\frac{1}{G}\sum_{g=1}^{G}\{\hat{\Sigma}^{(\mathrm{g})}(m + r^{(\mathrm{g})}) - e_i\}|_\infty \leq \lambda_1, |\hat{\Sigma}^{(\mathrm{g})}(m + r^{(\mathrm{g})}) - e_i|_\infty \leq \lambda_2, \sum_{g=1}^{G} r^{(\mathrm{g})} = 0, \tag{4}$$

where m, $r^{(1)}, \ldots, r^{(\mathrm{G})}$ are vectors in $\mathcal{R}^p$. We can show that solving the optimization problem (3) is equivalent to solving the p optimization problems in (4). The optimization problem in (4) can be further reformulated as a linear programming problem and the simplex method is used to solve this problem (Boyd and Vandenberghe, 2004). For our simulation study and the GBM data analysis, we obtain the solution of (3) using the efficient R-package *fastclime*, which provides a generic fast linear programming solver (Pang et al., 2014).

4.2 An ADMM Algorithm

In this section, we describe an alternating directions method of multipliers (ADMM) algorithm to solve (4) which can be potentially more scalable than the previously explained linear programming approach. We refer the reader to Boyd et al. (2010) for detailed explanation of ADMM algorithms and their convergence properties.

To reformulate (4) into an appropriate ADMM form, define $y = (m^{\mathrm{T}}, \nu r^{(1)\mathrm{T}}, \ldots, \nu r^{(\mathrm{G})\mathrm{T}})^{\mathrm{T}}$, $z_m = \sum_{g=1}^{G}\{\hat{\Sigma}^{(\mathrm{g})}(m + r^{(\mathrm{g})}) - e_i\}/G$, $z_g = \hat{\Sigma}^{(\mathrm{g})}(m + r^{(\mathrm{g})}) - e_i$, and $z = ({z_1}^{\mathrm{T}}, \ldots, {z_G}^{\mathrm{T}}, {z_m}^{\mathrm{T}})^{\mathrm{T}}$. Denote the $a \times a$ identity matrix as $I_{a\times a}$ and the $a \times b$ zero matrix as $O_{a\times b}$. Then the problem (4) can be rewritten as

$$\min |y|_1 \text{ s.t. } |z_m|_\infty \leq \lambda_1, |z_g|_\infty \leq \lambda_2, Ay - Bz = C, \text{ where} \tag{5}$$

$$A = \begin{pmatrix} \hat{\Sigma}^{(1)} & \nu^{-1}\hat{\Sigma}^{(1)} & O_{p\times p} & \cdots & O_{p\times p} \\ \hat{\Sigma}^{(2)} & O_{p\times p} & \nu^{-1}\hat{\Sigma}^{(2)} & \cdots & O_{p\times p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}^{(\mathrm{G})} & O_{p\times p} & O_{p\times p} & \cdots & \nu^{-1}\hat{\Sigma}^{(\mathrm{G})} \\ G^{-1}\sum_{g=1}^{G}\hat{\Sigma}^{(\mathrm{g})} & (\nu G)^{-1}\hat{\Sigma}^{(1)} & (\nu G)^{-1}\hat{\Sigma}^{(2)} & \cdots & (\nu G)^{-1}\hat{\Sigma}^{(\mathrm{G})} \\ O_{p\times p} & I_{p\times p} & I_{p\times p} & \cdots & I_{p\times p} \end{pmatrix},$$

$B = \begin{pmatrix} I_{(1+G)p\times(1+G)p} \\ O_{p\times(1+G)p} \end{pmatrix}$, and $C = ({e_i}^{\mathrm{T}}, \ldots, {e_i}^{\mathrm{T}}, {O_{p\times 1}})^{\mathrm{T}}$. The scaled augmented Lagrangian for (5) is given by

$$L(y, z, u) = |y|_1 + \frac{\rho}{2}||Ay - Bz - C + u||_2^2, \text{ s.t. } |z_m|_\infty \leq \lambda_1, |z_g|_\infty \leq \lambda_2,$$

where u is a $(2+G)p$-dimensional vector of dual variables. With the current solution z^k, u^k, the ADMM algorithm updates solutions sequentially as follows:

(a) $y^{k+1} = \text{argmin}_y L(y, z^k, u^k)$.

(b) $z^{k+1} = \text{argmin}_z L(y^{k+1}, z, u^k)$, s.t. $|z_m|_\infty \leq \lambda_1, |z_g|_\infty \leq \lambda_2$.

(c) $u^{k+1} = u^k + Ay^{k+1} - Bz^{k+1} - c$.

As $\text{argmin}_y L(y, z^k, u^k) = \text{argmin}_y \{|y|_1 + \frac{\rho}{2}||Ay - Bz^k - C + u^k||_2^2\}$, the step (a) can be viewed as an L_1 penalized least squares problem. Therefore, the step (a) can be solved using some existing algorithms for L_1 penalized least squares problems. In addition, one can show that the step (b) has a closed form of solution, $z^{k+1} = \min\{\max\{A'y^{k+1} - C' + (u^k)', -\lambda\}, \lambda\}$ where A' is the submatrix of A consisting of the first $(1+G)p$ rows, C' and $(u^k)'$ are the corresponding subvectors of C and u^k, and λ is a $(1+G)p$-dimensional vector of which the first Gp elements are λ_2 and the rest are λ_1. Note that scalability and computational speed of this ADMM algorithm largely depend on the algorithm used for the step (a) as the other steps have the explicit form of solutions.

4.3 Tuning Parameter Selection

To apply our method, we need to choose the tuning parameters, λ_1 and λ_2. In practice, we construct several models with many pairs of λ_1 and λ_2 satisfying $\lambda_1 \leq \lambda_2$ and evaluate them to determine the optimal pair. To evaluate each estimator, we measure the likelihood loss (LL) used in Cai et al. (2011) and its definition is

$$\text{LL} = \sum_{g=1}^{G} \text{tr}(\hat{\Sigma}_v^{(\text{g})} \hat{\Omega}^{(\text{g})}) - \log\{\det(\hat{\Omega}^{(\text{g})})\},$$

where $\hat{\Sigma}_v^{(\text{g})}$ is the sample covariance matrix of the gth group computed from an independent validation set. As mentioned in Section 2, the likelihood loss can be applicable for both Gaussian and some non-Gaussian data as it corresponds to the log-determinant Bregman divergence between the estimators and empirical precision matrices in the validation set. Among several pairs of tuning values, we select the pair which minimizes LL. If a validation set is not available, a K-fold cross-validation can be combined to this criterion. In particular, we first randomly split the data set into K parts of equal sizes. Denote the data in the kth part by $\{X_{(k)}^{(1)}, \ldots, X_{(k)}^{(\text{G})}\}$ which is used as a validation set for the kth estimator. For each k, with a given value of (λ_1, λ_2), we obtain estimators using all observations which do not belong to $\{X_{(k)}^{(1)}, \ldots, X_{(k)}^{(\text{G})}\}$ and denote them as $\{\hat{\Omega}_{(k)}^{(\text{G})}, \ldots, \hat{\Omega}_{(k)}^{(\text{G})}\}$. Then the likelihood loss (LL) is defined as

$$\text{LL} = \sum_{k=1}^{K} \sum_{g=1}^{G} \text{tr}(\hat{\Sigma}_{(k)}^{(\text{g})} \hat{\Omega}_{(k)}^{(\text{g})}) - \log\{\det(\hat{\Omega}_{(k)}^{(\text{g})})\},$$

where $\hat{\Sigma}_{(k)}^{(\text{g})}$ is the sample covariance matrix of the gth group using $X_{(k)}^{(\text{g})}$. Once the optimal pair is selected which minimizes LL, the final model is constructed using all data points with the selected pair.

5. Simulated Examples

In this section, we carry out simulation studies to assess the numerical performance of our proposed method. In particular, we compare the numerical performance of five methods: two separate methods and three joint methods. In separate approaches, each precision matrix is estimated separately via the CLIME estimator or the GLASSO estimator. For joint approaches, all precision matrices are estimated together using our JEMP estimator, the fused graphical lasso (FGL) estimator by Danaher et al. (2014), or the estimator by Guo et al. (2011), which we refer to as JOINT estimator hereafter. In our proposed method, ν is set to be $G^{-1/2}$. We also tried different values of ν such as G^{-1}, and the results are similar thus omitted. We consider three models as described below: the first two from Guo et al. (2011) and the last from Rothman et al. (2008); Cai et al. (2011). In all models, we set $p = 100$, $G = 3$ and $\Omega_0^{(\mathrm{g})} = \Omega_c + U^{(\mathrm{g})}$, where Ω_c is common in all groups and $U^{(\mathrm{g})}$ represents unique structure to the gth group. The common part, Ω_c, is generated as follows:

Model 1. Ω_c is a tridiagonal precision matrix. In particular, $\Sigma_c := \Omega_c^{-1} = (\sigma_{ij})$ is first constructed, where $\sigma_{ij} = \exp(-|d_i - d_j|/2)$, $d_1 < \ldots < d_p$, and $d_i - d_{i-1} \sim \mathrm{Unif}(0.5, 1), i = 2, \ldots, p$. Then let $\Omega_c = \Sigma_c^{-1}$.

Model 2. Ω_c is a 3 nearest-neighbor network. In particular, p points are randomly picked on a unit square and all pairwise distances among the points are calculated. Then we find 3 nearest neighbors for each point and a pair of symmetric entries in Ω_c corresponding to a pair of neighbors has a value randomly chosen from the interval $[-1, -0.5] \cup [0.5, 1]$.

Model 3. $\Omega_c = \Gamma + \delta I$, where each off-diagonal entry in Γ is generated independently from $0.5y$, with y following the Bernoulli distribution with success probability 0.02. Here, δ is selected so that the condition number of Ω_c is equal to p.

For each $U^{(\mathrm{g})}$, we randomly pick a pair of symmetric off-diagonal entries and replace them with values randomly chosen from the interval $[-1, -0.5] \cup [0.5, 1]$. We repeat this procedure until $\sum_{i<j} I(|u_{ij}^{(\mathrm{g})}| > 0) / \sum_{i<j} I(|\omega_{ij,c}| > 0) = \rho$, where $\Omega_c = (\omega_{ij,c})$ and $U^{(\mathrm{g})} = u_{ij}^{(\mathrm{g})}$. Therefore, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries. We consider four values of $\rho = 0, 0.25, 1$ and 4. To make the resulting precision matrices positive-definite, each diagonal element of each matrix $\Omega_0^{(\mathrm{g})}$ is replaced with 1.5 times the sum of the absolute values of the corresponding row. Finally, each matrix $\Omega_0^{(\mathrm{g})}$ is standardized to have unit diagonals. Note that in the case of $\rho = 1$ or 4, the true precision matrices are quite different from each other. From these cases, we can assess how joint methods work when the precision matrices are not similar. In addition, we also consider Model 4 below to assess how JEMP works when the precision matrices have different structures from each other.

Model 4. $\Omega_0^{(1)}$ is the tridiagonal precision matrix as in Model 1, $\Omega_0^{(2)}$ is the 3 nearest-neighbor network in Model 2, and $\Omega_0^{(3)}$ is the random network in Model 3.

For each group in each model, we generate a training sample of size $n = 100$ from either a multivariate normal distribution $N(0, \Sigma_0^{(\mathrm{g})})$ or a multivariate t-distribution with the covariance matrix $\Sigma_0^{(\mathrm{g})}$ and degrees of freedom of 3 or 5. In order to select optimal tuning parameters, an independent validation set of size $n = 100$ is also generated from the same distribution of the training sample. For each estimator, optimal tuning parameters are selected as described in Section 4. We replicate simulations 50 times for each model.

		$\rho = 0$		$\rho = 0.25$	
		EL	FL	EL	FL
Normal	CLIME	4.42 (0.02)	8.57 (0.03)	4.35 (0.02)	8.42 (0.03)
	GLASSO	3.70 (0.02)	6.90 (0.03)	3.60 (0.02)	6.73 (0.03)
	JOINT	3.43 (0.02)	6.64 (0.04)	3.41 (0.02)	6.61 (0.03)
	FGL	1.99 (0.02)	3.75 (0.03)	2.09 (0.02)	3.92 (0.03)
	JEMP	2.08 (0.02)	4.06 (0.04)	2.20 (0.02)	4.31 (0.04)
t (DF=5)	CLIME	5.75 (0.17)	10.63 (0.26)	5.81 (0.19)	10.75 (0.33)
	GLASSO	5.60 (0.09)	10.23 (0.16)	5.45 (0.09)	10.00 (0.16)
	JOINT	5.08 (0.11)	9.44 (0.15)	5.01 (0.12)	9.28 (0.19)
	FGL	3.47 (0.07)	6.12 (0.11)	3.46 (0.08)	6.12 (0.11)
	JEMP	3.21 (0.06)	6.14 (0.11)	3.41 (0.10)	6.52 (0.19)
t (DF=3)	CLIME	10.34 (0.83)	18.08 (1.05)	10.15 (0.91)	17.25 (1.06)
	GLASSO	11.87 (0.33)	24.10 (0.95)	11.78 (0.33)	24.21 (0.95)
	JOINT	8.84 (0.58)	15.16 (0.85)	8.95 (0.66)	15.17 (0.92)
	FGL	7.01 (0.24)	12.39 (0.52)	7.40 (0.31)	13.23 (0.66)
	JEMP	6.02 (0.33)	11.56 (0.73)	5.95 (0.30)	11.16 (0.62)

		$\rho = 1$		$\rho = 4$	
		EL	FL	EL	FL
Normal	CLIME	4.23 (0.02)	8.15 (0.03)	3.67 (0.01)	6.95 (0.03)
	GLASSO	3.37 (0.02)	6.33 (0.03)	2.57 (0.01)	4.96 (0.03)
	JOINT	3.27 (0.01)	6.40 (0.03)	2.51 (0.01)	4.95 (0.02)
	FGL	2.18 (0.01)	4.07 (0.02)	1.82 (0.01)	3.47 (0.02)
	JEMP	2.38 (0.01)	4.77 (0.04)	2.11 (0.01)	4.28 (0.02)
t (DF=5)	CLIME	5.53 (0.16)	10.12 (0.23)	4.83 (0.17)	8.72 (0.25)
	GLASSO	5.11 (0.09)	9.54 (0.17)	4.28 (0.09)	8.35 (0.19)
	JOINT	4.71 (0.10)	8.71 (0.14)	3.87 (0.12)	7.03 (0.16)
	FGL	3.31 (0.07)	5.95 (0.11)	2.54 (0.06)	4.68 (0.10)
	JEMP	3.32 (0.07)	6.40 (0.13)	2.78 (0.07)	5.35 (0.12)
t (DF=3)	CLIME	9.89 (0.86)	17.82 (1.16)	8.93 (0.91)	16.58 (1.28)
	GLASSO	11.32 (0.32)	23.77 (0.99)	10.42 (0.31)	23.70 (1.05)
	JOINT	9.27 (1.68)	14.23 (1.26)	7.14 (0.65)	11.90 (0.72)
	FGL	6.51 (0.25)	11.73 (0.56)	5.95 (0.27)	11.55 (0.67)
	JEMP	5.71 (0.29)	10.99 (0.73)	4.72 (0.24)	9.04 (0.49)

Table 1: Comparison summaries using Entropy loss (EL) and Frobenius loss (FL) over 50 replications for Model 1.

		$\rho = 0$		$\rho = 0.25$	
		EL	FL	EL	FL
Normal	CLIME	5.10 (0.02)	9.80 (0.04)	5.05 (0.02)	9.68 (0.04)
	GLASSO	4.50 (0.02)	8.07 (0.03)	4.44 (0.02)	7.98 (0.03)
	JOINT	3.89 (0.02)	7.42 (0.04)	4.13 (0.02)	7.84 (0.04)
	FGL	2.26 (0.02)	4.26 (0.03)	2.70 (0.02)	5.02 (0.03)
	JEMP	2.31 (0.02)	4.44 (0.03)	2.80 (0.02)	5.36 (0.03)
t (DF=5)	CLIME	6.60 (0.17)	12.03 (0.25)	6.62 (0.19)	12.09 (0.32)
	GLASSO	6.78 (0.09)	11.67 (0.15)	6.56 (0.09)	11.37 (0.14)
	JOINT	6.16 (0.10)	11.18 (0.16)	6.12 (0.14)	11.14 (0.23)
	FGL	4.03 (0.07)	6.88 (0.11)	4.28 (0.07)	7.30 (0.10)
	JEMP	3.74 (0.06)	6.98 (0.11)	4.15 (0.09)	7.72 (0.20)
t (DF=3)	CLIME	11.41 (0.87)	19.55 (1.06)	11.16 (0.93)	18.66 (1.09)
	GLASSO	13.16 (0.34)	24.31 (0.88)	12.90 (0.34)	24.29 (0.88)
	JOINT	10.14 (0.56)	16.96 (0.80)	10.24 (0.68)	17.03 (0.94)
	FGL	8.34 (0.28)	13.78 (0.55)	8.55 (0.31)	14.16 (0.59)
	JEMP	7.17 (0.36)	13.31 (0.84)	7.08 (0.31)	12.76 (0.61)

		$\rho = 1$		$\rho = 4$	
		EL	FL	EL	FL
Normal	CLIME	4.84 (0.02)	9.27 (0.04)	3.77 (0.01)	7.14 (0.03)
	GLASSO	4.07 (0.02)	7.42 (0.03)	2.68 (0.01)	5.09 (0.02)
	JOINT	3.99 (0.01)	7.72 (0.03)	2.63 (0.01)	5.16 (0.02)
	FGL	2.99 (0.01)	5.51 (0.02)	1.98 (0.01)	3.74 (0.01)
	JEMP	3.20 (0.01)	6.34 (0.04)	2.35 (0.01)	4.74 (0.02)
t (DF=5)	CLIME	6.14 (0.16)	11.22 (0.24)	4.95 (0.17)	8.96 (0.25)
	GLASSO	5.85 (0.09)	10.52 (0.16)	4.44 (0.09)	8.56 (0.18)
	JOINT	5.44 (0.10)	10.05 (0.15)	4.02 (0.12)	7.32 (0.16)
	FGL	4.07 (0.07)	7.17 (0.10)	2.68 (0.06)	4.91 (0.10)
	JEMP	4.11 (0.06)	7.87 (0.13)	3.00 (0.07)	5.77 (0.13)
t (DF=3)	CLIME	10.53 (0.88)	18.53 (1.15)	9.10 (0.92)	16.84 (1.29)
	GLASSO	12.11 (0.32)	23.89 (0.93)	10.59 (0.32)	23.77 (1.04)
	JOINT	10.00 (1.67)	15.26 (1.26)	7.27 (0.64)	12.10 (0.72)
	FGL	7.23 (0.25)	12.34 (0.52)	6.02 (0.26)	11.50 (0.64)
	JEMP	6.59 (0.31)	12.19 (0.70)	4.99 (0.26)	9.48 (0.53)

Table 2: Comparison summaries using Entropy loss (EL) and Frobenius loss (FL) over 50 replications for Model 2.

		$\rho = 0$		$\rho = 0.25$	
		EL	FL	EL	FL
Normal	CLIME	3.62 (0.02)	6.87 (0.03)	3.92 (0.02)	7.51 (0.04)
	GLASSO	2.60 (0.01)	5.03 (0.03)	3.03 (0.01)	5.78 (0.03)
	JOINT	2.53 (0.01)	4.97 (0.02)	2.99 (0.01)	5.89 (0.03)
	FGL	1.54 (0.01)	2.95 (0.02)	2.21 (0.01)	4.16 (0.02)
	JEMP	1.80 (0.01)	3.61 (0.03)	2.48 (0.01)	4.96 (0.03)
t (DF=5)	CLIME	4.77 (0.17)	8.68 (0.26)	5.23 (0.19)	9.63 (0.33)
	GLASSO	4.32 (0.09)	8.42 (0.20)	4.82 (0.09)	9.11 (0.18)
	JOINT	3.84 (0.12)	7.02 (0.16)	4.43 (0.15)	8.10 (0.21)
	FGL	2.54 (0.06)	4.68 (0.10)	3.11 (0.07)	5.62 (0.10)
	JEMP	2.60 (0.06)	4.99 (0.11)	3.35 (0.10)	6.44 (0.18)
t (DF=3)	CLIME	9.08 (0.84)	16.05 (1.07)	9.40 (0.92)	15.92 (1.06)
	GLASSO	10.64 (0.33)	24.09 (1.06)	11.14 (0.33)	24.26 (1.01)
	JOINT	7.54 (0.57)	13.03 (0.87)	8.35 (0.66)	14.09 (0.89)
	FGL	5.87 (0.26)	11.39 (0.65)	6.72 (0.30)	12.53 (0.70)
	JEMP	5.05 (0.37)	10.10 (0.93)	5.49 (0.30)	10.44 (0.66)

		$\rho = 1$		$\rho = 4$	
		EL	FL	EL	FL
Normal	CLIME	4.33 (0.02)	8.33 (0.03)	4.03 (0.02)	7.68 (0.03)
	GLASSO	3.52 (0.02)	6.54 (0.03)	3.00 (0.01)	5.67 (0.03)
	JOINT	3.50 (0.01)	6.86 (0.02)	2.94 (0.01)	5.78 (0.02)
	FGL	2.90 (0.01)	5.37 (0.02)	2.28 (0.01)	4.28 (0.01)
	JEMP	3.17 (0.01)	6.40 (0.02)	2.66 (0.01)	5.40 (0.02)
t (DF=5)	CLIME	5.64 (0.16)	10.31 (0.23)	5.20 (0.17)	9.42 (0.26)
	GLASSO	5.31 (0.09)	9.81 (0.17)	4.71 (0.09)	8.93 (0.18)
	JOINT	4.91 (0.11)	9.09 (0.14)	4.29 (0.12)	7.86 (0.17)
	FGL	3.66 (0.06)	6.53 (0.10)	2.98 (0.07)	5.40 (0.10)
	JEMP	3.93 (0.07)	7.56 (0.12)	3.27 (0.07)	6.32 (0.14)
t (DF=3)	CLIME	10.00 (0.87)	17.87 (1.16)	9.36 (0.88)	17.25 (1.26)
	GLASSO	11.60 (0.32)	23.89 (0.97)	10.89 (0.31)	23.79 (0.99)
	JOINT	9.52 (1.68)	14.60 (1.27)	7.57 (0.63)	12.59 (0.71)
	FGL	6.71 (0.24)	11.84 (0.52)	6.36 (0.26)	11.87 (0.61)
	JEMP	5.90 (0.26)	11.02 (0.59)	5.20 (0.26)	9.70 (0.51)

Table 3: Comparison summaries using Entropy loss (EL) and Frobenius loss (FL) over 50 replications for Model 3.

	Normal		t (DF=5)		t (DF=3)	
	EL	FL	EL	FL	EL	FL
CLIME	4.39 (0.02)	8.45 (0.04)	6.06 (0.39)	10.82 (0.43)	10.59 (1.03)	17.35 (1.08)
GLASSO	3.62 (0.02)	6.71 (0.03)	5.57 (0.11)	10.02 (0.14)	11.79 (0.43)	24.06 (1.29)
JOINT	3.68 (0.01)	7.16 (0.03)	5.24 (0.14)	9.56 (0.17)	8.28 (0.37)	13.83 (0.50)
FGL	3.12 (0.01)	5.75 (0.02)	3.85 (0.07)	6.84 (0.11)	7.08 (0.33)	12.26 (0.71)
JEMP	3.50 (0.01)	7.04 (0.02)	4.27 (0.08)	8.17 (0.14)	6.22 (0.29)	11.27 (0.60)

Table 4: Comparison summaries using Entropy loss (EL) and Frobenius loss (FL) over 50 replications for Model 4.

To compare performance of five different methods, we use the average entropy loss and the average Frobenius loss defined as,

$$\mathrm{EL} = G^{-1}\sum_{g=1}^{G}\left\{\mathrm{tr}(\Sigma_0^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}) - \log\det(\Sigma_0^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}) - p\right\},$$

$$\mathrm{FL} = G^{-1}\sum_{g=1}^{G} \parallel \Omega_0^{(\mathrm{g})} - \hat{\Omega}^{(\mathrm{g})} \parallel_F^2,$$

where $\| \,.\, \|_F$ is the Frobenius norm of a matrix.

Table 1 reports the results for Model 1. In terms of estimation accuracy, the three joint estimation methods, JEMP, FGL, and JOINT, outperform the two separate estimation methods while JEMP and FGL show better performance than JOINT. In Gaussian cases, FGL exhibits slightly smaller losses than JEMP. However, JEMP outperforms FGL in terms of entropy loss for some cases when the underlying distribution is t_5. If the true underlying distribution is t_3, then JEMP clearly outperforms FGL in both entropy loss and Frobenius loss for all cases. This indicates that our proposed JEMP can have some advantage in estimation for some non-Gaussian data. Overall, JEMP shows very competitive performance compared with other methods. Tables 2-3 report the results for Models 2 and 3 respectively. Performances of the methods show similar patterns as in Model 1. JEMP and FGL perform best while FGL is slightly better in Gaussian cases and JEMP has the best performance in the t_3 case.

Table 4 summarizes the results for Model 4 in which the true precision matrices have different structures. As in Models 1-3, our method outperforms JOINT, CLIME, and GLASSO for all cases. It shows competitive performance with FGL when the distribution is Gaussian or t_5. However, it outperforms FGL in the case of t_3 distribution. This indicates that our method works as well even when structures of precision matrices are different from each other. Note that the precision matrices in Model 4 share many zero components although their main structures are different. Joint methods can work better here since they encourage many common zeros to be estimated as zeros simultaneously.

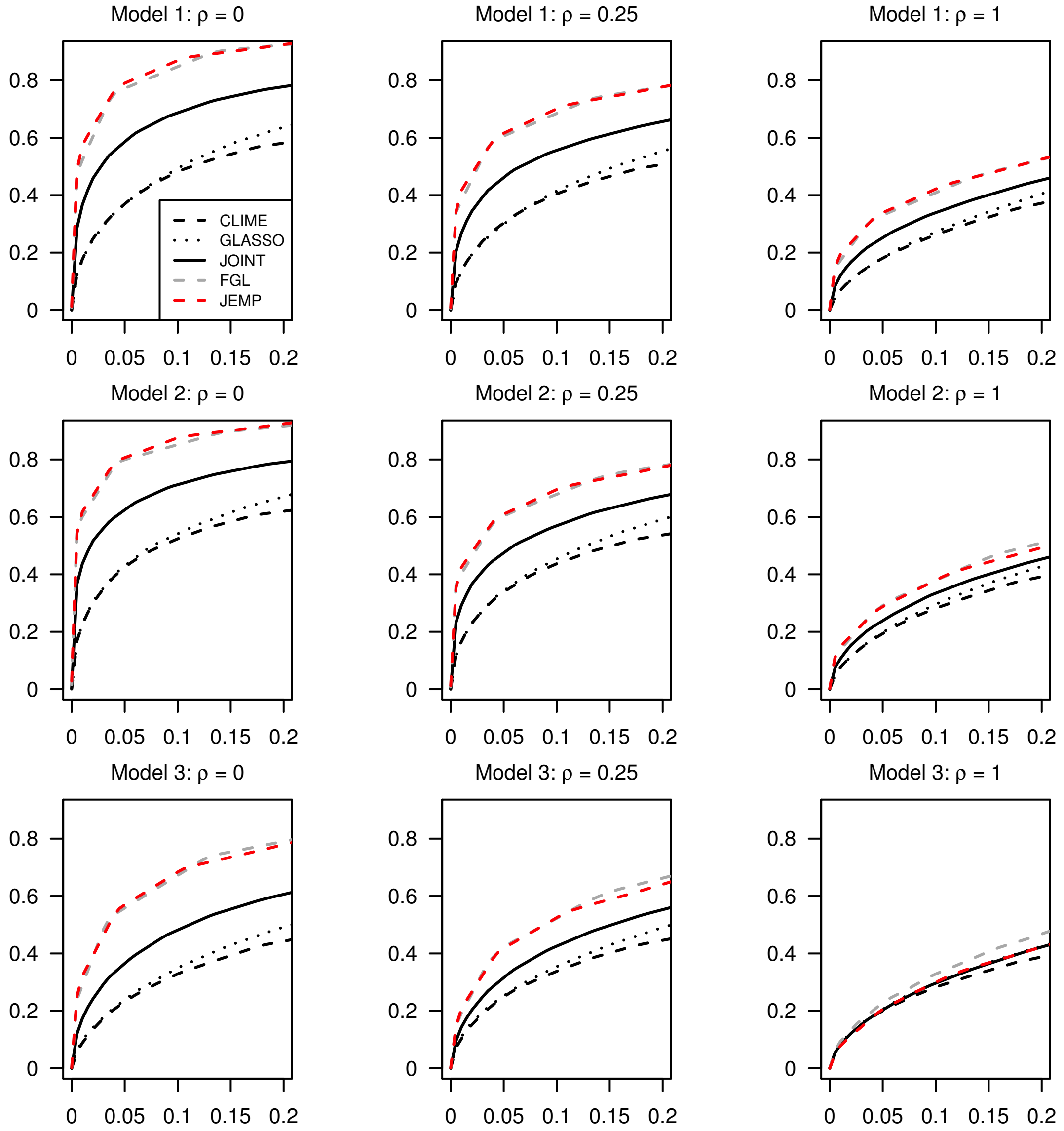

Figure 1: Receiver operating characteristic curves averaged over 50 replications from Gaussian distributions. In each panel, the horizontal and vertical axes are false positive rate and sensitivity respectively. Here, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries.

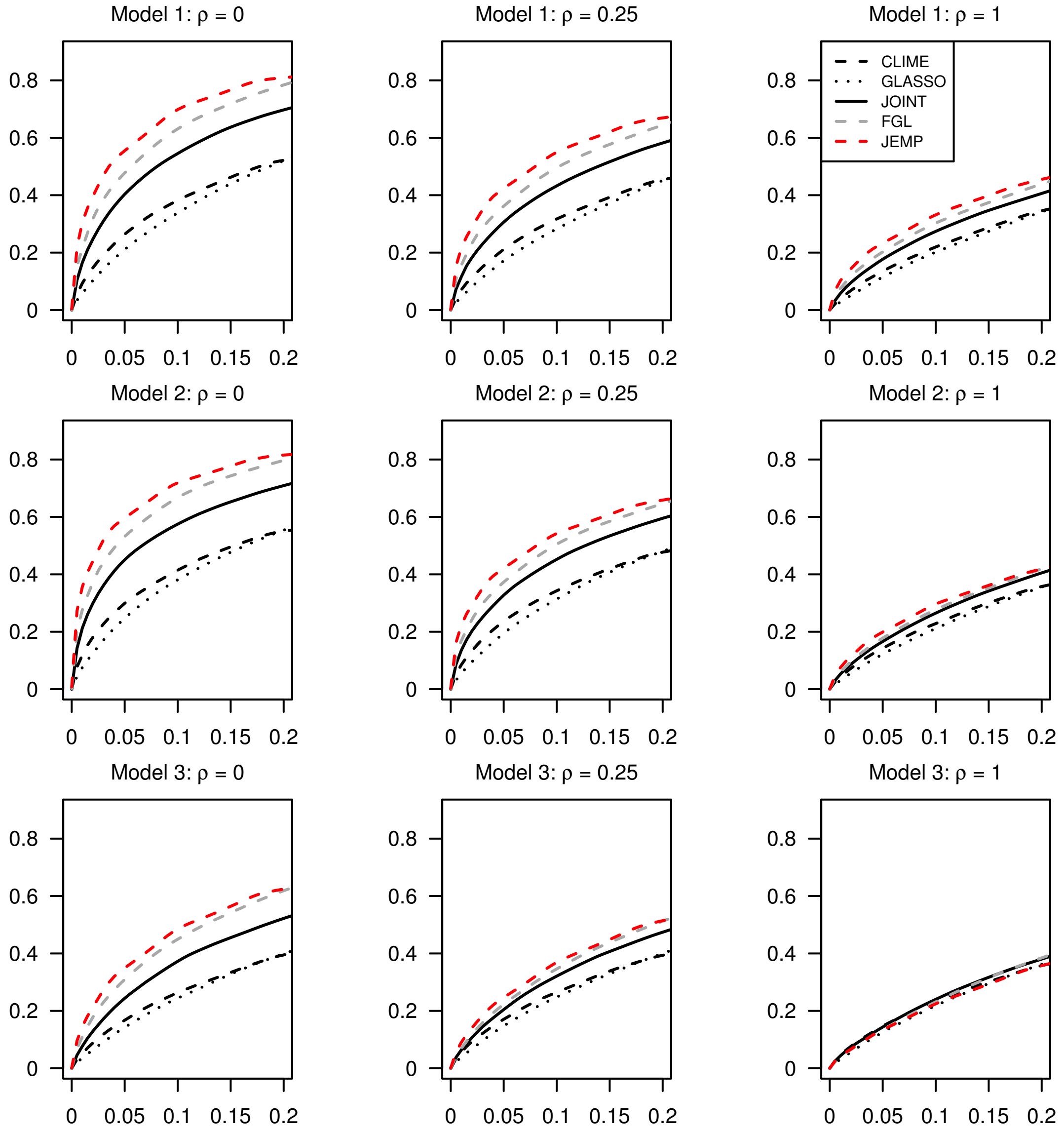

Figure 2: Receiver operating characteristic curves averaged over 50 replications from t_5 distributions. In each panel, the horizontal and vertical axes are false positive rate and sensitivity respectively. Here, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries.

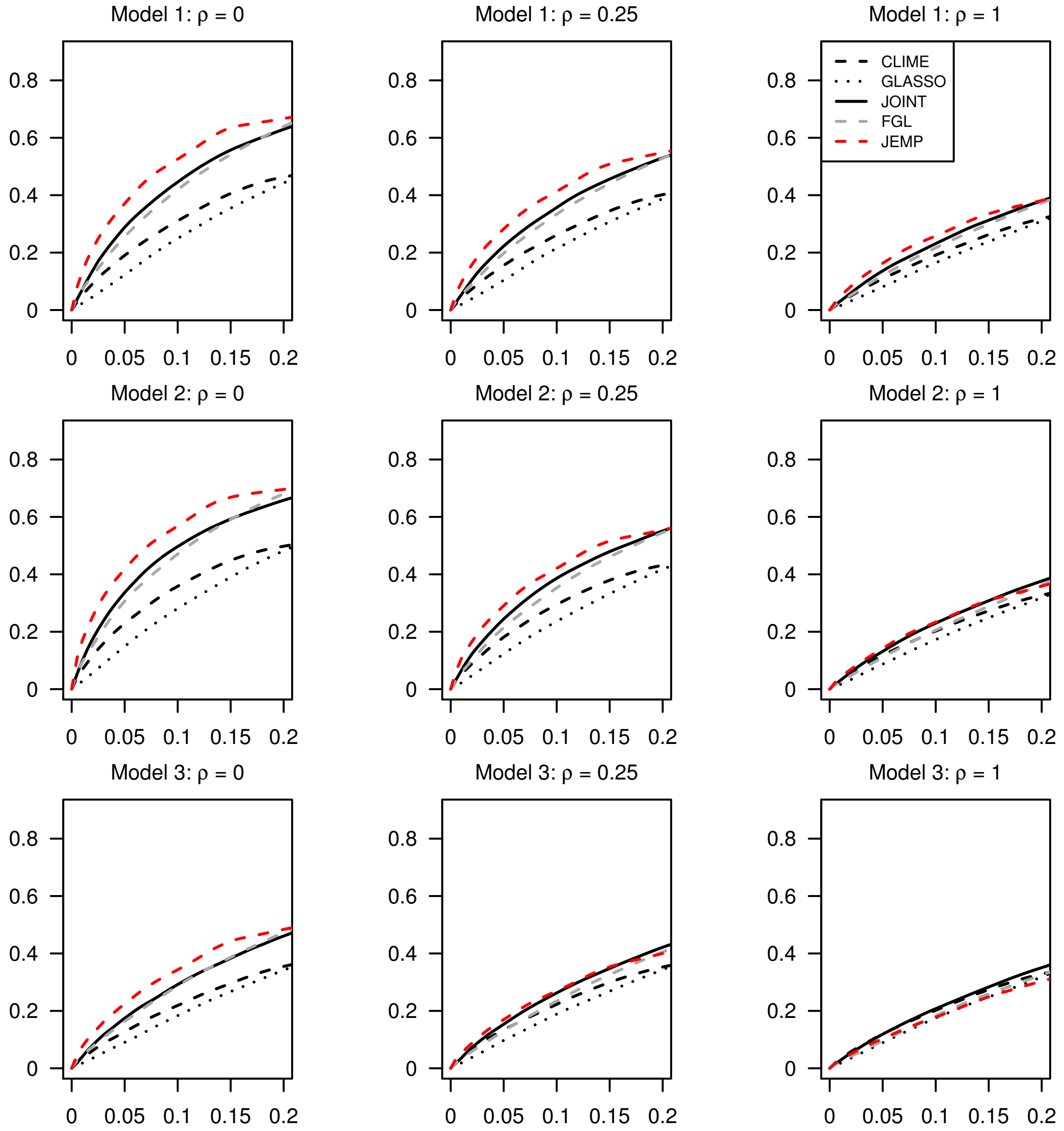

Figure 3: Receiver operating characteristic curves averaged over 50 replications from t_3 distributions. In each panel, the horizontal and vertical axes are false positive rate and sensitivity respectively. Here, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries.

Figures 1-3 show the estimated receiver operating characteristic (ROC) curves averaged over 50 replications. In the Gaussian case of Figure 1, JEMP and FGL show similar performance and outperform the others except the case of $\rho = 1$ in Model 3. In Figures 2 and 3 of multivariate t-distributions, it can be observed that JEMP has better ROC curves when $\rho = 0$ for all three models. It also shows better performance than the others when $\rho = 0.25$ for Models 1-2. When $\rho = 1$, all ROC curves move closer together. This is because the true precision matrices become much denser in terms of the number of edges and thus all methods have some difficulty in edge selection. Overall, our proposed JEMP estimator delivers competitive performance in terms of both estimation accuracy and selection.

Note that JEMP and FGL encourage the estimated precision matrices to be similar across all classes. This can be advantageous especially when the true precision matrices have many common values. Therefore, JEMP and FGL can have better performance than JOINT for such problems.

In terms of computational complexity, JEMP can be more intensive than separate estimation methods and JOINT as it involves a pair of tuning parameters (λ_1, λ_2) satisfying $\lambda_1 \leq \lambda_2$. The computational cost of JEMP can be potentially reduced using the ADMM algorithm discussed in Section 4 with a further improved algorithm for the least squares step.

6. Application on Glioblastoma Cancer Data

In this section, we apply our joint method to a Glioblastoma cancer data set. The data set consists of 17814 gene expression levels of 482 GBM patients. The patients were classified into four subtypes, namely, classical, mesenchymal, neural, and proneural with sample sizes of 127, 145, 85, and 125 respectively (Verhaak et al., 2010). These subtypes are shown to be different biologically, while at the same time, share similarities as well since they all belong to GBM cancer. In this application, we consider the signature genes reported by Verhaak et al. (2010). They established 210 signature genes for each subtype, which results 840 signature genes in total. These signature genes are highly distinctive for four subtypes and reported to have good predictive power for subtype prediction. In our analysis, the goal is to produce graphical presentation of relationships among these signature genes in each subtype based on the estimation of the precision matrices. Among the 840 signature genes, we excluded the genes with no subtype information or the genes with missing values. As a result, total 680 genes were included in our analysis. To produce interpretable graphical models using our JEMP estimator, we set the values of the tuning parameters as $\lambda_1 = 0.30$ and $\lambda_2 = 0.40$. JEMP estimated 214 edges shared among all subtypes, 9 edges present only in two subtypes, and 1 edge present only in three subtypes.

The resulting gene networks are shown in Figure 4. The black lines are the edges shared by all subtypes and the thick grey lines are the unique edges present only in two or three subtypes. It is noticeable that most of edges are black lines, which means that they appear in all subtypes. This indicates that the networks of the signature genes reported by Verhaak et al. (2010) may be very similar across all subtypes as they all belong to GBM cancer.

All of the small red network's genes in the upper region belong to the ZNF gene family. This network includes ZNF211, ZNF227, ZNF228, ZNF235, ZNF419, and ZNF671. These are known to be involved in making zinc finger proteins, which are regulatory proteins

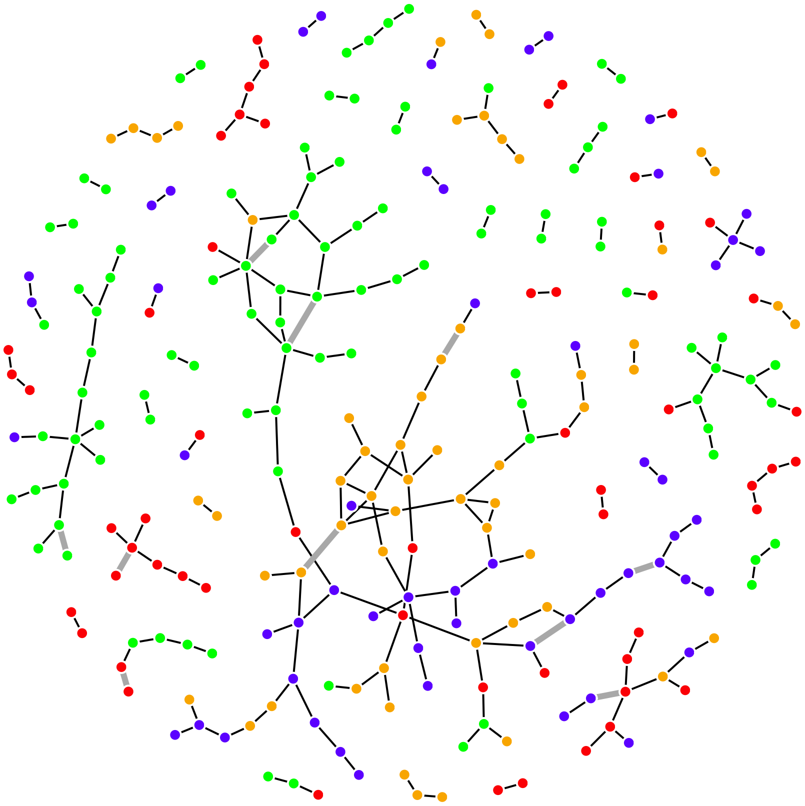

Figure 4: Graphical presentation of conditional dependence structures among genes using our estimator of precision matrices. The black lines are the edges shared in all subtypes and the thick grey lines are the unique edges present only in two or three subtypes. The red, green, blue and orange genes are classical, mesenchymal, proneural and neural genes respectively (Verhaak et al., 2010).

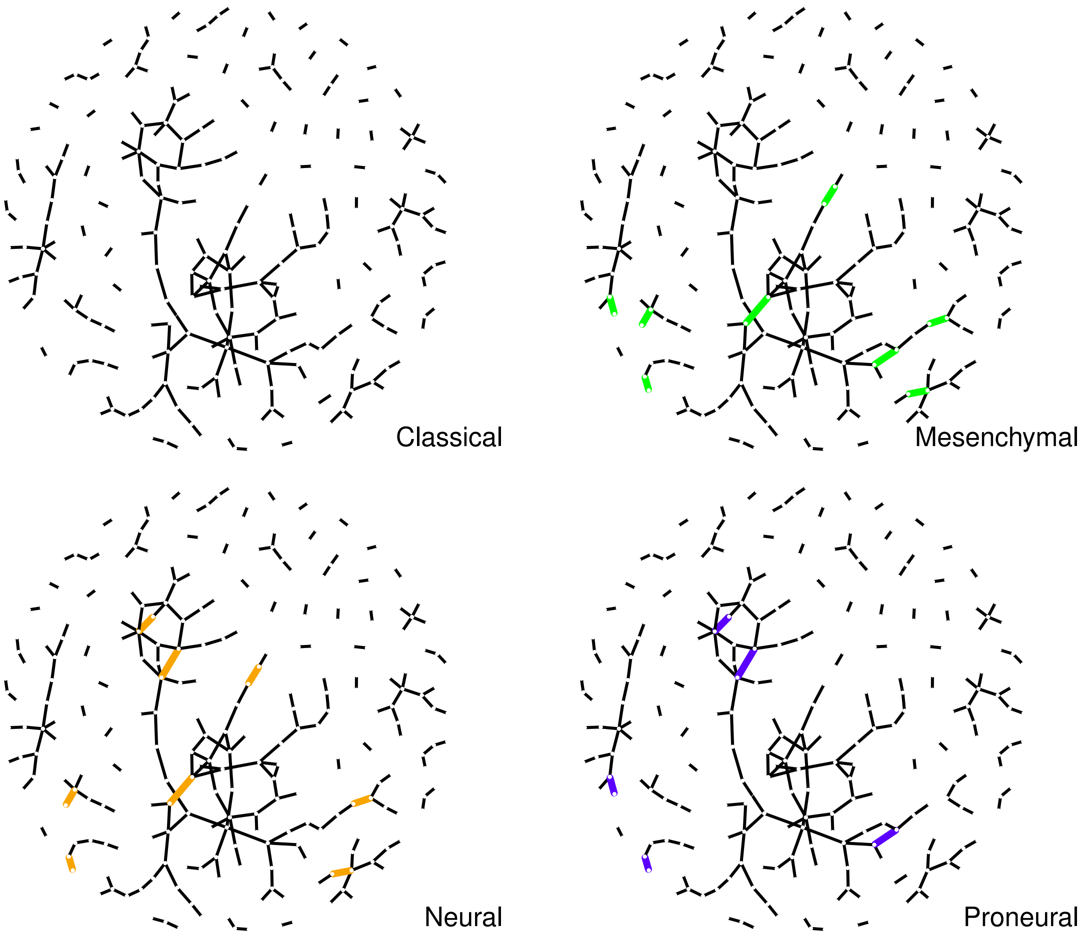

Figure 5: Four gene networks corresponding to four subtypes of the GMB cancer. In each network, the black lines are the edges shared in all subtypes. The colored lines are the edge shared only in two or three subtypes.

that are related to many cellular functions. As they are all involved in the same biological process, it may seem reasonable that this network is shared in all GBM subtypes.

The red genes are signature genes for the classical subtype. Likewise, green, blue and orange genes are the mesenchymal, proneural and neural signature genes respectively. Each class of signature genes tends to have more links with the genes in the same class. This is expected because each class of signature genes is more likely to be highly co-expressed.

Each estimated network for each subtype is depicted in Figure 5. The black lines are the edges shared by all subtypes and the colored lines are the edges appearing only in two or three subtypes. One interesting edge is the one between EGFR and MEOX2. It does not appear in the classical subtype while it is shared by all the other subtypes. EGFR is known to be involved in cell proliferation and Verhaak et al. (2010) demonstrated the essential role of this gene in GBM tumor genesis. Furthermore, high rates of EGFR alteration were claimed in the classical subtype. Therefore, studying the relationship between EGFR and MEOX2 can be an interesting direction for future investigation as only the classical subtype lacks this edge.

There are 9 edges appearing only in two subtypes. These include SCG3 and ACSBG1, GRIK5 and BTBD2, NCF4 and CSTA, IFI30 and BATF, HK3 and SLC11A1, ACSBG1 and SCG3, GPM6A and OLIG2, C1orf61 and CKB, and PPFIA2 and GRM1. It would be also interesting to investigate these relationships further as they are unique only in two subtypes. For example, the edge between OLIG2 and GPM6A does not appear in the proneural subtype while it is shared by Neural and Mesenchymal subtypes. High expression of OLIG2 was observed in the proneural subtype (Verhaak et al., 2010), which can down-regulate the tumor suppressor p21. Therefore, it may be helpful to investigate the relationship between OLIG2 and GPM6A for understanding the effect of OLIG2 in the proneural subtype.

Acknowledgments

The authors would like to thank the Action Editor Professor Francis Bach and three reviewers for their constructive comments and suggestions. The authors were supported in part by NIH/NCI grant R01 CA-149569, NIH/NCI P01 CA-142538, and NSF grant DMS-1407241.

Appendix A.

Write $\Sigma_0^{(g)} = (\sigma_{ij,0}^{(g)})$ and $\hat{\Sigma}^{(g)} = (\hat{\sigma}_{ij}^{(g)})$. Let $m_{j,0}$ and $r_{j,0}^{(g)}$ be the jth columns of M_0 and $R_0^{(g)}$ respectively. Define the jth columns of $\hat{M}$ and $\hat{R}^{(g)}$ as $\hat{m}_j$ and $\hat{r}_j^{(g)}$ respectively. We first state some results established by Cai et al. (2011) in the proof of their Theorem 1.

Lemma 4 *Suppose Condition 1 holds. For any fixed $g = 1, \ldots, G$, with probability greater than $1 - 4p^{-\tau}$,*

$$\max_{ij} |\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)}| \leq C_0 \left(\frac{\log p}{n} \right)^{1/2},$$

where C_0 is given in Theorem 1.

Proof [Proof of Theorem 1] It follows from Lemma 4 that

$$\max_{ij} |\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})}| \le \lambda_2/(3C_M) \quad \text{for all } g = 1, \ldots, G, \tag{6}$$

with probability greater than $1 - 4Gp^{-\tau}$. All following arguments assume (6) holds. First, we have that

$$\begin{aligned} |(\hat{\Omega}_1^{(\mathrm{g})} - \Omega_0^{(\mathrm{g})})e_j|_\infty &= |\Omega_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}\hat{\Omega}_1^{(\mathrm{g})} - I)e_j|_\infty \le ||\Omega_0^{(\mathrm{g})}||_{L_1}|(\Sigma_0^{(\mathrm{g})}\hat{\Omega}_1^{(\mathrm{g})} - I)e_j|_\infty \\ &\le C_M \left\{ |(\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})})\hat{\Omega}_1^{(\mathrm{g})}e_j|_\infty + |(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}_1^{(\mathrm{g})} - I)e_j|_\infty \right\} \\ &\le C_M |\hat{\Omega}_1^{(\mathrm{g})}e_j|_1 |\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})}|_\infty + C_M\lambda_2 \\ &\le |\hat{\Omega}_1^{(\mathrm{g})}e_j|_1 \lambda_2/3 + C_M\lambda_2, \end{aligned}$$

for all $g = 1, \ldots, G$. Second, note that $\{M_0, R_0^{(1)}, \ldots, R_0^{(\mathrm{G})}\}$ is a feasible solution of (3) as $|I - \hat{\Sigma}^{(\mathrm{g})}(M_0 + R_0^{(\mathrm{g})})|_\infty = |(\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})})\Omega_0^{(\mathrm{g})}|_\infty \le ||\Omega_0^{(\mathrm{g})}||_{L_1}|\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})}|_\infty \le C_M\lambda_2/(3C_M) < \lambda_2$ and $\lambda_1 = \lambda_2$. Therefore, we have that

$$\begin{aligned} \sum_{g=1}^G |(\hat{\Omega}_1^{(\mathrm{g})} - \Omega_0^{(\mathrm{g})})e_j|_\infty &\le \sum_{g=1}^G |\hat{\Omega}_1^{(\mathrm{g})}e_j|_1 \lambda_2/3 + GC_M\lambda_2 \le G\left\{ |\hat{m}_j|_1 + G^{-1}\sum_{g=1}^G |\hat{r}_j^{(\mathrm{g})}|_1 \right\} \lambda_2/3 + GC_M\lambda_2 \\ &\le G\left\{ |m_{j,0}|_1 + G^{-1}\sum_{g=1}^G |r_{j,0}^{(\mathrm{g})}|_1 \right\} \lambda_2/3 + GC_M\lambda_2 \\ &\le G3C_M\lambda_2/3 + GC_M\lambda_2 = 2GC_M\lambda_2 = 6GC_M^2C_0(\log p/n)^{1/2}. \end{aligned}$$

By the inequality

$$\max_{ij}\left(\frac{1}{G}\sum_{g=1}^G |\hat{\omega}_{ij}^{(\mathrm{g})} - \omega_{ij,0}^{(\mathrm{g})}| \right) \le \max_j \frac{1}{G}\sum_{g=1}^G |(\hat{\Omega}_1^{(\mathrm{g})} - \Omega_0^{(\mathrm{g})})e_j|_\infty \le 6C_M^2C_0\left(\frac{\log p}{n}\right)^{1/2},$$

the proof is completed. ■

Lemma 5 *With probability greater than* $1 - 2(1+G)p^{-\tau}$*, the following holds:*

$$\max_{ij} |\sum_{g=1}^G (\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})})| \le C_0\left(\frac{G\log p}{n}\right)^{1/2}.$$

Proof We adopt a similar technique used in Cai et al. (2011) for the proof of their Theorem 1. Without loss of generality, we assume that $\mu_i^{(\mathrm{g})} = 0$ for all i and g. Let $y_{kij}^{(\mathrm{g})} := x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})} - E(x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})})$. Define $\bar{x}_i^{(\mathrm{g})} := \sum_{k=1}^n x_{ki}^{(\mathrm{g})}/n; i = 1, \ldots, p, g = 1, \ldots, G$. Then $\sum_{g=1}^G (\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})}) = \sum_{g=1}^G \left(\sum_{k=1}^n y_{kij}^{(\mathrm{g})}/n - \bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})} \right)$. Let $t := \eta(\log p)^{1/2}(nG)^{-1/2}$ and $C_1 := 2 + \tau + \eta^{-1}K^2$. Using the Markov's inequality and the inequality $|\exp(s) - 1 - s| \le s^2\exp\{\max(s, 0)\}$ for any $s \in \mathcal{R}$, we can show that

$$
\begin{aligned}
&\mathrm{pr}\left\{\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})} \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} \\
&= \mathrm{pr}\left\{\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})} \geq \eta^{-1}C_1\left(nG\log p\right)^{1/2}\right\} \\
&\leq \exp\left\{-t\eta^{-1}C_1(nG\log p)^{1/2}\right\} E\left\{\exp\left(t\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})}\right)\right\} \\
&= \exp\left\{-C_1\log p\right\}\prod_{g=1}^{G}\prod_{k=1}^{n} E\left\{\exp(ty_{kij}^{(\mathrm{g})})\right\} \\
&= \exp\left[-C_1\log p + \sum_{g=1}^{G} n\log\left\{E\left(e^{ty_{kij}^{(\mathrm{g})}}\right)\right\}\right] \\
&\leq \exp\left[-C_1\log p + \sum_{g=1}^{G} n\left\{E\left(e^{ty_{kij}^{(\mathrm{g})}}\right) - 1\right\}\right] \\
&= \exp\left[-C_1\log p + \sum_{g=1}^{G} n\left\{E\left(e^{ty_{kij}^{(\mathrm{g})}} - ty_{kij}^{(\mathrm{g})} - 1\right)\right\}\right] \\
&\leq \exp\left\{-C_1\log p + \sum_{g=1}^{G} nt^2 E\left({y_{kij}^{(\mathrm{g})}}^{2} e^{|ty_{kij}^{(\mathrm{g})}|}\right)\right\} \\
&\leq \exp\left\{-C_1\log p + \sum_{g=1}^{G}(\eta G)^{-1}K^2\log p\right\}. \qquad (7)
\end{aligned}
$$

The last inequality (7) holds since

$$
nt^2 E\left({y_{kij}^{(\mathrm{g})}}^{2} e^{|ty_{kij}^{(\mathrm{g})}|}\right) = (\eta G)^{-1}(\log p) E\left\{\left(\eta^{3/2}|y_{kij}^{(\mathrm{g})}|\right)^2 e^{t|y_{kij}^{(\mathrm{g})}|}\right\}
$$

and

$$
\begin{aligned}
E\left\{\left(\eta^{3/2}|y_{kij}^{(\mathrm{g})}|\right)^2 e^{t|y_{kij}^{(\mathrm{g})}|}\right\} &\leq E\left\{e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}e^{t|y_{kij}^{(\mathrm{g})}|}\right\} \leq E\left\{e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}\right\} \\
&\leq E\left\{e^{\eta|y_{kij}^{(\mathrm{g})}|}\right\} \leq E\left\{e^{\eta|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}| + \eta E\left(|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}|\right)}\right\} \\
&\leq \left\{E\left(e^{\eta|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}|}\right)\right\}^2 \leq \left\{E\left(e^{\eta {x_{ki}^{(\mathrm{g})}}^{2}/2 + \eta {x_{kj}^{(\mathrm{g})}}^{2}/2}\right)\right\}^2 \\
&\leq E\left(e^{\eta {x_{ki}^{(\mathrm{g})}}^{2}}\right)E\left(e^{\eta {x_{kj}^{(\mathrm{g})}}^{2}}\right) \leq K^2.
\end{aligned}
$$

From (7), it follows that

$$
\mathrm{pr}\left\{\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})} \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} \leq \exp\left\{-C_1\log p + \eta^{-1}K^2\log p\right\} \leq p^{-(\tau+2)}.
$$

Therefore, we have

$$\mathrm{pr}\left\{\max_{ij}\left|\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})}\right| \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} \leq 2p^{-\tau}. \tag{8}$$

Next, let $C_2 = 2+\tau+\eta^{-1}(eK)^2$. Cai et al. (2011) showed in the proof of their Theorem 1 that

$$\mathrm{pr}\left(\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2\log p/n\right) \leq 2p^{-\tau-1}.$$

Using this result, we have that

$$\begin{aligned}\mathrm{pr}\left(\max_{ij}|\sum_{g=1}^{G}\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2 G\log p/n\right) &\leq \mathrm{pr}\left(\sum_{g=1}^{G}\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2 G\log p/n\right)\\ &\leq \sum_{g=1}^{G}\mathrm{pr}\left(\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2\log p/n\right)\\ &\leq \sum_{g=1}^{G} 2p^{-\tau-1} \leq 2Gp^{-\tau} \end{aligned} \tag{9}$$

By (8), (9) and the inequality $C_0 > \eta^{-1}C_1 + \eta^{-2}C_2^2(G\log p/n)^{1/2}$, we see that

$$\begin{aligned}&\mathrm{pr}\left\{\max_{ij}|\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})})| \geq C_0\left(\frac{G\log p}{n}\right)^{1/2}\right\}\\ &\leq \mathrm{pr}\left\{\max_{ij}\left|\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})}\right| \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\}\\ &+ \mathrm{pr}\left(\max_{ij}|\sum_{g=1}^{G}\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2 G\log p/n\right)\\ &\leq 2(1+G)p^{-\tau}.\end{aligned}$$

The proof is completed. ■

Proof [Proof of Theorem 2] By Lemma 4 and 5, we see that

$$\max_{ij}|\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})})| \leq C_0\left(\frac{G\log p}{n}\right)^{1/2} \text{and} \max_{ij}|\hat{\sigma}_{ij}^{(\mathrm{g})} - \sigma_{ij,0}^{(\mathrm{g})}| \leq C_0\left(\frac{\log p}{n}\right)^{1/2}, \tag{10}$$

for all $g = 1, \ldots, G$ with probability greater than $1 - 2(1+3G)p^{-\tau}$. All following arguments assume (10) holds. Note that $\{M_0, R_0^{(1)}, \ldots, R_0^{(\mathrm{G})}\}$ is a feasible solution of (3) as

$$\begin{aligned}|I - \hat{\Sigma}^{(\mathrm{g})}(M_0 + R_0^{(\mathrm{g})})|_\infty = |(\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})})\Omega_0^{(\mathrm{g})}|_\infty &\leq ||\Omega_0^{(\mathrm{g})}||_{L_1}|\Sigma_0^{(\mathrm{g})} - \hat{\Sigma}^{(\mathrm{g})}|_\infty\\ &\leq C_M C_0(\log p/n)1/2 = \lambda_2\end{aligned}$$

and

$$
\begin{aligned}
&|G^{-1}\sum_{g=1}^{G}\left\{I-\hat{\Sigma}^{(\mathrm{g})}(M_0+R_0^{(\mathrm{g})})\right\}|_\infty \\
&\le |G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})M_0|_\infty+|G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})R_0^{(\mathrm{g})}|_\infty \\
&\le ||M_0||_{L_1}|G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})|_\infty+G^{-1}\sum_{g=1}^{G}||R_0^{(\mathrm{g})}||_{L_1}|\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})}|_\infty \\
&\le C_MC_0\left\{\log p/(nG)\right\}^{1/2}+C_RC_0\left\{\log p/(nG)\right\}^{1/2}=\lambda_1.
\end{aligned}
$$

Now, we find an upper bound of $|G(\hat{M}-M_0)e_j|_\infty=|\sum_{g=1}^{G}(\hat{\Omega}_1^{(\mathrm{g})}-\Omega_0^{(\mathrm{g})})e_j|_\infty$. In particular, we use

$$
|\sum_{g=1}^{G}(\hat{\Omega}_1^{(\mathrm{g})}-\Omega_0^{(\mathrm{g})})e_j|_\infty\le|\sum_{g=1}^{G}\Omega_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{\Omega}_1^{(\mathrm{g})}e_j|_\infty+|\sum_{g=1}^{G}\Omega_0^{(\mathrm{g})}(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}_1^{(\mathrm{g})}-I)e_j|_\infty. \quad (11)
$$

First, consider the first term in the right-hand side of (11). We can show that

$$
\begin{aligned}
|\sum_{g=1}^{G}\Omega_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{\Omega}_1^{(\mathrm{g})}e_j|_\infty &\le |\sum_{g=1}^{G}M_0(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{m}_j|_\infty+|\sum_{g=1}^{G}M_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{r}_j^{(\mathrm{g})}|_\infty \\
&\quad+|\sum_{g=1}^{G}R_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{m}_j|_\infty+|\sum_{g=1}^{G}R_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{r}_j^{(\mathrm{g})}|_\infty \\
&\le ||M_0||_{L_1}\left\{|\sum_{g=1}^{G}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})|_\infty|\hat{m}_j|_1+\sum_{g=1}^{G}|\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})}|_\infty|\hat{r}_j^{(\mathrm{g})}|_1\right\} \\
&\quad+\sum_{g=1}^{G}|R_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})|_\infty|\hat{m}_j|_1+\sum_{g=1}^{G}|R_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})|_\infty|\hat{r}_j^{(\mathrm{g})}|_1.
\end{aligned}
$$

Using the assumptions $||R_0^{(\mathrm{g})}||_{L_1}\le C_R$ and $\sum_{g=1}^{G}||R_0^{(\mathrm{g})}||_{L_1}\le G^{1/2}C_R$, we have

$$
\begin{aligned}
|\sum_{g=1}^{G}\Omega_0^{(\mathrm{g})}(\Sigma_0^{(\mathrm{g})}-\hat{\Sigma}^{(\mathrm{g})})\hat{\Omega}_1^{(\mathrm{g})}e_j|_\infty &\le C_MC_0(G\log p/n)^{1/2}|\hat{m}_j|_1+C_MC_0(\log p/n)^{1/2}\sum_{g=1}^{G}|\hat{r}_j^{(\mathrm{g})}|_1 \\
&\quad+C_RC_0(G\log p/n)^{1/2}|\hat{m}_j|_1+C_RC_0(\log p/n)^{1/2}\sum_{g=1}^{G}|\hat{r}_j^{(\mathrm{g})}|_1 \\
&\le C_0(C_M+C_R)(G\log p/n)^{1/2}(|\hat{m}_j|_1+G^{-1/2}\sum_{g=1}^{G}|\hat{r}_j^{(\mathrm{g})}|_1) \\
&\le C_0(C_M+C_R)(G\log p/n)^{1/2}(|m_{j,0}|_1+G^{-1/2}\sum_{g=1}^{G}|r_{j,0}^{(\mathrm{g})}|_1) \\
&\le C_0(C_M+C_R)^2(G\log p/n)^{1/2}. \quad (12)
\end{aligned}
$$

For the second term in the right-hand side of (11), note that

$$
\begin{aligned}
&|\sum_{g=1}^{G}\Omega_0^{(\mathrm{g})}(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}_1^{(\mathrm{g})}-I)e_j|_\infty \\
&\le |\sum_{g=1}^{G}M_0(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}-I)e_j|_\infty + |\sum_{g=1}^{G}R_0^{(\mathrm{g})}(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}-I)e_j|_\infty \\
&\le ||M_0||_{L_1}|\sum_{g=1}^{G}(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}-I)e_j|_\infty + \sum_{g=1}^{G}||R_0^{(\mathrm{g})}||_{L_1}|(\hat{\Sigma}^{(\mathrm{g})}\hat{\Omega}^{(\mathrm{g})}-I)e_j|_\infty \\
&\le C_M\lambda_1 + G^{1/2}C_R\lambda_2 = C_0C_M(C_M+2C_R)(G\log p/n)^{1/2}. \qquad (13)
\end{aligned}
$$

By (11), (12), (13) and the equality $|\hat{M}-M_0|_\infty = \max_j|(\hat{M}-M_0)e_j|_\infty$, we have

$$
|\hat{M}-M_0|_\infty \le C_0(2C_M^2+4C_MC_R+C_R^2)\left(\frac{\log p}{nG}\right)^{1/2}.
$$

The proof is completed. ■

Proof [Proof of Theorem 3] By Theorem 1, we see that

$$
\max_{ij}\sum_{g=1}^{G}|\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| \le 2GC_M\lambda_2 \le \delta_n, \qquad (14)
$$

with probability greater than $1-4Gp^{-\tau}$. We show that $S_0=\hat{S}$ when (14) holds. For any $(i,j,g)\notin S_0$, we have $|\hat{\omega}_{ij}^{(\mathrm{g})}| = |\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| \le \sum_{g=1}^{G}|\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| \le \delta_n$. Therefore, we see $\tilde{\omega}_{ij}^{(\mathrm{g})}=0$, which implies $\hat{S}\subset S_0$. On the other hand, for any $(i,j,g)\in S_0$, we have $|\hat{\omega}_{ij}^{(\mathrm{g})}| \ge |\omega_{ij,0}^{(\mathrm{g})}| - |\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| \ge |\omega_{ij,0}^{(\mathrm{g})}| - \sum_{g=1}^{G}|\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| > \delta_n$. Therefore, we see that $\tilde{\omega}_{ij}^{(\mathrm{g})}\neq 0$, which implies $S_0\subset\hat{S}$. In summary, we see that $S_0=\hat{S}$ if (14) holds, which implies that $\mathrm{pr}(S_0=\hat{S}) \ge \mathrm{pr}(\max_{ij}\sum_{g=1}^{G}|\hat{\omega}_{ij}^{(\mathrm{g})}-\omega_{ij,0}^{(\mathrm{g})}| \le \delta_n)$. ■

References

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76:373–379, 2014.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multitask learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, Seattle, Washington, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3:521–541, 2009.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98:1–15, 2011.

Jean Honorio and Dimitris Samaras. Simultaneous and group-sparse multi-task learning of gaussian graphical models. *arXiv:1207.4255*, 2012.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37:4254–4278, 2009.

Wonyul Lee, Ying Du, Wei Sun, David Neil Hayes, and Yufeng Liu. Multiple response regression for gaussian mixture models with known labels. *Statistical Analysis and Data Mining*, 5:493–508, 2012.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Haotian Pang, Han Liu, and Robert Vanderbei. *fastclime: A fast solver for parameterized lp problems and constrained l_1-minimization approach to sparse precision matrix estimation*, 2014. R package version 1.2.4.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104:735–746, 2009.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael OKelly, Pablo Tamayo, Barbara A. Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, D. Neil Hayes, and The Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17: 98–110, 2010.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.