SIAM J. OPTIM.
Vol. 25, No. 2, pp. 916–943

FUSED MULTIPLE GRAPHICAL LASSO*

SEN YANG†, ZHAOSONG LU‡, XIAOTONG SHEN§, PETER WONKA†, AND JIEPING YE¶

Abstract. In this paper, we consider the problem of estimating multiple graphical models simultaneously using the fused lasso penalty, which encourages adjacent graphs to share similar structures. A motivating example is the analysis of brain networks of Alzheimer's disease using neuroimaging data. Specifically, we may wish to estimate a brain network for the normal controls (NC), a brain network for the patients with mild cognitive impairment (MCI), and a brain network for Alzheimer's patients (AD). We expect the two brain networks for NC and MCI to share common structures but not to be identical to each other; similarly for the two brain networks for MCI and AD. The proposed formulation can be solved using a second-order method. Our key technical contribution is to establish the necessary and sufficient condition for the graphs to be decomposable. Based on this key property, a simple screening rule is presented, which decomposes the large graphs into small subgraphs and allows an efficient estimation of multiple independent (small) subgraphs, dramatically reducing the computational cost. We perform experiments on both synthetic and real data; our results demonstrate the effectiveness and efficiency of the proposed approach.

Key words. fused multiple graphical lasso, screening, second-order method

AMS subject classifications. 90C22, 90C25, 90C47, 65K05, 62J10

DOI. 10.1137/130936397

1. Introduction.

Undirected graphical models explore the relationships among a set of random variables through their joint distribution. The estimation of undirected graphical models has applications in many domains, such as computer vision, biology, and medicine [12, 18, 51]. One instance is the analysis of gene expression data. As shown in many biological studies, genes tend to work in groups based on their biological functions, and there exist some regulatory relationships between genes [6]. Such biological knowledge can be represented as a graph, where nodes are the genes, and edges describe the regulatory relationships. Graphical models provide a useful tool for modeling these relationships and can be used to explore gene activities. One of the most widely used graphical models is the Gaussian graphical model (GGM), which assumes the variables to be Gaussian distributed [2, 54]. In the framework of the GGM, the problem of learning a graph is equivalent to estimating the inverse of the covariance matrix (precision matrix), since the nonzero off-diagonal elements of the precision matrix represent edges in the graph [2, 54].

In recent years many research efforts have focused on estimating the precision matrix and the corresponding graphical model (see, for example, [2, 11, 17, 18, 24, 25, 28, 29, 32, 34, 38, 54]). Meinshausen and Bühlmann [34] estimated edges for

*Received by the editors September 10, 2013; accepted for publication (in revised form) January 16, 2015; published electronically May 7, 2015. This work was supported in part by research grants from the NIH (R01 LM010730) and NSF (IIS-0953662, III-1421057, and III-1421100).
http://www.siam.org/journals/siopt/25-2/93639.html

†School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287 (senyang@asu.edu, peter.wonka@asu.edu).
‡Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 156, Canada (zhaosong @sfu.ca).
§School of Statistics, University of Minnesota, Minneapolis, MN 55455 (xshen@umn.edu).
¶Department of Computational Medicine and Bioinformatics and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2218 (jpye@umich.edu).

each node in the graph by fitting a lasso problem [42] using the remaining variables as predictors. Yuan and Lin [54] and Banerjee, el Ghaoui, and d'Aspremont [2] proposed a penalized maximum likelihood model using ℓ_1 regularization to estimate the sparse precision matrix. Numerous methods have been developed for solving this model. For example, d'Aspremont, Banerjee, and el Ghaoui [9] and Lu [28, 29] studied Nesterov's smooth gradient methods [36] for solving this problem or its dual. Banerjee, el Ghaoui, and d'Aspremont [2] and Friedman, Hastie, and Tibshirani [11] proposed block coordinate ascent methods for solving the dual problem. The latter method [11] is widely referred to as graphical lasso (GLasso). Mazumder and Hastie [32] proposed a new algorithm called DP-GLasso, each step of which is a box-constrained QP problem. Scheinberg and Rish [40] proposed a coordinate descent method for solving this model in a greedy approach. Yuan [55] and Scheinberg, Ma, and Goldfarb [39] applied the alternating direction method of multipliers (ADMM) [4] to solve this problem. Li and Toh [24] and Yuan and Lin [54] proposed to solve this problem using interior point methods. Wang, Sun, and Toh [46], Hsieh et al. [17], Olsen et al. [38], and Dinh, Kyrillidis, and Cevher [10] studied the Newton method for solving this model. The main challenge of estimating a sparse precision matrix for the problems with a large number of nodes (variables) is its intensive computation. Witten, Friedman, and Simon [49] and Mazumder and Hastie [31] independently derived a necessary and sufficient condition for the solution of a single graphical lasso to be block diagonal (subject to some rearrangement of variables). This can be used as a simple screening test to identify the associated blocks, and the original problem can thus be decomposed into a group of smaller sized but independent problems corresponding to these blocks. When the number of blocks is large, it can achieve massive computational gain. However, these formulations assume that observations are independently drawn from a single Gaussian distribution. In many applications the observations may be drawn from multiple Gaussian distributions; in this case, multiple graphical models need to be estimated.

There are some recent works on the estimation of multiple precision matrices [8, 12, 13, 14, 21, 22, 35, 56]. Guo et al. [12] proposed a method to jointly estimate multiple graphical models using a hierarchical penalty. However, their model is not convex. Honorio and Samaras [14] proposed a convex formulation to estimate multiple graphical models using the $\ell_{1,\infty}$ regularizer. Hara and Washio [13] introduced a method to learn common substructures among multiple graphical models. Danaher, Wang, and Witten [8] estimated multiple precision matrices simultaneously using a pairwise fused penalty and grouping penalty. ADMM was used to solve the problem, but it requires computing multiple eigendecompositions at each iteration. Mohan et al. [35] proposed estimating multiple precision matrices based on the assumption that the network differences are generated from node perturbations. Compared with single graphical model learning, learning multiple precision matrices jointly is even more challenging. Recently, a necessary and sufficient condition for multiple graphs to be decomposable was proposed in [8]. However, such necessary and sufficient condition was restricted to two graphs only when the fused penalty is used. It is not clear whether this condition can be extended to the more general case with more than two graphs, which is the case in brain network modeling.

There are several types of fused penalties that can be used for estimating multiple (more than two) graphs such as the pairwise fused penalty and sequential fused penalty [43]. In this paper we set out to address the sequential fused case first, because we work on practical applications that can be more appropriately formulated using the sequential formulation. Specifically, we consider the problem of estimating multiple graphical models by maximizing a penalized log likelihood with ℓ_1 and sequential fused regularization. The ℓ_1 regularization yields a sparse solution, and the fused regularization encourages adjacent graphs to be similar. The graphs considered in this paper have a natural order, which is common in many applications. A motivating example is the modeling of brain networks for Alzheimer's disease using neuroimaging data such as Positron emission tomography (PET). In this case, we want to estimate graphical models for three groups: normal controls (NC), patients of mild cognitive impairment (MCI), and Alzheimer's patients (AD). These networks are expected to share some common connections, but they are not identical. Furthermore, the networks are expected to evolve over time, in the order of disease progression from NC to MCI to AD. Estimating the graphical models separately fails to exploit the common structures among them. It is thus desirable to jointly estimate the three networks (graphs). Our key technical contribution is to establish the necessary and sufficient condition for the solution of the fused multiple graphical lasso (FMGL) to be block diagonal. The duality theory and several other tools in linear programming are used to derive the necessary and sufficient condition. Based on this crucial property of the FMGL, we develop a screening rule which enables efficient estimation of large multiple precision matrices for the FMGL. The proposed screening rule can be combined with any algorithms to reduce the computational cost. We employ a second-order method [17, 23, 44] to solve the FMGL, where each step is solved by the spectral projected gradient method [30, 50]. In addition, we propose an active set identification scheme to identify the variables to be updated in each step of the second-order method, which reduces the computation cost of each step. We conduct experiments on both synthetic and real data; our results demonstrate the effectiveness and efficiency of the proposed approach.

1.1. Notation. In this paper, $\Re$ stands for the set of all real numbers, $\Re^n$ denotes the n-dimensional Euclidean space, and the set of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$. All matrices are presented in bold format. The space of symmetric matrices is denoted by $\mathcal{S}^n$. If $\mathbf{X} \in \mathcal{S}^n$ is positive semidefinite (resp., definite), we write $\mathbf{X} \succeq 0$ (resp., $\mathbf{X} \succ 0$). Also, we write $\mathbf{X} \succeq \mathbf{Y}$ to mean $\mathbf{X} - \mathbf{Y} \succeq 0$. The cone of positive semidefinite matrices in $\mathcal{S}^n$ is denoted by $\mathcal{S}^n_+$. Given matrices $\mathbf{X}$ and $\mathbf{Y}$ in $\Re^{m\times n}$, the standard inner product is defined by $\langle \mathbf{X}, \mathbf{Y} \rangle := \mathrm{tr}(\mathbf{X}\mathbf{Y}^T)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. $\mathbf{X} \circ \mathbf{Y}$ and $\mathbf{X} \otimes \mathbf{Y}$ mean the Hadamard and Kronecker product, respectively, of $\mathbf{X}$ and $\mathbf{Y}$. We denote the identity matrix by $\mathbf{I}$, whose dimension should be clear from the context. The determinant and the minimal eigenvalue of a real symmetric matrix $\mathbf{X}$ are denoted by $\det(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$, respectively. Given a matrix $\mathbf{X} \in \Re^{n\times n}$, $\mathrm{diag}(\mathbf{X})$ denotes the vector formed by the diagonal of $\mathbf{X}$; that is, $\mathrm{diag}(\mathbf{X})_i = \mathbf{X}_{ii}$ for $i = 1, \ldots, n$. $\mathrm{Diag}(\mathbf{X})$ is the diagonal matrix which has the same diagonal as $\mathbf{X}$. $\mathrm{vec}(\mathbf{X})$ is the vectorization of $\mathbf{X}$. In addition, $\mathbf{X} > 0$ means that all entries of $\mathbf{X}$ are positive.

The rest of the paper is organized as follows. We introduce the formulation of the FMGL in section 2. The screening rule is presented in section 3. The proposed second-order method is presented in section 4. The experimental results are shown in section 5. We conclude the paper in section 6.

2. Fused multiple graphical lasso. Assume we are given K data sets, $x^{(k)} \in \Re^{n_k \times p}$, $k = 1, \ldots, K$, with $K \geq 2$, where n_k is the number of samples and p is the number of features. The p features are common for all K data sets, and all $\sum_{k=1}^K n_k$ samples are independent. Furthermore, the samples within each data set $x^{(k)}$ are identically distributed with a p-variate Gaussian distribution with zero mean and positive definite covariance matrix $\mathbf{\Sigma}^{(k)}$, and there are many conditionally independent pairs of features; i.e., the precision matrix $\mathbf{\Theta}^{(k)} = (\mathbf{\Sigma}^{(k)})^{-1}$ should be sparse. For notational simplicity, we assume that $n_1 = \cdots = n_K = n$. Denote the sample covariance matrix for each data set $x^{(k)}$ as $\mathbf{S}^{(k)}$ with $\mathbf{S}^{(k)} = \frac{1}{n}(x^{(k)})^T x^{(k)}$, and $\mathbf{\Theta} = (\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(K)})$. Then the negative log likelihood for the data takes the form of

$$\sum_{k=1}^{K} \left(-\log\det(\mathbf{\Theta}^{(k)}) + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}) \right). \tag{2.1}$$

Clearly, minimizing (2.1) leads to the maximum likelihood estimate (MLE), $\widehat{\mathbf{\Theta}}^{(k)} = (\mathbf{S}^{(k)})^{-1}$. However, the MLE fails when $\mathbf{S}^{(k)}$ is singular. Furthermore, the MLE is usually dense. The ℓ_1 regularization has been employed to induce sparsity, resulting in the sparse inverse covariance estimation [2, 11, 53]. In this paper, we employ both the ℓ_1 regularization and the fused regularization for simultaneously estimating multiple graphs. The ℓ_1 regularization leads to a sparse solution, and the fused penalty encourages $\mathbf{\Theta}^{(k)}$ to be similar to its neighbors. Mathematically, we solve the following formulation:

$$\min_{\mathbf{\Theta}^{(k)} \succ 0, k=1,\ldots,K} \sum_{k=1}^{K} \left(-\log\det(\mathbf{\Theta}^{(k)}) + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}) \right) + P(\mathbf{\Theta}), \tag{2.2}$$

where

$$P(\mathbf{\Theta}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\mathbf{\Theta}_{ij}^{(k)}| + \lambda_2 \sum_{k=1}^{K-1} \sum_{i \neq j} |\mathbf{\Theta}_{ij}^{(k)} - \mathbf{\Theta}_{ij}^{(k+1)}|,$$

and $\lambda_1 > 0$ and $\lambda_2 > 0$ are positive regularization parameters. This model is referred to as the fused multiple graphical lasso (FMGL).

To ensure the existence of a solution for problem (2.2), we assume throughout this paper that $\mathrm{diag}(\mathbf{S}^{(k)}) > 0$, $k = 1, \ldots, K$. Recall that $\mathbf{S}^{(k)}$ is a sample covariance matrix, and hence $\mathrm{diag}(\mathbf{S}^{(k)}) \geq 0$. The diagonal entries may, however, not be strictly positive. But we can always add a small perturbation (say 10^{-8}) to ensure that the above assumption holds. The following theorem shows that under this assumption the FMGL (2.2) has a unique solution. A rigorous proof is given in the appendix.

THEOREM 2.1. *Under the assumption that* $\mathrm{diag}(\mathbf{S}^{(k)}) > 0$*,* $k = 1, \ldots, K$*, problem* (2.2) *has a unique optimal solution.*

3. The screening rule for FMGL.

Due to the presence of the log determinant, it is challenging to solve formulations involving the penalized log-likelihood efficiently. The existing methods for single graphical lasso are not scalable to problems with a large number of features because of the high computational complexity. Recent studies have shown that the graphical model may contain many connected components, which are disjoint from each other, due to the sparsity of the graphical model; i.e., the corresponding precision matrix has a block diagonal structure (subject to some rearrangement of features). To reduce the computational complexity, it is advantageous to first identify the block structure and then compute the diagonal blocks of the precision matrix instead of the whole matrix. Danaher, Wang, and Witten [8] developed a similar necessary and sufficient condition for fused graphical lasso with two graphs; thus the block structure can be identified. However, it remains a challenge to derive the necessary and sufficient condition for the solution of FMGL to be block diagonal for $K > 2$ graphs.

In this section, we first present a theorem demonstrating that FMGL can be decomposable once its solution has a block diagonal structure. Then we derive a necessary and sufficient condition for the solution of FMGL to be block diagonal for an arbitrary number of graphs.

Let $C_1, \ldots, C_L$ be a partition of the p features into L nonoverlapping sets, with $C_l \cap C_{l'} = \emptyset$ for all $l \neq l'$ and $\bigcup_{l=1}^{L} C_l = \{1, \ldots, p\}$. We say that the solution $\widehat{\mathbf{\Theta}}$ of FMGL (2.2) is block diagonal with L known blocks consisting of features in the sets C_l, $l = 1, \ldots, L$, if there exists a permutation matrix $\mathbf{U} \in \Re^{p\times p}$ such that each estimation precision matrix takes the form of

$$\widehat{\mathbf{\Theta}}^{(k)} = \mathbf{U} \begin{pmatrix} \widehat{\mathbf{\Theta}}_1^{(k)} & & \\ & \ddots & \\ & & \widehat{\mathbf{\Theta}}_L^{(k)} \end{pmatrix} \mathbf{U}^T, \qquad k = 1, \ldots, K. \tag{3.1}$$

For simplicity of presentation, we assume throughout this paper that $\mathbf{U} = \mathbf{I}$.

The following decomposition result for problem (2.2) is straightforward. Its proof is thus omitted.

THEOREM 3.1. *Suppose that the solution $\widehat{\mathbf{\Theta}}$ of FMGL* (2.2) *is block diagonal with L known C_l, $l = 1, \ldots, L$; i.e., each estimated precision matrix has the form* (3.1) *with $\mathbf{U} = \mathbf{I}$. Let $\widehat{\mathbf{\Theta}}_l = (\widehat{\mathbf{\Theta}}_l^{(1)}, \ldots, \widehat{\mathbf{\Theta}}_l^{(K)})$ for $l = 1, \ldots, L$. Then we have*

$$\widehat{\mathbf{\Theta}}_l = \arg\min_{\mathbf{\Theta}_l \succ 0} \sum_{k=1}^{K} \Big(-\log\det(\mathbf{\Theta}_l^{(k)}) + \mathrm{tr}(\mathbf{S}_l^{(k)} \mathbf{\Theta}_l^{(k)}) \Big) + P(\mathbf{\Theta}_l), \quad l = 1, \ldots, L, \tag{3.2}$$

where $\mathbf{\Theta}_l^{(k)}$ and $\mathbf{S}_l^{(k)}$ are the $|C_l| \times |C_l|$ symmetric submatrices of $\mathbf{\Theta}^{(k)}$ and $\mathbf{S}^{(k)}$, respectively, corresponding to the lth diagonal block, for $k = 1, \ldots, K$, and $\mathbf{\Theta}_l = (\mathbf{\Theta}_l^{(1)}, \ldots, \mathbf{\Theta}_l^{(K)})$ for $l = 1, \ldots, L$.

The above theorem demonstrates that if a large-scale FMGL problem has a block diagonal solution, it can then be decomposed into a group of smaller sized FMGL problems. The computational cost for the latter problems can be much cheaper. Now one natural question is how to efficiently identify the block diagonal structure of the FMGL solution before solving the problem. We address this question in the remaining part of this section.

The following theorem provides a necessary and sufficient condition for the solution of the FMGL to be block diagonal with L blocks C_l, $l = 1, \ldots, L$, which is a key for developing an efficient decomposition scheme for solving FMGL. Since its proof requires some substantial development of other technical results, we shall postpone the proof until the end of this section.

THEOREM 3.2. *The FMGL* (2.2) *has a block diagonal solution $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, with L known blocks C_l, $l = 1, \ldots, L$, if and only if $\mathbf{S}^{(k)}$, $k = 1, \ldots, K$, satisfy the following inequalities:*

$$\begin{cases} |\sum_{k=1}^{t} \mathbf{S}_{ij}^{(k)}| \le t\lambda_1 + \lambda_2, \\ |\sum_{k=0}^{t-1} \mathbf{S}_{ij}^{(r+k)}| \le t\lambda_1 + 2\lambda_2, \quad 2 \le r \le K - t, \\ |\sum_{k=1}^{t} \mathbf{S}_{ij}^{(K-t+k)}| \le t\lambda_1 + \lambda_2, \\ |\sum_{k=1}^{K} \mathbf{S}_{ij}^{(k)}| \le K\lambda_1 \end{cases} \tag{3.3}$$

for $t = 1, \ldots, K-1$, $i \in C_l, j \in C_{l'}, l \neq l'$.

One immediate consequence of Theorem 3.2 is that the conditions (3.3) can be used as a screening rule to identify the block diagonal structure of the FMGL solution. The steps for this rule are described as follows:

1. Construct an adjacency matrix $\mathbf{E} = \mathbf{I}_{p \times p}$. Set $\mathbf{E}_{ij} = \mathbf{E}_{ji} = 0$ if $\mathbf{S}_{ij}^{(k)}$, $k = 1, \ldots, K$, satisfy the conditions (3.3). Otherwise, set $\mathbf{E}_{ij} = \mathbf{E}_{ji} = 1$.
2. Identify the connected components of the adjacency matrix $\mathbf{E}$ (for example, it can be done by calling the MATLAB function "graphconncomp").

In view of Theorem 3.2, it is not hard to observe that the resulting connected components are the partition of the p features into nonoverlapping sets. It then follows from Theorem 3.1 that a large-scale FMGL problem can be decomposed into a group of smaller sized FMGL problems restricted to the features in each connected component. The computational cost for the latter problems can be much lower. Therefore, this approach may enable us to solve large-scale FMGL problems very efficiently.

In the remainder of this section we provide a proof for Theorem 3.2. Before proceeding, we establish several technical lemmas as follows.

LEMMA 3.3. *Given any two arbitrary index sets* $I \subseteq \{1, \ldots, n\}$ *and* $J \subseteq \{1, \ldots, n-1\}$, *let* $\bar{I}$ *and* $\bar{J}$ *be the complement of* I *and* J *with respect to* $\{1, \ldots, n\}$ *and* $\{1, \ldots, n-1\}$, *respectively. Define*

$$P_{I,J} = \left\{ y \in \Re^n : y_I \geq 0,\ y_{\bar{I}} \leq 0,\ y_J - y_{J+1} \geq 0,\ y_{\bar{J}} - y_{\bar{J}+1} \leq 0 \right\}, \tag{3.4}$$

where $J + 1 = \{j + 1 : j \in J\}$ *and* $\bar{J} + 1 = \{j + 1 : j \in \bar{J}\}$. *Then, the following statements hold:*

(i) *Either* $P_{I,J} = \{0\}$ *or* $P_{I,J}$ *is unbounded.*

(ii) 0 *is the unique extreme point of* $P_{I,J}$.

(iii) *Suppose that* $P_{I,J}$ *is unbounded. Then,* $\emptyset \neq \mathrm{ext}(P_{I,J}) \subseteq Q$, *where* $\mathrm{ext}(P_{I,J})$ *denotes the set of all extreme rays of* $P_{I,J}$ *and*

$$Q := \{\alpha(\underbrace{0, \ldots, 0}_{m}, \underbrace{1, \ldots, 1}_{l}, 0, \ldots, 0)^T \in \Re^n : \alpha \neq 0, m \geq 0, 1 \leq l \leq n\}. \tag{3.5}$$

Proof. (i) We observe that $0 \in P_{I,J}$. If $P_{I,J} \neq \{0\}$, then there exists $0 \neq y \in P_{I,J}$. Hence, $\{\alpha y : \alpha \geq 0\} \subseteq P_{I,J}$, which implies that $P_{I,J}$ is unbounded.

(ii) It is easy to see that $0 \in P_{I,J}$ and, moreover, that there exist n linearly independent active inequalities at 0. Hence, 0 is an extreme point of $P_{I,J}$. On the other hand, suppose y is an arbitrary extreme point of $P_{I,J}$. Then there exist n linearly independent active inequalities at y, which, together with the definition of $P_{I,J}$, immediately implies $y = 0$. Therefore, 0 is the unique extreme point of $P_{I,J}$.

(iii) Suppose that $P_{I,J}$ is unbounded. By statement (ii), we know that $P_{I,J}$ has a unique extreme point. Using Minkowski's resolution theorem (e.g., see [3]), we conclude that $\mathrm{ext}(P_{I,J}) \neq \emptyset$. Let $d \in \mathrm{ext}(P_{I,J})$ be arbitrarily chosen. Then $d \neq 0$. It follows from (3.4) that d satisfies the inequalities

$$d_I \geq 0, \quad d_{\bar{I}} \leq 0, \quad d_J - d_{J+1} \geq 0, \quad d_{\bar{J}} - d_{\bar{J}+1} \leq 0, \tag{3.6}$$

and moreover, the number of independent active inequalities at d is $n-1$. If all entries of d are nonzero, then d must satisfy $d_J - d_{J+1} = 0$ and $d_{\bar{J}} - d_{\bar{J}+1} = 0$ (with a total number $n-1$), which implies $d_1 = d_2 = \cdots = d_n$ and thus $d \in Q$. We now assume that d has at least one zero entry. Then, there exist positive integers k, $\{m_i\}_{i=1}^k$, and $\{n_i\}_{i=1}^k$ satisfying $m_i \leq n_i < m_{i+1} \leq n_{i+1}$ for $i = 1, \ldots, k-1$ such that

$$\{i : d_i = 0\} = \{m_1, \ldots, n_1\} \cup \{m_2, \ldots, n_2\} \cup \cdots \cup \{m_k, \ldots, n_k\}. \tag{3.7}$$

One can immediately observe that

$$d_{m_i} = \cdots = d_{n_i} = 0, \quad d_j - d_{j+1} = 0, \quad m_i \le j \le n_i - 1,\ 1 \le i \le k. \tag{3.8}$$

We next divide the rest of proof into four cases.

Case (a): $m_1 = 1$ and $n_k = n$. In view of (3.7), one can observe that $d_{m_i-1} - d_{m_i} \neq 0$ and $d_{n_{i-1}} - d_{n_{i-1}+1} \neq 0$ for $i = 2, \ldots, k$. We then see from (3.6) that, except for the active inequalities given in (3.8), all other possible active inequalities at d are

$$d_j - d_{j+1} = 0, \quad n_{i-1} < j < m_i - 1, \qquad 2 \le i \le k \tag{3.9}$$

(with a total number $\sum_{i=2}^{k}(m_i - n_{i-1} - 2)$). Notice that the total number of independent active inequalities given in (3.8) is $\sum_{i=1}^{k}(n_i - m_i + 1)$. Hence, the number of independent active inequalities at d is at most

$$\sum_{i=1}^{k}(n_i - m_i + 1) + \sum_{i=2}^{k}(m_i - n_{i-1} - 2) = n_k - m_1 - k + 2 = n - k + 1.$$

Recall that the number of independent active inequalities at d is $n-1$. Hence, we have $n-k+1 \ge n-1$, which implies $k \le 2$. Due to $d \neq 0$, we observe that $k \neq 1$ holds for this case. Also, we know that $k > 0$. Hence, $k = 2$. We then see that all possible active inequalities described in (3.9) must be active at d, which, together with $k = 2$, immediately implies that $d \in Q$.

Case (b): $m_1 = 1$ and $n_k < n$. Using (3.7), we observe that $d_{m_i-1} - d_{m_i} \neq 0$ for $i = 2, \ldots, k$ and $d_{n_i} - d_{n_i+1} \neq 0$ for $i = 1, \ldots, k$. In view of these relations and an argument similar to that in case (a), one can see that the number of independent active inequalities at d is at most

$$\sum_{i=1}^{k}(n_i - m_i + 1) + \sum_{i=2}^{k}(m_i - n_{i-1} - 2) + n - n_k - 1 = n - m_1 - k + 1 = n - k.$$

As in case (a), we can conclude from the above relation that $k = 1$ and $d \in Q$.

Case (c): $m_1 > 1$ and $n_k = n$. By (3.7), one can observe that $d_{m_i-1} - d_{m_i} \neq 0$ for $i = 1, \ldots, k$ and $d_{n_i} - d_{n_i+1} \neq 0$ for $i = 1, \ldots, k-1$. Using these relations and an argument similar to that in case (a), we see that the number of independent active inequalities at d is at most

$$m_1 - 2 + \sum_{i=1}^{k}(n_i - m_i + 1) + \sum_{i=2}^{k}(m_i - n_{i-1} - 2) = n_k - k = n - k.$$

As in case (a), we can conclude from the above relation that $k = 1$ and $d \in Q$.

Case (d): $m_1 > 1$ and $n_k < n$. From (3.7), one can observe that $d_{m_i-1} - d_{m_i} \neq 0$ for $i = 1, \ldots, k$ and $d_{n_i} - d_{n_i+1} \neq 0$ for $i = 1, \ldots, k$. By virtue of these relations and an argument similar to that in case (a), one can see that the number of independent active inequalities at d is at most

$$m_1 - 2 + \sum_{i=1}^{k}(n_i - m_i + 1) + \sum_{i=2}^{k}(m_i - n_{i-1} - 2) + n - n_k - 1 = n - k - 1.$$

Recall that $k \ge 1$ and the number of independent active inequalities at d is $n-1$. Hence, this case cannot occur.

Combining the above four cases, we conclude that $\mathrm{ext}(P_{I,J}) \subseteq Q$. ∎

LEMMA 3.4. *Let P_{IJ} and Q be defined in* (3.4) *and* (3.5), *respectively. Then,*

$$\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1,\dots,n\},\ J \subseteq \{1,\dots,n-1\}\} = Q.$$

Proof. It follows from Lemma 3.3(iii) that

$$\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1,\dots,n\},\ J \subseteq \{1,\dots,n-1\}\} \ \subseteq\ Q.$$

We next show that

$$\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1,\dots,n\},\ J \subseteq \{1,\dots,n-1\}\} \ \supseteq\ Q.$$

Indeed, let $d \in Q$ be arbitrarily chosen. Then, there exist $\alpha \neq 0$ and positive integers m_1 and n_1 satisfying $1 \leq m_1 \leq n_1$ such that $d_i = \alpha$ for $m_1 \leq i \leq n_1$ and the rest of the d_i's are 0. If $\alpha > 0$, it is not hard to see that $d \in \mathrm{ext}(P_{I,J})$ with $I = \{1,\dots,n\}$ and $J = \{m_1,\dots,n-1\}$. Similarly, if $\alpha < 0$, $d \in \mathrm{ext}(P_{I,J})$ with $I = \emptyset$ and J being the complement of $\bar{J} = \{m_1,\dots,n-1\}$. Hence, $d \in \cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1,\dots,n\},\ J \subseteq \{1,\dots,n-1\}\}$. ∎

LEMMA 3.5. *Let $x \in \Re^n$, λ_1, $\lambda_2 \geq 0$ be given, and let*

$$f(y) := x^T y - \lambda_1 \sum_{i=1}^{n} |y_i| - \lambda_2 \sum_{i=1}^{n-1} |y_i - y_{i+1}|.$$

Then, $f(y) \leq 0$ for all $y \in \Re^n$ if and only if x satisfies the following inequalities:

$$\begin{cases} |\sum_{j=1}^{k} x_j| \leq k\lambda_1 + \lambda_2, \\ |\sum_{j=0}^{k-1} x_{i+j}| \leq k\lambda_1 + 2\lambda_2, \quad 2 \leq i \leq n-k, \\ |\sum_{j=1}^{k} x_{n-k+j}| \leq k\lambda_1 + \lambda_2, \\ |\sum_{j=1}^{n} x_j| \leq n\lambda_1 \end{cases}$$

for $k = 1,\dots,n-1$.

Proof. Let $P_{I,J}$ be defined in (3.4) for any $I \subseteq \{1,\dots,n\}$ and $J \subseteq \{1,\dots,n-1\}$. We observe the following:

(a) $\Re^n = \cup\{P_{I,J} :\ I \subseteq \{1,\dots,n\},\ J \subseteq \{1,\dots,n-1\}\}$.

(b) $f(y) \leq 0$ for all $y \in \Re^n$ if and only if $f(y) \leq 0$ for all $y \in P_{I,J}$, and every $I \subseteq \{1,\dots,n\}$ and $J \subseteq \{1,\dots,n-1\}$.

(c) $f(y)$ is a linear function of y when restricted to the set $P_{I,J}$ for every $I \subseteq \{1,\dots,n\}$ and $J \subseteq \{1,\dots,n-1\}$.

If $P_{I,J}$ is bounded, we have $P_{I,J} = \{0\}$ and $f(y) = 0$ for $y \in P_{I,J}$. Suppose that $P_{I,J}$ is unbounded. By Lemma 3.3 and Minkowski's resolution theorem, $P_{I,J}$ equals the finitely generated cone by $\mathrm{ext}(P_{I,J})$. It then follows that $f(y) \leq 0$ for all $y \in P_{I,J}$ if and only if $f(d) \leq 0$ for all $d \in \mathrm{ext}(P_{I,J})$. Using these facts and Lemma 3.4, we see that $f(y) \leq 0$ for all $y \in \Re^n$ if and only if $f(d) \leq 0$ for all $d \in Q$, where Q is defined in (3.5). By the definitions of Q and f, we further observe that $f(y) \leq 0$ for all $y \in \Re^n$ if and only if $f(d) \leq 0$ for all

$$d \in \left\{ \pm(\underbrace{0,\dots,0}_{m},\underbrace{1,\dots,1}_{l},0,\dots,0)^T \in \Re^n : m \geq 0, 1 \leq l \leq n \right\},$$

which together with the definition of f immediately implies that the conclusion of this lemma holds. ∎

LEMMA 3.6. *Let $x \in \Re^n$, λ_1, $\lambda_2 \geq 0$ be given. The linear system*

$$\begin{cases} x_1 + \lambda_1\gamma_1 + \lambda_2 \upsilon_1 = 0, \\ x_i + \lambda_1\gamma_i + \lambda_2(\upsilon_i - \upsilon_{i-1}) = 0, \quad 2 \leq i \leq n-1, \\ x_n + \lambda_1\gamma_n - \lambda_2\upsilon_{n-1} = 0, \\ -1 \leq \gamma_i \leq 1, \quad i = 1, \ldots, n, \\ -1 \leq \upsilon_i \leq 1, \quad i = 1, \ldots, n-1, \end{cases} \tag{3.10}$$

has a solution (γ, υ) if and only if $(x, \lambda_1, \lambda_2)$ satisfies the following inequalities:

$$\begin{cases} |\sum_{j=1}^{k} x_j| \leq k\lambda_1 + \lambda_2, \\ |\sum_{j=0}^{k-1} x_{i+j}| \leq k\lambda_1 + 2\lambda_2, \quad 2 \leq i \leq n-k, \\ |\sum_{j=1}^{k} x_{n-k+j}| \leq k\lambda_1 + \lambda_2, \\ |\sum_{j=1}^{n} x_j| \leq n\lambda_1 \end{cases}$$

for $k = 1, \ldots, n-1$.

Proof. The linear system (3.10) has a solution if and only if the linear program

$$\min_{\gamma,\upsilon}\{0^T\gamma + 0^T\upsilon : (\gamma, \upsilon) \text{ satisfies } (3.10)\} \tag{3.11}$$

has an optimal solution. The Lagrangian dual of (3.11) is

$$\max_{y} \min_{\gamma,\upsilon} \left\{ x^T y + \lambda_1 \sum_{i=1}^{n} y_i\gamma_i + \lambda_2 \sum_{i=1}^{n-1} (y_i - y_{i+1})\upsilon_i : \ -1 \leq \gamma, \upsilon \leq 1 \right\},$$

which is equivalent to

$$\max_{y} f(y) := x^T y - \lambda_1 \sum_{i=1}^{n} |y_i| - \lambda_2 \sum_{i=1}^{n-1} |y_i - y_{i+1}|. \tag{3.12}$$

By the Lagrangian duality theory, problem (3.11) has an optimal solution if and only if its dual problem (3.12) has optimal value 0, which is equivalent to $f(y) \leq 0$ for all $y \in \Re^n$. The conclusion of this lemma then immediately follows from Lemma 3.5. ∎

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. For the sake of convenience, we denote the inverse of $\widehat{\mathbf{\Theta}}^{(k)}$ as $\widehat{\mathbf{W}}^{(k)}$ for $k = 1, \ldots, K$. By the first-order optimality conditions, we observe that $\widehat{\mathbf{\Theta}}^{(k)} \succ 0$, $k = 1, \ldots, K$, is the optimal solution of problem (2.2) if and only if it satisfies

$$-\widehat{\mathbf{W}}_{ii}^{(k)} + \mathbf{S}_{ii}^{(k)} = 0, \quad 1 \leq k \leq K, \tag{3.13}$$

$$-\widehat{\mathbf{W}}_{ij}^{(1)} + \mathbf{S}_{ij}^{(1)} + \lambda_1\gamma_{ij}^{(1)} + \lambda_2\upsilon_{ij}^{(1,2)} = 0, \tag{3.14}$$

$$-\widehat{\mathbf{W}}_{ij}^{(k)} + \mathbf{S}_{ij}^{(k)} + \lambda_1\gamma_{ij}^{(k)} + \lambda_2(-\upsilon_{ij}^{(k-1,k)} + \upsilon_{ij}^{(k,k+1)}) = 0, \quad 2 \leq k \leq K-1, \tag{3.15}$$

$$-\widehat{\mathbf{W}}_{ij}^{(K)} + \mathbf{S}_{ij}^{(K)} + \lambda_1\gamma_{ij}^{(K)} - \lambda_2\upsilon_{ij}^{(K-1,K)} = 0 \tag{3.16}$$

for all $i, j = 1, \ldots, p,\ i \neq j$, where $\gamma_{ij}^{(k)}$ is a subgradient of $|\mathbf{\Theta}_{ij}^{(k)}|$ at $\mathbf{\Theta}_{ij}^{(k)} = \widehat{\mathbf{\Theta}}_{ij}^{(k)}$ and $\upsilon_{ij}^{(k,k+1)}$ is a subgradient of $|\mathbf{\Theta}_{ij}^{(k)} - \mathbf{\Theta}_{ij}^{(k+1)}|$ with respect to $\mathbf{\Theta}_{ij}^{(k)}$ at $(\mathbf{\Theta}_{ij}^{(k)}, \mathbf{\Theta}_{ij}^{(k+1)}) = (\widehat{\mathbf{\Theta}}_{ij}^{(k)}, \widehat{\mathbf{\Theta}}_{ij}^{(k+1)})$; that is, $\upsilon_{ij}^{(k,k+1)} = 1$ if $\widehat{\mathbf{\Theta}}_{ij}^{(k)} > \widehat{\mathbf{\Theta}}_{ij}^{(k+1)}$, $\upsilon_{ij}^{(k,k+1)} = -1$ if $\widehat{\mathbf{\Theta}}_{ij}^{(k)} < \widehat{\mathbf{\Theta}}_{ij}^{(k+1)}$, and $\upsilon_{ij}^{(k,k+1)} \in [-1, 1]$ if $\widehat{\mathbf{\Theta}}_{ij}^{(k)} = \widehat{\mathbf{\Theta}}_{ij}^{(k+1)}$.

Necessity. Suppose that $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, is a block diagonal optimal solution of problem (2.2) with L known blocks C_l, $l = 1, \ldots, L$. Note that $\widehat{\mathbf{W}}^{(k)}$ has the same block diagonal structure as $\widehat{\mathbf{\Theta}}^{(k)}$. Hence, $\widehat{\mathbf{W}}_{ij}^{(k)} = \widehat{\mathbf{\Theta}}_{ij}^{(k)} = 0$ for $i \in C_l$, $j \in C_{l'}$, $l \neq l'$. This together with (3.14)–(3.16) implies that for each $i \in C_l$, $j \in C_{l'}$, $l \neq l'$, there exist $(\gamma_{ij}^{(k)}, \upsilon_{ij}^{(k,k+1)})$, $k = 1, \ldots, K-1$, and $\gamma_{ij}^{(K)}$ such that

$$\begin{aligned}
&\mathbf{S}_{ij}^{(1)} + \lambda_1 \gamma_{ij}^{(1)} + \lambda_2 \upsilon_{ij}^{(1,2)} = 0, \\
&\mathbf{S}_{ij}^{(k)} + \lambda_1 \gamma_{ij}^{(k)} + \lambda_2 (-\upsilon_{ij}^{(k-1,k)} + \upsilon_{ij}^{(k,k+1)}) = 0, \quad 2 \leq k \leq K-1, \\
&\mathbf{S}_{ij}^{(K)} + \lambda_1 \gamma_{ij}^{(K)} - \lambda_2 \upsilon_{ij}^{(K-1,K)} = 0, \\
&-1 \leq \gamma_{ij}^{(k)} \leq 1,\ 1 \leq k \leq K, \\
&-1 \leq \upsilon_{ij}^{(k,k+1)} \leq 1, \quad 1 \leq k \leq K-1.
\end{aligned} \tag{3.17}$$

Using (3.17) and Lemma 3.6, we see that (3.3) holds for $t = 1, \ldots, K-1$, $i \in C_l$, $j \in C_{l'}$, $l \neq l'$.

Sufficiency. Suppose that (3.3) holds for $t = 1, \ldots, K-1$, $i \in C_l$, $j \in C_{l'}$, $l \neq l'$. It then follows from Lemma 3.6 that for each $i \in C_l$, $j \in C_{l'}$, $l \neq l'$ there exist $(\gamma_{ij}^{(k)}, \upsilon_{ij}^{(k,k+1)})$, $k = 1, \ldots, K-1$, and $\gamma_{ij}^{(K)}$ such that (3.17) holds. Now let $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, be a block diagonal matrix as defined in (3.1) with $\mathbf{U} = \mathbf{I}$, where $\widehat{\mathbf{\Theta}}_l = (\widehat{\mathbf{\Theta}}_l^{(1)}, \ldots, \widehat{\mathbf{\Theta}}_l^{(K)})$ is given by (3.2) for $l = 1, \ldots, L$. Also, let $\widehat{\mathbf{W}}^{(k)}$ be the inverse of $\widehat{\mathbf{\Theta}}^{(k)}$ for $k = 1, \ldots, K$. Since $\widehat{\mathbf{\Theta}}_l$ is the optimal solution of problem (3.2), the first-order optimality conditions imply that (3.13)–(3.16) hold for all $i, j \in C_l$, $i \neq j$, $l = 1, \ldots, L$. Notice that $\widehat{\mathbf{\Theta}}_{ij}^{(k)} = \widehat{\mathbf{W}}_{ij}^{(k)} = 0$ for every $i \in C_l$, $j \in C_{l'}$, $l \neq l'$. Using this fact and (3.17), we observe that (3.13)–(3.16) also hold for all $i \in C_l$, $j \in C_{l'}$, $l \neq l'$. It then follows that $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, is an optimal solution of problem (2.2). In addition, $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, is block diagonal with L known blocks C_l, $l = 1, \ldots, L$. The conclusion thus holds. ∎

3.1. Extension to other regularizations. We show how to establish a similar necessary and sufficient condition for general fused regularization (i.e., graph fused regularization). Denote $G = (V, E)$ as an undirected graph, where the nodes are $V = \{1, \ldots, K\}$ and E is a set of edges. Assume that there is no redundancy in E (i.e., if $(u, v) \in E$, $(v, u) \notin E$). Then we define the graph fused regularization by

$$P(\mathbf{\Theta}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\mathbf{\Theta}_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \sum_{(u,v) \in E} |\mathbf{\Theta}_{ij}^{(u)} - \mathbf{\Theta}_{ij}^{(v)}|. \tag{3.18}$$

Clearly, the sequential fused and pairwise fused regularization are special cases of the graph fused regularization. The graph fused regularization is decomposable based on the connected components of the given graph G. Without loss of generality, we assume that G has only *one connected component*, which means that there exists an edge across any two-set partition of V. The technique used in the sequential fused case can be extended to the case of graph fused regularization. The key is to prove results similar to those in Lemmas 3.3 and 3.4 for graph fused regularization.

Denote $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$ as the set of subgraphs in graph G such that each subgraph G_m has only one connected component. For example, a fully connected graph with 3 nodes has 7 such subgraphs. According to the assumption that G has only one connected component, we have $G \in \mathcal{G}$. Let $\mathcal{V} = \{V_1, V_2, \ldots, V_M\}$, where V_m represents the nodes of subgraph G_m. Then we have the following results.

LEMMA 3.7. *Given an undirected graph $G = (V, E)$, where the nodes are $V = \{1, \ldots, n\}$ and E is a set of edges of size $|E|$. Given any two arbitrary index sets $I \subseteq \{1, \ldots, n\}$, $J \subseteq \{1, \ldots, |E|\}$, let $\bar{I}$ and $\bar{J}$ be the complement of I and J with respect to $\{1, \ldots, n\}$ and $\{1, \ldots, |E|\}$, respectively. Define*

$$P_{I,J} = \{y \in \Re^n : y_I \geq 0,\ y_{\bar{I}} \leq 0,\ y_u - y_v \geq 0\ \forall (u,v) \in E_J,\\ y_u - y_v \leq 0\ \forall (u,v) \in E_{\bar{J}}\}, \tag{3.19}$$

where E_J and $E_{\bar{J}}$ denote the sets of edges whose indexes are in J and $\bar{J}$, respectively. Then, the following statements hold:

(i) *Either $P_{I,J} = \{0\}$ or $P_{I,J}$ is unbounded.*

(ii) 0 *is the unique extreme point of $P_{I,J}$.*

(iii) *Suppose that $P_{I,J}$ is unbounded. Then, $\emptyset \neq \operatorname{ext}(P_{I,J}) \subseteq Q$, where*

$$Q := \left\{ \alpha d \in \Re^n : \alpha \neq 0, d_i = \begin{cases} 1, & i \in V_m, \\ 0, & i \notin V_m, \end{cases} \quad \forall V_m \in \mathcal{V} \right\}. \tag{3.20}$$

(iv) $\cup\{\operatorname{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\} = Q$.

The proof is given in the appendix. After we obtain the set of all extreme rays, the remaining steps can be proved in the same manner as in the fused case. Let $|E_{\backslash V_m}|$ be the number of edges across V_m and its complement, and let $|V_m|$ be the number of nodes in V_m. Then the necessary and sufficient condition for graph fused regularization is

$$\left| \sum_{k=1}^{|V_m|} \mathbf{S}_{ij}^{(u_k)} \right| \leq |V_m|\lambda_1 + |E_{\backslash V_m}|\lambda_2, \quad u_k \in V_m,\ \forall V_m \in \mathcal{V}. \tag{3.21}$$

The complexity of verifying the necessary and sufficient condition for an arbitrary graph is exponential due to all possible subgraphs with only one connected component. Exploring the structure of the given graph may reduce redundancy of the conditions (3.21). We defer this to future work.

3.2. Screening rule for general structured multiple graphical lasso (SMGL). We consider the following general SMGL:

$$\min_{\mathbf{\Theta}^{(k)} \succ 0, k=1,\ldots,K} \sum_{k=1}^{K} \left(-\log\det(\mathbf{\Theta}^{(k)}) + \operatorname{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}) \right) + \sum_{i \neq j} \phi(\mathbf{\Theta}_{ij}), \tag{3.22}$$

where $\mathbf{\Theta}_{ij} = (\mathbf{\Theta}_{ij}^{(1)}, \ldots, \mathbf{\Theta}_{ij}^{(K)})^T \in \Re^K$ and $\phi(x)$ is a convex regularization that encourages estimated graph models to have a certain structure. Besides fused and graph regularizations, there are other examples including but not limited to the following:

- Overlapping group regularization:

$$\phi(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i=1}^{g} \|x_{G_i}\|_2,$$

where G_i, $i = 1, \dots, g$, are g groups such that $\bigcup_{i=1}^{g} G_i = \{1, \dots, K\}$. Different groups may overlap.

- Tree structured group regularization:

$$\phi(x) = \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_j^i \|x_{G_j^i}\|_2,$$

where w_j^i is a positive weight and the groups G_j^i, $j = 1, \dots, n_i$, $i = 1, \dots, d$, exhibit a tree structure [26].

THEOREM 3.8. *The SMGL* (3.22) *has a block diagonal solution* $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \dots, K$, *with* L *blocks* C_l, $l = 1, \dots, L$, *if and only if* 0 *is the optimal solution of the following problem:*

$$\min_{x} \frac{1}{2} \|x + \mathbf{S}_{ij}\|_2^2 + \phi(x) \tag{3.23}$$

for $i \in C_l$, $j \in C_{l'}$, $l \neq l'$.

The proof can be found in the appendix. Theorem 3.8 can be used as a screening rule for the SMGL. If (3.23) has a closed form solution as in the case of tree structured group regularization [26], the screening rule results in an exact block diagonal structure. However, if (3.23) does not have a closed form solution, the screening rule may not identify an exact block diagonal structure due to numerical error. Although the identified structure may be inexact, it can still be used to find a good initial solution, as shown in [16]. An interesting future direction is to study the error bound between the identified and exact block diagonal structures.

4. Second-order method. The screening rule proposed in section 3 is capable of partitioning all features into a group of smaller sized blocks. Accordingly, a large-scale FMGL (2.2) can be decomposed into a number of smaller sized FMGL problems. For each block l we need to compute its individual precision matrix $\mathbf{\Theta}_l^{(k)}$ by solving the FMGL (2.2) with $\mathbf{S}^{(k)}$ replaced by $\mathbf{S}_l^{(k)}$. In this section, we show how to solve those single block FMGL problems efficiently. For simplicity of presentation, we assume throughout this section that the FMGL (2.2) has only one block; that is, $L = 1$.

We now propose a second-order method to solve the FMGL (2.2). For simplicity of notation, we let $\mathbf{\Theta} := (\mathbf{\Theta}^{(1)}, \dots, \mathbf{\Theta}^{(K)})$ and use t to denote the Newton iteration index. Let $\mathbf{\Theta}_t = (\mathbf{\Theta}_t^{(1)}, \dots, \mathbf{\Theta}_t^{(K)})$ be the approximate solution obtained at the tth Newton iteration.

The optimization problem (2.2) can be rewritten as

$$\min_{\mathbf{\Theta} \succ 0} F(\mathbf{\Theta}) := \sum_{k=1}^{K} f_k(\mathbf{\Theta}^{(k)}) + P(\mathbf{\Theta}), \tag{4.1}$$

where

$$f_k(\mathbf{\Theta}^{(k)}) = -\log\det(\mathbf{\Theta}^{(k)}) + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}).$$

In the second-order method, we approximate the objective function $F(\mathbf{\Theta})$ at the current iterate $\mathbf{\Theta}_t$ by a "quadratic" model $Q_t(\mathbf{\Theta})$:

$$\min_{\mathbf{\Theta}} Q_t(\mathbf{\Theta}) := \sum_{k=1}^{K} q_k(\mathbf{\Theta}^{(k)}) + P(\mathbf{\Theta}), \tag{4.2}$$

where q_k is the quadratic approximation of f_k at $\mathbf{\Theta}_t^{(k)}$; that is,

$$q_k(\mathbf{\Theta}^{(k)}) = \frac{1}{2}\mathrm{tr}(\mathbf{W}_t^{(k)}\mathbf{D}^{(k)}\mathbf{W}_t^{(k)}\mathbf{D}^{(k)}) + \mathrm{tr}((\mathbf{S}^{(k)} - \mathbf{W}_t^{(k)})\mathbf{D}^{(k)}) + f_k(\mathbf{\Theta}_t^{(k)})$$

with $\mathbf{W}_t^{(k)} = (\mathbf{\Theta}_t^{(k)})^{-1}$ and $\mathbf{D}^{(k)} = \mathbf{\Theta}^{(k)} - \mathbf{\Theta}_t^{(k)}$. Suppose that $\bar{\mathbf{\Theta}}_{t+1}$ is the optimal solution of (4.2). Then we obtain the Newton search direction

$$\mathbf{D} = \bar{\mathbf{\Theta}}_{t+1} - \mathbf{\Theta}_t. \tag{4.3}$$

We shall mention that the subproblem (4.2) can be suitably solved by the nonmonotone spectral projected gradient (NSPG) method (see, for example, [30, 50]). It was shown by Lu and Zhang [30] that the NSPG method is locally linearly convergent. Numerous computational studies have demonstrated that the NSPG method is very efficient, though its global convergence rate is so far unknown. When applied to (4.2), the NSPG method requires solving proximal subproblems in the form of

$$prox_{\alpha P}(\mathbf{Z}_r) := \arg\min_{\mathbf{\Theta}} \frac{1}{2}\|\mathbf{\Theta} - \mathbf{Z}_r\|_F^2 + \alpha P(\mathbf{\Theta}), \tag{4.4}$$

where r represents the rth iteration in NSPG, $\|\mathbf{\Theta} - \mathbf{Z}_r\|_F^2 = \sum_{k=1}^K \|\mathbf{\Theta}^{(k)} - \mathbf{Z}_r^{(k)}\|_F^2$, $\mathbf{Z}_r = \mathbf{\Theta}_r - \alpha\mathbf{G}_r$, and $\mathbf{G}_r^{(k)} = \mathbf{S}^{(k)} - 2\mathbf{W}_t^{(k)} + \mathbf{W}_t^{(k)}\mathbf{\Theta}_r^{(k)}\mathbf{W}_t^{(k)}$. Denote $\mathbf{R} = \mathbf{\Theta}_r - \mathbf{\Theta}_{r-1}$ and $\bar{\alpha} = \sum_{k=1}^K \mathrm{tr}(\mathbf{R}^{(k)}\mathbf{W}_t^{(k)}\mathbf{R}^{(k)}\mathbf{W}_t^{(k)}) / \sum_{k=1}^K \|\mathbf{R}^{(k)}\|_F^2$. Then α is given by $\alpha = \max(\alpha_{min}, \min(1/\bar{\alpha}, \alpha_{max}))$, where $[\alpha_{min}, \alpha_{max}]$ is a given safeguard [30, 50].

By the definition of $P(\mathbf{\Theta})$, it is not hard to see that problem (4.4) can be decomposed into a set of independent and smaller sized problems,

$$\min_{\mathbf{\Theta}_{ij}^{(k)}, k=1,\ldots,K} \frac{1}{2}\sum_{k=1}^K (\mathbf{\Theta}_{ij}^{(k)} - \mathbf{Z}_{r,ij}^{(k)})^2 + \alpha_1 \sum_{k=1}^K |\mathbf{\Theta}_{ij}^{(k)}| + \alpha_2 \sum_{k=1}^{K-1} |\mathbf{\Theta}_{ij}^{(k)} - \mathbf{\Theta}_{ij}^{(k+1)}| \tag{4.5}$$

for all $i > j$, $(\alpha_1, \alpha_2) = \alpha(\lambda_1, \lambda_2)$, and for $i = j$, $\alpha_1, \alpha_2 = 0$, $j = 1, \ldots, p$. Problem (4.5) is known as the fused lasso signal approximator, which can be solved very efficiently and exactly [7, 27]. In addition, these smaller problems are independent from each other and thus can be solved in parallel.

Given the current search direction $\mathbf{D} = (\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(K)})$ that is computed above, we need to find a suitable step length $\beta \in (0, 1]$ to ensure a sufficient reduction in the objective function of (2.2) and positive definiteness of the next iterate $\mathbf{\Theta}_{t+1}^{(k)} = \mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)}$, $k = 1, \ldots, K$. In the context of the standard (single) graphical lasso, Hsieh et al. [17] have shown that a step length satisfying the above requirements always exists. We can similarly prove that the desired step length also exists for the FMGL (2.2) (the poof is similar to that in [17] and is thus omitted).

LEMMA 4.1. *Let $\mathbf{\Theta}_t = (\mathbf{\Theta}_t^{(1)}, \ldots, \mathbf{\Theta}_t^{(K)})$ be such that $\mathbf{\Theta}_t^{(k)} \succ 0$ for $k = 1, \ldots, K$, and let $\mathbf{D} = (\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(K)})$ be the associated Newton search direction computed according to* (4.2). *Suppose $\mathbf{D} \neq 0$.*[1] *Then there exists a $\bar{\beta} > 0$ such that $\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)} \succ 0$ and the sufficient reduction condition*

$$F(\mathbf{\Theta}_t + \beta\mathbf{D}) \leq F(\mathbf{\Theta}_t) + \sigma\beta\delta_t \tag{4.6}$$

[1] It is well known that if $\mathbf{D} = 0$, then $\mathbf{\Theta}_t$ is the optimal solution of problem (2.2).

holds for all $0 < \beta < \bar{\beta}$, *where* $\sigma \in (0, 1/2)$ *is a given constant and*

$$\delta_t = \sum_{k=1}^{K} \mathrm{tr}((\mathbf{S}^{(k)} - \mathbf{W}_t^{(k)})\mathbf{D}^{(k)}) + P(\mathbf{\Theta}_t + \mathbf{D}) - P(\mathbf{\Theta}_t). \tag{4.7}$$

By virtue of Lemma 4.1, we can adopt the well-known Armijo backtracking line search rule [44] to select a step length $\beta \in (0, 1]$ so that $\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)} \succ 0$ and (4.6) holds. In particular, we choose β to be the largest number of the sequence $\{1, 1/2, \ldots, 1/2^i, \ldots\}$ that satisfies these requirements. We can use the Cholesky factorization to check the positive definiteness of $\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)}$, $k = 1, \ldots, K$. In addition, the associated terms $\log\det(\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)})$ and $(\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)})^{-1}$ can be efficiently computed as a byproduct of the Cholesky decomposition of $\mathbf{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)}$.

4.1. Active set identification. Given the large number of unknown variables in (4.2), it is advantageous to minimize (4.2) in a reduced space. In the case of a single graph ($K = 1$), problem (4.2) degenerates to a lasso problem of size p^2. Hsieh et al. [17] proposed a strategy to determine a subset of variables that are allowed to be updated in each Newton iteration for single graphical lasso. Specifically, the p^2 variables in single graphical lasso are partitioned into two sets, including a free set $\mathcal{F}$ and an active set $\mathcal{A}$, based on the gradient at the start of each Newton iteration, and then the minimization is performed only on the variables in $\mathcal{F}$. We call this technique "active set identification" in this paper. Due to the sparsity of the precision matrix, the size of $\mathcal{F}$ is usually much smaller than p^2. Moreover, it has been shown in the single graph case that the size of $\mathcal{F}$ will decrease quickly [17]. The active set identification can thus improve the computational efficiency. This technique was also successfully used in [20, 37, 38, 52]. We show that active set identification can be extended to the FMGL based on the results established in section 3.

Denote the gradient of f_k at the tth iteration by $\widetilde{\mathbf{G}}_t^{(k)} = \mathbf{S}^{(k)} - \mathbf{W}_t^{(k)}$, and its (i, j)th element by $\widetilde{\mathbf{G}}_{t,ij}^{(k)}$. Then we have the following result.

Lemma 4.2. *For* $\mathbf{\Theta}_t$ *in the* t*th iteration, define the active set* $\mathcal{A}$ *as*

$$\mathcal{A} = \{(i, j) | \mathbf{\Theta}_{t,ij}^{(1)} = \cdots = \mathbf{\Theta}_{t,ij}^{(K)} = 0 \text{ and } \widetilde{\mathbf{G}}_{t,ij}^{(1)}, \ldots, \widetilde{\mathbf{G}}_{t,ij}^{(K)} \text{ satisfy the inequalities below}\}:$$

$$\begin{cases} |\sum_{k=1}^{u} \widetilde{\mathbf{G}}_{t,ij}^{(k)}| < u\lambda_1 + \lambda_2, \\ |\sum_{k=0}^{u-1} \widetilde{\mathbf{G}}_{t,ij}^{(r+k)}| < u\lambda_1 + 2\lambda_2, \quad 2 \le r \le K - u, \\ |\sum_{k=1}^{u} \widetilde{\mathbf{G}}_{t,ij}^{(K-u+k)}| < u\lambda_1 + \lambda_2, \\ |\sum_{k=1}^{K} \widetilde{\mathbf{G}}_{t,ij}^{(k)}| < K\lambda_1 \end{cases} \tag{4.8}$$

for $u = 1, \ldots, K - 1$.

Then, the solution of the following problem is $\mathbf{D}^{(1)} = \cdots = \mathbf{D}^{(K)} = 0$:

$$\min_{\mathbf{D}} Q_t(\mathbf{\Theta}_t + \mathbf{D}) \quad \text{such that } \mathbf{D}_{ij}^{(1)} = \cdots = \mathbf{D}_{ij}^{(K)} = 0,\ (i, j) \notin \mathcal{A}. \tag{4.9}$$

Proof. Consider problem (4.9), which can be reformulated to

$$\begin{aligned} \min_{\mathbf{D}} \ & \sum_{k=1}^{K} \left(\frac{1}{2}\mathrm{vec}(\mathbf{D}^{(k)})^T \mathbf{H}_t^{(k)} \mathrm{vec}(\mathbf{D}^{(k)}) + \mathrm{vec}(\widetilde{\mathbf{G}}_t^{(k)})^T \mathrm{vec}(\mathbf{D}^{(k)}) \right) \\ & \quad + P(\mathbf{\Theta}_t + \mathbf{D}) \\ \text{s.t.} \ & \mathbf{D}_{ij}^{(1)} = \cdots = \mathbf{D}_{ij}^{(K)} = 0, \quad (i, j) \notin \mathcal{A}, \end{aligned} \tag{4.10}$$

where $\mathbf{H}_t^{(k)} = \mathbf{W}_t^{(k)} \otimes \mathbf{W}_t^{(k)}$. Because of the constraint $\mathbf{D}_{ij}^{(1)} = \cdots = \mathbf{D}_{ij}^{(K)} = 0$, $(i,j) \notin \mathcal{A}$, we consider only the variables in the set $\mathcal{A}$. According to Lemma 3.6, it is easy to see that $\mathbf{D}_\mathcal{A} = 0$ satisfies the optimality condition of the following problem:

$$\min_{\mathbf{D}_\mathcal{A}} \sum_{k=1}^{K} \mathrm{vec}(\widetilde{\mathbf{G}}_{t,\mathcal{A}}^{(k)})^T \mathrm{vec}(\mathbf{D}_\mathcal{A}^{(k)}) + P(\mathbf{D}_\mathcal{A}).$$

Since $\sum_{k=1}^{K} \mathrm{vec}(\mathbf{D}^{(k)})^T \mathbf{H}_t^{(k)} \mathrm{vec}(\mathbf{D}^{(k)}) \geq 0$, the optimal solution of (4.9) is given by $\mathbf{D}^{(1)} = \cdots = \mathbf{D}^{(K)} = 0$. ∎

Lemma 4.2 provides an active set identification scheme to partition the variables into the free set $\mathcal{F}$ and the active set $\mathcal{A}$. Lemma 4.2 shows that when the variables in the free set $\mathcal{F}$ are fixed, no update is needed for the variables in the active set $\mathcal{A}$. The resulting second-order method with active set identification for solving the FMGL is summarized in Algorithm 1.

Algorithm 1: Proposed Second-Order Method for FMGL.

Input: $\mathbf{S}^{(k)}$, $k = 1, \ldots, K, \lambda_1, \lambda_2$
Output: $\mathbf{\Theta}^{(k)}$, $k = 1, \ldots, K$
Initialization: $\mathbf{\Theta}_0^{(k)} = (\mathrm{Diag}(\mathbf{S}^{(k)}))^{-1}$;
while *Not Converged* **do**
  Determine the sets of free and fixed indices $\mathcal{F}$ and $\mathcal{A}$ using Lemma 4.2.
  Compute the Newton direction $\mathbf{D}^{(k)}$, $k = 1, \ldots, K$, by solving (4.2) and (4.3) over the free variables $\mathcal{F}$.
  Choose $\mathbf{\Theta}_{t+1}^{(k)}$ by performing the Armijo backtracking line search along $\mathbf{\Theta}_t^{(k)} + \beta \mathbf{D}^{(k)}$ for $k = 1, \ldots, K$.
end
return $\mathbf{\Theta}^{(k)}$, $k = 1, \ldots, K$;

4.2. Convergence. Convergence of proximal Newton-type methods has been studied in previous literature [5, 17, 23, 41, 44]. Under the assumption that the subproblems are solved exactly, a local quadratic convergence rate can be achieved when the exact Hessian is used (i.e., the proximal Newton method) [17, 23, 44]. When an approximate Hessian is used (i.e., the proximal quasi-Newton method), the local convergence rate is linear or superlinear [23, 44]. We show that the FMGL algorithm (with active set identification) falls into the proximal quasi-Newton framework. Denote the approximate Hessian by

$$\widetilde{\mathbf{H}}_t^{(k)} = \begin{pmatrix} \mathbf{H}_{t,\mathcal{F}}^{(k)} & \\ & \mathbf{H}_{t,\mathcal{A}}^{(k)} \end{pmatrix}, \tag{4.11}$$

where $\mathbf{H}_{t,\mathcal{J}}^{(k)}$ is the submatrix of the exact Hessian $\mathbf{H}_t^{(k)}$ with variables in $\mathcal{J}$. Using $\widetilde{\mathbf{H}}_t^{(k)}$ instead, the subproblem (4.2) can be decomposed into the following two problems:

$$\begin{aligned} \min_{\mathbf{D}_\mathcal{J}} \sum_{k=1}^{K} \Bigg(& \frac{1}{2} \mathrm{vec}(\mathbf{D}_\mathcal{J}^{(k)})^T \mathbf{H}_{t,\mathcal{J}}^{(k)} \mathrm{vec}(\mathbf{D}_\mathcal{J}^{(k)}) + \mathrm{vec}(\widetilde{\mathbf{G}}_{t,\mathcal{J}}^{(k)})^T \mathrm{vec}(\mathbf{D}_\mathcal{J}^{(k)}) \Bigg) \\ & + P(\mathbf{\Theta}_{t,\mathcal{J}} + \mathbf{D}_\mathcal{J}), \qquad \mathcal{J} = \mathcal{F},\ \mathcal{A}. \end{aligned} \tag{4.12}$$

Consider the problem with respect to the variables in $\mathcal{A}$:

$$\min_{\mathbf{D}_\mathcal{A}} \sum_{k=1}^{K} \left(\frac{1}{2}\mathrm{vec}(\mathbf{D}_\mathcal{A}^{(k)})^T \mathbf{H}_{t,\mathcal{A}}^{(k)} \mathrm{vec}(\mathbf{D}_\mathcal{A}^{(k)}) + \mathrm{vec}(\widetilde{\mathbf{G}}_{t,\mathcal{A}}^{(k)})^T \mathrm{vec}(\mathbf{D}_\mathcal{A}^{(k)}) \right) + P(\mathbf{\Theta}_{t,\mathcal{A}} + \mathbf{D}_\mathcal{A}),$$

which is equivalent to problem (4.10). According to the definition of the active set $\mathcal{A}$, it follows from Lemma 4.2 that the optimal solution is $\mathbf{D}_\mathcal{A}^{(k)} = 0,\ k = 1, \ldots, K$. Thus, the FMGL in Algorithm 1 is a proximal quasi-Newton method. The global convergence to the unique optimal solution is therefore guaranteed [23].

In the case when the subproblems are solved inexactly (i.e., *inexact* FMGL), we can adopt the following adaptive stopping criterion proposed in [5, 23] to achieve the global convergence:

$$\|M_{\tau\bar{q}}(\bar{\mathbf{\Theta}})\| \le \eta_t \|M_{\tau\bar{f}}(\mathbf{\Theta}_t)\|, \quad Q_t^H(\bar{\mathbf{\Theta}}) - Q_t^H(\mathbf{\Theta}_t) \le \zeta(L_t(\bar{\mathbf{\Theta}}) - L_t(\mathbf{\Theta}_t)), \tag{4.13}$$

for some $\tau > 0$, where $\bar{\mathbf{\Theta}}$ is an *inexact* solution of the subproblem, $\eta_t \in (0, 1)$ is a forcing term, $\zeta \in (\sigma, 1/2)$, $L_t(\mathbf{\Theta})$ is defined by

$$L_t(\mathbf{\Theta}) = \bar{f}(\mathbf{\Theta}_t) + \mathrm{vec}(\nabla \bar{f}(\mathbf{\Theta}))^T \mathrm{vec}(\mathbf{\Theta} - \mathbf{\Theta}_t) + P(\mathbf{\Theta}),$$

and the composite gradient step $M_{\tau\bar{f}}(\mathbf{\Theta})$ is defined by

$$M_{\tau\bar{f}}(\mathbf{\Theta}) = \frac{1}{\tau}\left(\mathbf{\Theta} - prox_{\tau P}(\mathbf{\Theta} - \tau\nabla\bar{f}(\mathbf{\Theta}))\right).$$

The functions $\bar{q}(\mathbf{\Theta})$ and $\bar{f}(\mathbf{\Theta})$ are defined by

$$\bar{q}(\mathbf{\Theta}) = \sum_{k=1}^{K} q_k^H(\mathbf{\Theta}^{(k)}), \qquad \bar{f}(\mathbf{\Theta}) = \sum_{k=1}^{K} f_k(\mathbf{\Theta}^{(k)}).$$

The superscript in Q_t^H and q_k^H represents the "quadratic" approximate functions Q_t and q_k using the approximate Hessian in (4.11) rather than the exact Hessian. According to the definition of $\mathcal{A}$ (i.e., $\mathbf{D}_\mathcal{A} = 0$ and $\mathbf{\Theta}_{t,\mathcal{A}} = 0$), the adaptive stopping criterion in (4.13) can only be verified over the variables in the free set $\mathcal{F}$. Following [5], the sufficient reduction condition in the line search of inexact FMGL uses $L_t(\mathbf{\Theta}_t + \beta\mathbf{D}) - L_t(\mathbf{\Theta}_t)$ instead of $\beta\delta_t$ in (4.6).

Although the global convergence of inexact proximal Newton-type (including Newton and quasi-Newton) methods is guaranteed, it is still challenging to prove a convergence rate for inexact proximal quasi-Newton methods such as inexact FMGL where an approximate Hessian is used. The local convergence rate of the inexact proximal Newton method has been studied in [5, 23]. However, those proofs require the Hessian to be exact, which is not the case in inexact FMGL. It is worth noting that Jiang, Sun, and Toh [19] and Scheinberg and Tang [41] have recently shown a sublinear global convergence rate for inexact proximal quasi-Newton methods. In order to have such global convergence rate, the method in [19] requires stricter conditions on the approximate Hessian, while the method in [41] uses a prox-parameter updating mechanism instead of line search for acceptance of iterates [41]. It is difficult to apply their techniques to our method, since the conditions in [19, 41] for the global convergence rates may not hold for inexact FMGL. The property of the selected active set $\mathcal{A}$ and the special structure of the approximate Hessian may be the key to establishing a faster local convergence rate for inexact FMGL. We defer these analyses to future work.

5. Experimental results. In this section, we evaluate the proposed algorithm and screening rule on synthetic datasets and two real datasets: ADHD-200 [33] and FDG-PET [48] images. The experiments are performed on a PC with a quad-core Intel 2.67 GHz CPU and 9 GB memory.

5.1. Simulation. We conduct experiments to demonstrate the effectiveness of the proposed screening rule and the efficiency of our FMGL method. The following algorithms are included in our comparisons:

- FMGL: the proposed second-order method in Algorithm 1.
- ADMM: the ADMM method.
- FMGL-S: FMGL with screening.
- ADMM-S: ADMM with screening.

Both FMGL and ADMM are written in MATLAB, and they are available online.[2] Since both methods involve solving (4.4), which involves a double loop, we implement the subroutine for solving (4.4) in C for a fair comparison.

The synthetic covariance matrices are generated as follows. We first generate K block diagonal ground truth precision matrices $\mathbf{\Theta}^{(k)}$ with L blocks, and each block $\mathbf{\Theta}_l^{(k)}$ is of size $(p/L) \times (p/L)$. Each $\mathbf{\Theta}_l^{(k)}$, $l = 1, \dots, L$, $k = 1, \dots, K$, has random sparsity structures. We control the number of nonzeros in each $\mathbf{\Theta}_l^{(k)}$ to be about $10p/L$ so that the total number of nonzeros in the K precision matrices is $10Kp$. Given the precision matrices, we draw $5p$ samples from each Gaussian distribution to compute the sample covariance matrices. The fused penalty parameter λ_2 is fixed to 0.1, and the ℓ_1 regularization parameter λ_1 is selected so that the total number of nonzeros in the solution is about $10Kp$.

5.1.1. Convergence. We first explore the convergence behavior of FMGL with different stopping criteria in NSPG. Three stopping criteria are considered:

- 1E-6: stop when the relative error $\frac{\max\{\|\mathbf{\Theta}_r^{(k)} - \mathbf{\Theta}_{r-1}^{(k)}\|_\infty\}}{\max\{\|\mathbf{\Theta}_{r-1}^{(k)}\|_\infty\}} \le 1e\text{-}6$.
- Exact: the subproblems are solved accurately as in [23]. (More precisely, NSPG stops when $\frac{\max\{\|\mathbf{\Theta}_r^{(k)} - \mathbf{\Theta}_{r-1}^{(k)}\|_\infty\}}{\max\{\|\mathbf{\Theta}_{r-1}^{(k)}\|_\infty\}} \le 1e\text{-}12$).
- Adaptive: stop when the adaptive stopping criterion (4.13) is satisfied. The forcing term η_k is chosen as in [23].

We plot the relative error of objective value versus Newton iterations and computational time on a synthetic dataset ($K = 5,\ L = 1,\ p = 500$) in Figure 1. We observe from Figure 1 that the exact stopping criterion has the fastest convergence with respect to Newton iterations. Considering computational time, the adaptive criterion has the best convergence behavior. Although the criterion 1E-6 has almost the same convergence behavior as the exact criterion in the first few steps, FMGL with this constant stopping criterion converges slower when the approximated solution is close enough to the optimal solution. We also include the convergence of ADMM in Figure 1. We can see that ADMM converges much more slowly than does FMGL.

5.1.2. Screening. We conduct experiments to show the effectiveness of the proposed screening rule. NSPG is terminated using the adaptive stop criterion. FMGL is terminated when the relative error of the objective value is smaller than $1e$-5, and ADMM stops when it achieves an objective value equal to or smaller than that of FMGL. The results presented in Table 1 show that FMGL is consistently faster than

[2]http://www.yelab.net/software/MGL/

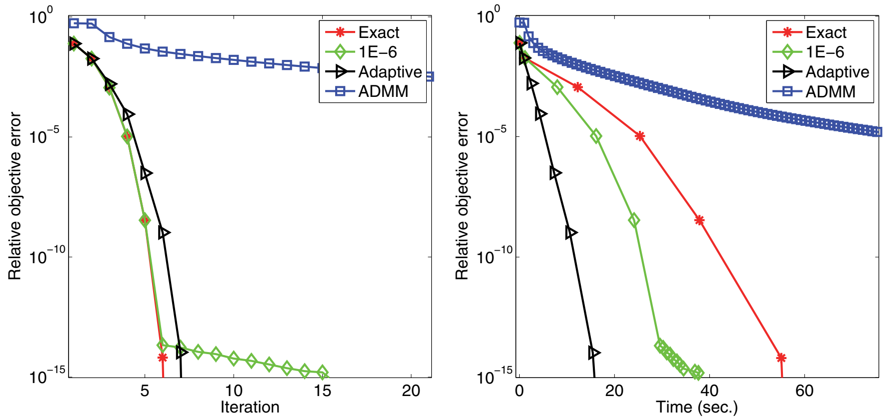


FIG. 1. *Convergence behavior of FMGL with three stopping criteria (exact, adaptive, and* 1*E*-6*) and ADMM.*

TABLE 1

Comparison of the proposed FMGL and ADMM methods with and without screening in terms of average computational time (seconds). FMGL-S and ADMM-S are FMGL and ADMM with screening, respectively. p stands for the dimension, K is the number of graphs, L is the number of blocks, and λ_1 is the ℓ_1 regularization parameter. The fused penalty parameter λ_2 is fixed to 0.1*. $\|\mathbf{\Theta}\|_0$ represents the total number of nonzero entries in ground truth precision matrices $\mathbf{\Theta}^{(k)}$, $k = 1, \dots, K$, and $\|\mathbf{\Theta}^*\|_0$ is the number of nonzeros in the solution.*

Data and parameter setting						Computational time			
p	K	L	$\|\mathbf{\Theta}\|_0$	λ_1	$\|\mathbf{\Theta}^*\|_0$	FMGL-S	FMGL	ADMM-S	ADMM
500	2	5	9848	0.08	9810	**0.44**	4.13	13.30	100.79
1000			20388	0.088	19090	**2.25**	17.88	57.44	617.88
500	5		24866	0.055	23304	**0.97**	12.23	32.40	286.98
1000			50598	0.054	44030	**5.16**	50.95	174.91	1595.91
500	10		49092	0.051	45474	**2.33**	24.35	63.75	458.51
1000			100804	0.046	84310	**10.27**	111.78	302.86	2966.72
500	2	10	9348	0.07	9386	**0.32**	4.87	6.82	105.01
1000			19750	0.08	20198	**0.76**	17.93	25.62	674.28
500	5		23538	0.055	22900	**0.77**	14.96	15.09	256.33
1000			49184	0.054	45766	**1.92**	53.96	64.31	1314.18
500	10		47184	0.051	47814	**1.66**	52.32	29.86	455.43
1000			98564	0.046	94566	**4.44**	126.26	128.52	2654.24

ADMM. Moreover, the screening rule can achieve great computational gain. The speedup with the screening rule is about 10 and 20 times for $L = 5$ and 10, respectively.

5.2. Real data.

5.2.1. ADHD-200. Attention Deficit Hyperactivity Disorder (ADHD) affects at least 5–10% of school-age children with annual costs exceeding 36 billion/year in the United States. The ADHD-200 project has released resting-state functional magnetic resonance images (fMRI) of 491 typically developing children and 285 ADHD children, aiming to encourage research on ADHD. The data used in this experiment is preprocessed using the NIAK pipeline and downloaded from neurobureau.[3] More details about the preprocessing strategy can be found in the same website. The dataset

[3]http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:NIAKPipeline

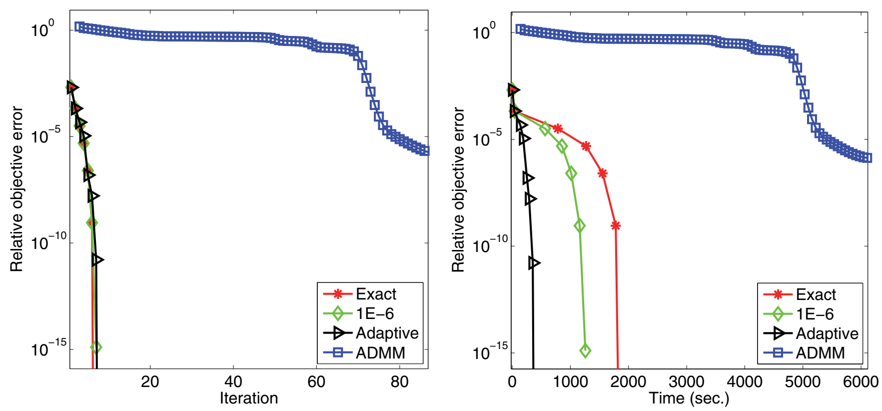


FIG. 2. *Comparison of FMGL with three stopping criteria and ADMM in terms of objective value curve on the ADHD-*200 *dataset. The dimension p is* 2834*, and the number of graphs K is* 3*.*

we choose includes 116 typically developing children (TDC), 29 ADHD-Combined (ADHD-C), and 49 ADHD-Inattentive (ADHD-I). There are 231 time series and 2834 brain regions for each subject. We want to estimate the graphs of the three groups simultaneously. The sample covariance matrix is computed using all data from the same group. Since the number of brain regions p is 2834, obtaining the precision matrices is computationally intensive. We use this data to test the effectiveness of the proposed screening rule. λ_1 and λ_2 are set to 0.6 and 0.015. The comparison of FMGL with three stopping criteria and ADMM in terms of the objective value curve is shown in Figure 2. The result shows that FMGL converges much faster than ADMM. To obtain a solution of precision 1e-5, the computational times of FMGL (Adaptive), FMGL (1E-6), FMGL (Exact), and ADMM are 252.78, 855.86, 1269.75 and 5410.48 seconds, respectively. However, with the screening, the computational times of FMGL-S (Adaptive), FMGL-S (1E-6), FMGL-S (Exact), and ADMM-S are reduced to 4.02, 12.51, 19.55, and 80.52 seconds, respectively, demonstrating the superiority of the proposed screening rule. The obtained solution has 1443 blocks. The largest one including 634 nodes is shown in Figure 3.

The block structures of the FMGL solution are the same as those identified by the screening rule. The screening rule can be used to analyze the rough structures of the graphs. The cost of identifying blocks using the screening rule is negligible compared to that of estimating the graphs. For high-dimensional data such as ADHD-200, it is practical to use the screening rule to identify the block structure before estimating the large graphs. We use the screening rule to identify block structures on ADHD-200 data with varying λ_1 and λ_2. The size distribution is shown in Figure 4. We can observe that the number of blocks increases, and the size of blocks decreases, when the regularization parameter value increases.

5.2.2. FDG-PET. In this experiment, we use FDG-PET images from 74 Alzheimer's disease (AD), 172 mild cognitive impairment (MCI), and 81 normal control (NC) subjects downloaded from the Alzheimer's disease neuroimaging initiative (ADNI) database [48]. The different regions of the whole brain volume can be represented by 116 anatomical volumes of interest (AVOI), defined by automated anatomical labeling (AAL) [45]. Then we extracted data from each of the 116 AVOIs and derived the average of each AVOI for each subject. The 116 AVOIs can be categorized into

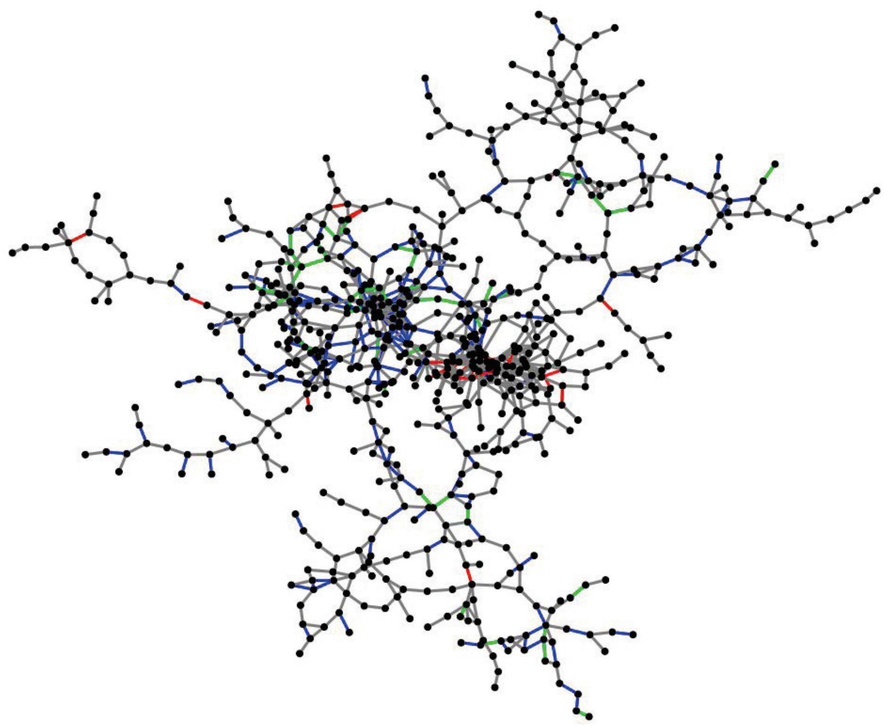
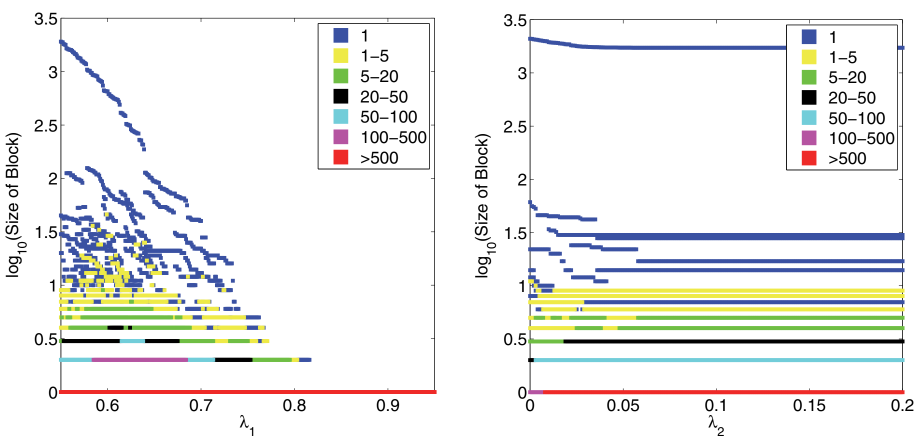


Fig. 3. *A subgraph of ADHD-*200 *identified by FMGL with the proposed screening rule. The grey edges are common edges among the three graphs; the red, green, and blue edges (see color online) are the specific edges for TDC, ADHD-I, and ADHD-C, respectively.*



(a) (b)

Fig. 4. *The size distribution of blocks (in the logarithmic scale) identified by the proposed screening rule. The color represents the number of blocks of a specified size.* (a) λ_1 *varies from* 0.5 *to* 0.95 *with* λ_2 *fixed to* 0.015. (b) λ_2 *varies from* 0 *to* 0.2 *with* λ_1 *fixed to* 0.55.

10 groups: prefrontal lobe, other parts of the frontal lobe, parietal lobe, occipital lobe, thalamus, insula, temporal lobe, corpus striatum, cerebellum, and vermis. More

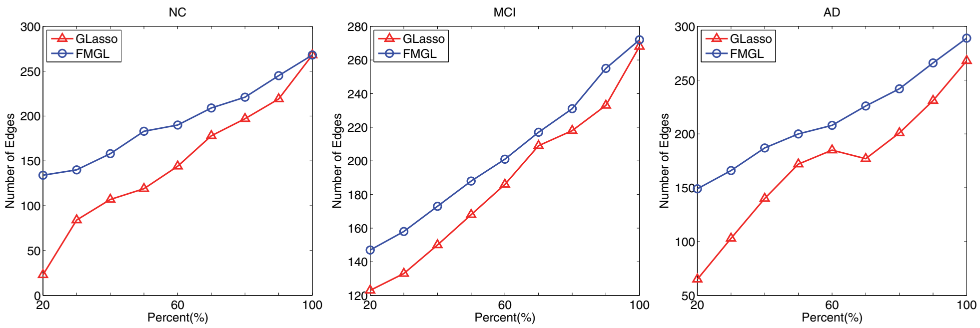


FIG. 5. *The average number of stable edges detected by FMGL and GLasso in NC (left), MCI (middle), and AD (right) of* 500 *replications. Sample size varies from* 20% *to* 100% *with a step of* 10%.

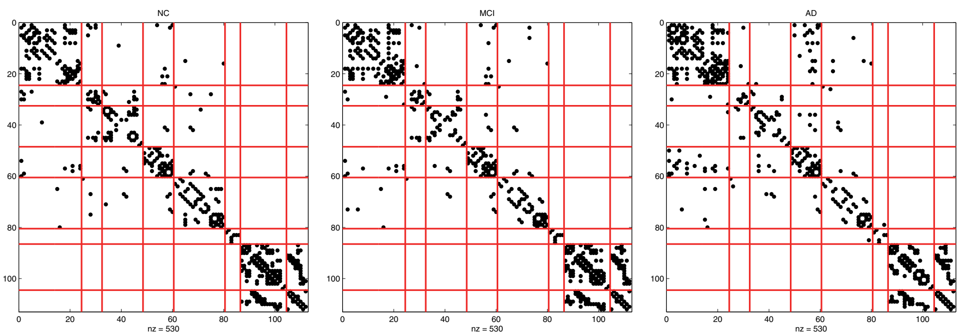


FIG. 6. *Brain connection models with* 265 *edges: NC (left), MCI (middle), and AD (right). In each figure, the diagonal blocks are prefrontal lobe, other parts of frontal lobe, parietal lobe, occipital lobe, temporal lobe, corpus striatum, cerebellum, and vermis, respectively.*

details about the categories can be found in [45, 47]. We remove two small groups (thalamus and insula) containing only 4 AVOIs in our experiments.

To examine whether FMGL can effectively utilize the information of common structures, we randomly select g percent samples from each group, where g varies from 20 to 100 with a step size of 10. For each g, λ_2 is fixed to 0.1, and λ_1 is adjusted to make sure that the number of edges in each group is about the same. We perform 500 replications for each g. The edges with probability larger than 0.85 are considered as stable edges. The results showing the numbers of stable edges are summarized in Figure 5. We can observe that FMGL is more stable than GLasso. When the sample size is too small (say 20%), there are only 20 stable edges in the graph of NC obtained by GLasso. But the graph of NC obtained by FMGL still has about 140 stable edges, illustrating the superiority of FMGL in stability.

The brain connectivity models obtained by FMGL are shown in Figure 6. We can see that the number of connections within the prefrontal lobe significantly increases and the number of connections within the temporal lobe significantly decreases from NC to AD, which is supported by previous findings [1, 15]. The connections between the prefrontal and occipital lobes increase from NC to AD, and connections within cerebellum decrease. We can also find that the adjacent graphs are similar, indicating that FMGL can identify the common structures but also keep the meaningful differences.

6. Conclusion. In this paper, we have considered simultaneously estimating multiple graphical models by maximizing a fused penalized log likelihood. We have derived a set of necessary and sufficient conditions for the FMGL solution to be block diagonal for an arbitrary number of graphs. A screening rule has been developed to enable the efficient estimation of large multiple graphs. The second-order method is employed to solve the FMGL, which is shown to be equivalent to a proximal quasi-Newton method. The global convergence of the proposed method with an adaptive stopping criterion is guaranteed. An active set identification scheme has been proposed to identify the variables to be updated during the Newton iterations, thus reducing the computation. Numerical experiments on synthetic and real data demonstrate the efficiency and effectiveness of the proposed method and the screening rule. We plan to further explore the convergence properties of the second-order methods when the subproblems are solved inexactly. Due to the active set identification scheme, the proposed second-order method is suitable for warm-start techniques. A good initial solution can further speed up the computation. As part of future work, we plan to explore how to efficiently find a good initial solution to further improve the efficiency of the proposed method. One possibility is to use divide-and-conquer techniques [16].

Appendix A. Supporting proofs.

A.1. Uniqueness of the FMGL solution. To prove Theorem 2.1, we first establish a technical lemma regarding the existence of a solution for a standard graphical lasso problem.

LEMMA A.1. *Let* $\mathbf{S} \in \mathcal{S}_+^p$ *and* $\mathbf{\Lambda} \in \mathcal{S}^p$ *be such that* $\mathrm{Diag}(\mathbf{S}) + \mathbf{\Lambda} > 0$ *and* $\mathrm{diag}(\mathbf{\Lambda}) \geq 0$. *Consider the problem*

$$\min_{\mathbf{X}\succ 0} \underbrace{-\log\det(\mathbf{X}) + \mathrm{tr}(\mathbf{S}\mathbf{X}) + \sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}|}_{f(\mathbf{X})}. \tag{A.1}$$

Then the following statements hold:

(a) *Problem* (A.1) *has a unique optimal solution.*

(b) *The sublevel set* $\mathcal{L} = \{\mathbf{X} \succ 0 : f(\mathbf{X}) \leq \alpha\}$ *is compact for any* $\alpha \geq f^*$, *where* f^* *is the optimal value of* (A.1).

Proof. (a) Let $\mathcal{U} = \{\mathbf{U} \in \mathcal{S}^p : \mathbf{U}_{ij} \in [-1, 1]\ \forall i, j\}$. Consider the problem

$$\max_{\mathbf{U}\in\mathcal{U}} \{\log\det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) : \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} \succ 0\}. \tag{A.2}$$

We first claim that the feasible region of problem (A.2) is nonempty, or equivalently, there exists $\bar{\mathbf{U}} \in \mathcal{U}$ such that $\lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \bar{\mathbf{U}}) > 0$. Indeed, one can observe that

$$\begin{aligned}
\max_{\mathbf{U}\in\mathcal{U}} \lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) &= \max_{t,\mathbf{U}\in\mathcal{U}} \{t : \mathbf{\Lambda} \circ \mathbf{U} + \mathbf{S} - t\mathbf{I} \succeq 0\} \\
&= \min_{\mathbf{X}\succeq 0} \max_{t,\mathbf{U}\in\mathcal{U}} \{t + \mathrm{tr}(\mathbf{X}(\mathbf{\Lambda} \circ \mathbf{U} + \mathbf{S} - t\mathbf{I}))\} \\
&= \min_{\mathbf{X}\succeq 0} \left\{ \mathrm{tr}(\mathbf{S}\mathbf{X}) + \sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| :\ \mathrm{tr}(\mathbf{X}) = 1 \right\},
\end{aligned} \tag{A.3}$$

where the second equality follows from the Lagrangian duality since its associated Slater condition is satisfied. Let $\Omega := \{\mathbf{X} \in \mathcal{S}^p : \mathrm{tr}(\mathbf{X}) = 1,\ \mathbf{X} \succeq 0\}$. By the assumption $\mathrm{Diag}(\mathbf{S}) + \mathbf{\Lambda} > 0$, we see that $\mathbf{\Lambda}_{ij} > 0$ for all $i \neq j$ and $\mathbf{S}_{ii} + \mathbf{\Lambda}_{ii} > 0$ for every i. Since $\Omega \subset \mathcal{S}_+^p$, we have $\mathrm{tr}(\mathbf{SX}) \geq 0$ for all $\mathbf{X} \in \Omega$. If there exists some $k \neq l$ such that $\mathbf{X}_{kl} > 0$, then $\sum_{i \neq j} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| > 0$, and hence

$$\mathrm{tr}(\mathbf{SX}) + \sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| \ > \ 0 \quad \forall \mathbf{X} \in \Omega. \tag{A.4}$$

Otherwise, one has $\mathbf{X}_{ij} = 0$ for all $i \neq j$, which, together with the facts that $\mathbf{S}_{ii}+\mathbf{\Lambda}_{ii} > 0$ for all i and $\mathrm{tr}(\mathbf{X}) = 1$, implies that for all $\mathbf{X} \in \Omega$,

$$\mathrm{tr}(\mathbf{SX}) + \sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| \ = \ \sum_{i}(\mathbf{S}_{ii} + \mathbf{\Lambda}_{ii})\mathbf{X}_{ii} \ \geq \ \mathrm{tr}(\mathbf{X}) \min_i(\mathbf{S}_{ii} + \mathbf{\Lambda}_{ii}) \ > \ 0.$$

Hence, (A.4) again holds. Combining (A.3) with (A.4), one can then see that $\max_{\mathbf{U} \in \mathcal{U}} \lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) > 0$. Therefore, problem (A.2) has at least a feasible solution.

We next show that problem (A.2) has an optimal solution. Let $\bar{\mathbf{U}}$ be a feasible point of (A.2), and

$$\bar{\Omega} := \{\mathbf{U} \in \mathcal{U} : \log\det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) \geq \log\det(\mathbf{S} + \mathbf{\Lambda} \circ \bar{\mathbf{U}}),\ \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} \succ 0\}.$$

One can observe that $\{\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} : \mathbf{U} \in \mathcal{U}\}$ is compact. Using this fact, it is not hard to see that $\log\det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) \to -\infty$ as $\mathbf{U} \in \mathcal{U}$ and $\lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) \downarrow 0$. Thus there exists some $\delta > 0$ such that

$$\bar{\Omega} \subseteq \{\mathbf{U} \in \mathcal{U} : \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} \succeq \delta I\},$$

which implies that

$$\bar{\Omega} = \{\mathbf{U} \in \mathcal{U} : \log\det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) \geq \log\det(\mathbf{S} + \mathbf{\Lambda} \circ \bar{\mathbf{U}}),\ \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} \succeq \delta I\}.$$

Hence, $\bar{\Omega}$ is a compact set. In addition, one can observe that problem (A.2) is equivalent to

$$\max_{\mathbf{U} \in \bar{\Omega}} \log\det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}).$$

The latter problem clearly has an optimal solution and so does problem (A.2).

Finally, we show that $\mathbf{X}^* = (\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}^*)^{-1}$ is the unique optimal solution of (A.1), where $\mathbf{U}^*$ is an optimal solution of (A.2). Since $\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}^* \succ 0$, we have $\mathbf{X}^* \succ 0$. By the definitions of $\mathcal{U}$ and $\mathbf{X}^*$, and the first-order optimality conditions of (A.2) at $\mathbf{U}^*$, one can have

$$\mathbf{U}^*_{ij} = \begin{cases} 1 & \text{if } \mathbf{X}^*_{ij} > 0, \\ \beta \in [-1, 1] & \text{if } \mathbf{X}^*_{ij} = 0, \\ -1 & \text{otherwise.} \end{cases}$$

It follows that $\mathbf{\Lambda} \circ \mathbf{U}^* \in \partial(\sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}|)$ at $\mathbf{X} = \mathbf{X}^*$, where $\partial(\cdot)$ stands for the subdifferential of the associated convex function. For convenience, let $f(\mathbf{X})$ denote the objective function of (A.1). Then we have

$$-(\mathbf{X}^*)^{-1} + \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}^* \in \partial f(\mathbf{X}^*),$$

which, together with $\mathbf{X}^* = (\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}^*)^{-1}$, implies that $0 \in \partial f(\mathbf{X}^*)$. Hence, $\mathbf{X}^*$ is an optimal solution of (A.1), and moreover, it is unique due to the strict convexity of $-\log\det(\cdot)$.

(b) By statement (a), problem (A.1) has a finite optimal value f^*. Hence, the above sublevel set $\mathcal{L}$ is nonempty. We can observe that for any $\mathbf{X} \in \mathcal{L}$,

$$\begin{aligned}\frac{1}{2}\sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| &= f(\mathbf{X}) - \underbrace{\left[-\log\det(\mathbf{X}) + \mathrm{tr}(\mathbf{SX}) + \frac{1}{2}\sum_{ij}\mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}|\right]}_{\underline{f}(\mathbf{X})} \\ &\le \alpha - \underline{f}^*, \end{aligned} \tag{A.5}$$

where $\underline{f}^* := \inf\{\underline{f}(\mathbf{X}) : \mathbf{X} \succ 0\}$. By the assumption $\mathrm{Diag}(\mathbf{S}) + \mathbf{\Lambda} > 0$, one has $\mathrm{Diag}(\mathbf{S}) + \mathbf{\Lambda}/2 > 0$. This together with statement (a) yields $\underline{f}^* \in \Re$. Notice that $\mathbf{\Lambda}_{ij} > 0$ for all $i \neq j$. This relation and (A.5) imply that $\mathbf{X}_{ij}$ is bounded for all $\mathbf{X} \in \mathcal{L}$ and $i \neq j$. In addition, it is well known that $\det(\mathbf{X}) \le \mathbf{X}_{11}\mathbf{X}_{22}\cdots\mathbf{X}_{pp}$ for all $\mathbf{X} \succeq 0$. Using this relation, the definition of $f(\cdot)$, and the boundedness of $\mathbf{X}_{ij}$ for all $\mathbf{X} \in \mathcal{L}$ and $i \neq j$, we have that for every $\mathbf{X} \in \mathcal{L}$,

$$\begin{aligned}\sum_i -\log(\mathbf{X}_{ii}) + (\mathbf{S}_{ii} + \Lambda_{ii})\mathbf{X}_{ii} &\le f(\mathbf{X}) - \sum_{i\neq j}(\mathbf{S}_{ij}\mathbf{X}_{ij} + \Lambda_{ij}|\mathbf{X}_{ij}|) \\ &\le \alpha - \sum_{i\neq j}(\mathbf{S}_{ij}\mathbf{X}_{ij} + \Lambda_{ij}|\mathbf{X}_{ij}|) \le \delta \end{aligned} \tag{A.6}$$

for some $\delta > 0$. In addition, notice from the assumption that $\mathbf{S}_{ii} + \Lambda_{ii} > 0$ for all i, and hence

$$-\log(\mathbf{X}_{ii}) + (\mathbf{S}_{ii} + \Lambda_{ii})\mathbf{X}_{ii} \;\ge\; 1 + \min_k \log(\mathbf{S}_{kk} + \Lambda_{kk}) \;=:\; \sigma$$

for all i. This relation together with (A.6) implies that for every $\mathbf{X} \in \mathcal{L}$ and all i,

$$-\log(\mathbf{X}_{ii}) + (\mathbf{S}_{ii} + \Lambda_{ii})\mathbf{X}_{ii} \;\le\; \delta - (p-1)\sigma,$$

and hence $\mathbf{X}_{ii}$ is bounded for all i and $\mathbf{X} \in \mathcal{L}$. We thus conclude that $\mathcal{L}$ is bounded. In view of this result and the definition of f, it is not hard to see that there exists some $\nu > 0$ such that $\lambda_{\min}(\mathbf{X}) \ge \nu$ for all $\mathbf{X} \in \mathcal{L}$. Hence, one has

$$\mathcal{L} = \{\mathbf{X} \succeq \nu I : f(\mathbf{X}) \le \alpha\}.$$

By the continuity of f on $\{\mathbf{X} : \mathbf{X} \succeq \nu I\}$, it follows that $\mathcal{L}$ is closed. Hence, $\mathcal{L}$ is compact. ∎

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Since $\lambda_1 > 0$ and $\mathrm{diag}(\mathbf{S}^{(k)}) > 0$, $k = 1, \ldots, K$, it follows from Lemma A.1 that there exists some δ such that for each $k = 1, \ldots, K$,

$$-\log\det(\mathbf{\Theta}^{(k)}) + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}) + \lambda_1 \sum_{i\neq j}|\mathbf{\Theta}_{ij}^{(k)}| \;\ge\; \delta \qquad \forall \mathbf{\Theta}^{(k)} \succ 0.$$

For convenience, let $h(\mathbf{\Theta})$ denote the objective function of (2.2), and $\bar{\mathbf{\Theta}} = (\bar{\mathbf{\Theta}}^{(1)}, \ldots, \bar{\mathbf{\Theta}}^{(K)})$ an arbitrary feasible point of (2.2). Let

$$\begin{aligned}\Omega &= \left\{\mathbf{\Theta} = (\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(K)}) : h(\mathbf{\Theta}) \le h(\bar{\mathbf{\Theta}}),\ \mathbf{\Theta}^{(k)} \succ 0,\ k = 1, \ldots, K\right\}, \\ \Omega_k &= \left\{\mathbf{\Theta}^{(k)} \succ 0 : -\log\det(\mathbf{\Theta}^{(k)}) + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Theta}^{(k)}) + \lambda_1 \sum_{i\neq j}|\mathbf{\Theta}_{ij}^{(k)}| \;\le\; \bar{\delta}\right\}\end{aligned}$$

for $k = 1, \ldots, K$, where $\bar{\bar{\delta}} = h(\bar{\mathbf{\Theta}}) - (K-1)\delta$. Then it is not hard to observe that $\Omega \subseteq \bar{\Omega} := \Omega_1 \times \cdots \times \Omega_K$. Moreover, problem (2.2) is equivalent to

$$\min_{\mathbf{\Theta} \in \bar{\Omega}} h(\mathbf{\Theta}). \tag{A.7}$$

In view of Lemma A.1, we know that Ω_k is compact for all k, which implies that $\bar{\Omega}$ is also compact. Notice that h is continuous and strictly convex on $\bar{\Omega}$. Hence, problem (A.7) has a unique optimal solution and so does problem (2.2). ▯

A.2. Proof of Lemma 3.7.

Proof. (i) and (ii) can be proved in a similar way as used for Lemma 3.3.

(iii) Similar to Lemma 3.3, we can show that $\mathrm{ext}(P_{I,J}) \neq \emptyset$. Next we show that $\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\} \subseteq Q$. Denote G_J and $G_{\bar{J}}$ as the subgraphs with edges only in E_J and $E_{\bar{J}}$, respectively. Accordingly, $\mathcal{G}_J$ represents the set of all possible subgraphs with only one connected component in G_J, and $\mathcal{V}_J$ denotes the corresponding node sets of $\mathcal{G}_J$. Then we have $\mathcal{V}_J \cup \mathcal{V}_{\bar{J}} \subseteq \mathcal{V}$. Moreover, $\cup\{\mathcal{V}_J \cup \mathcal{V}_{\bar{J}}, J \subseteq \{1, \ldots, |E|\}\} = \mathcal{V}$.

Let $d \in \cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\}$. Then $d \neq 0$, and the number of independent active inequalities at d is $n-1$. It is clear that the maximum number of independent active inequalities restricted to the nodes in $V_m \in \mathcal{V}$ is $|V_m|$, which is achieved when $d_i = 0$ for all $i \in V_m$. If $d_i \neq 0$ for all $i \in V_m, V_m \neq \emptyset$, it is not hard to show that the maximum number of independent active inequalities restricted to V_m is $|V_m| - 1$, which is achieved when $d_i = d_j$ for all $i, j \in V_m$. Suppose that there exist two nonempty and nonoverlapping sets V_l and V_m such that $d_i = d_j \neq 0$ for all $i, j \in V_l$ and $d_i = d_j \neq 0$ for all $i, j \in V_m$. We consider the following two cases: (a) there is no edge across V_l and V_m. In this case, the maximum number of independent active inequalities is $|V_m| - 1 + |V_l| - 1 + n - |V_m| - |V_l| = n - 2$. (b) $d_i \neq d_j,\ i \in V_l, j \in V_m$; thus inequalities from the edges across V_l and V_m are inactive. In this case, the maximum number of independent active inequalities is $|V_m| - 1 + |V_l| - 1 + n - |V_m| - |V_l| = n-2$. This is a contradiction to the definition of extreme ray d. Combining the arguments above, we show that all nodes in V with a nonzero value in d form a set in $\mathcal{V}$. Therefore, $\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\} \subseteq Q$.

(iv) Let $d \in Q$ be arbitrarily chosen. Then, there exist $\alpha \neq 0$ and a $V_m \in \mathcal{V}$ such that $d_i = \alpha,\ i \in V_m$, and the rest of d_i's are 0. If $\alpha > 0$, it is not hard to see that $d \in \mathrm{ext}(P_{I,J})$ with $I = \{1, \ldots, n\}$ and J such that $E_J = \{(u,v) : u, v \in V_m,\ (u,v) \in E\} \cup \{(u,v) : u \in V_m, v \in \bar{V}_m, (u,v) \in E\}$, where $\bar{V}_m$ is the complement of V_m. If $\alpha < 0$, $d \in \mathrm{ext}(P_{I,J})$ with $I = \emptyset$ and J such that $E_J = \{(u,v) : u, v \in V_m,\ (u,v) \in E\} \cup \{(u,v) : u \in \bar{V}_m, v \in V_m, (u,v) \in E\}$. Hence, $d \in \cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\}$. Combining this with (iii), we have $\cup\{\mathrm{ext}(P_{I,J}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\}\} = Q$. ▯

A.2.1. Proof of Theorem 3.8.

Proof. By the first-order optimality conditions, $\widehat{\mathbf{\Theta}}^{(k)} \succ 0$, $k = 1, \ldots, K$, is the optimal solution of problem (3.22) if and only if it satisfies

$$-\widehat{\mathbf{W}}_{ii}^{(k)} + \mathbf{S}_{ii}^{(k)} = 0, \quad 1 \leq k \leq K, \tag{A.8}$$

$$-\widehat{\mathbf{W}}_{ij} + \mathbf{S}_{ij} + \partial\phi_{ij} = 0, \tag{A.9}$$

for all $i, j = 1, \ldots, p,\ i \neq j$, where $\widehat{\mathbf{W}}_{ij} = (\widehat{\mathbf{W}}_{ij}^{(1)}, \ldots, \widehat{\mathbf{W}}_{ij}^{(K)})^T$, $\mathbf{S}_{ij} = (\mathbf{S}_{ij}^{(1)}, \ldots, \mathbf{S}_{ij}^{(K)})^T$, and $\partial\phi_{ij}$ is a subgradient of $\phi(\mathbf{\Theta}_{ij})$ at $\mathbf{\Theta}_{ij} = \widehat{\mathbf{\Theta}}_{ij}$.

Suppose that $\widehat{\mathbf{\Theta}}^{(k)}$, $k = 1, \ldots, K$, is a block diagonal optimal solution of problem (3.22) with L known blocks C_l, $l = 1, \ldots, L$. $\widehat{\mathbf{W}}_{ij}^{(k)} = \widehat{\mathbf{\Theta}}_{ij}^{(k)} = 0$ for $i \in C_l,\ j \in C_{l'},\ l \neq l'$. This together with (A.9) implies that for each $i \in C_l,\ j \in C_{l'},\ l \neq l'$ there exists a $\partial\phi_{ij}$ such that

$$\mathbf{S}_{ij} + \partial\phi_{ij} = 0,$$

which directly shows that 0 is the optimal solution of (3.23). Sufficiency can be proved in a way similar to that used for Theorem 3.2. $\square$

REFERENCES

[1] N. P. Azari, S. I. Rapoport, C. L. Grady, M. B. Schapiro, J. A. Salerno, A. Gonzales-Aviles, and B. Horwitz, *Patterns of interregional correlations of cerebral glucose metabolic rates in patients with dementia of the Alzheimer type*, Neurodegeneration, 1 (1992), pp. 101–111.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data*, J. Mach. Learning Res., 9 (2008), pp. 485–516.

[3] D. Bertsekas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Nashua, NH, 1997.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learning, 3 (2011), pp. 1–122.

[5] R. H. Byrd, J. Nocedal, and F. Oztoprak, *An inexact successive quadratic approximation method for convex $L1$ regularized optimization*, arXiv:1309.3529, 2013.

[6] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis*, Mol. Syst. Biol., 3 (2007), doi: 10.1038/msb4100180.

[7] L. Condat, *A direct algorithm for 1d total variation denoising*, Signal Process. Lett., 20 (2013), pp. 1054–1057.

[8] P. Danaher, P. Wang, and D. M. Witten, *The joint graphical lasso for inverse covariance estimation across multiple classes*, J. Roy. Statist. Soc. Ser. B, 76 (2014), pp. 373–397.

[9] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, *First-order methods for sparse covariance selection*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 56–66.

[10] Q. Dinh, A. Kyrillidis, and V. Cevher, *A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions*, in Proceedings of the 30th International Conference on Machine Learning (ICML 2013), 2013, pp. 271–279.

[11] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[12] J. Guo, E. Levina, G. Michailidis, and J. Zhu, *Joint estimation of multiple graphical models*, Biometrika, 98 (2011), pp. 1–15.

[13] S. Hara and T. Washio, *Common substructure learning of multiple graphical Gaussian models*, MLKDD, (2011), pp. 1–16.

[14] J. Honorio and D. Samaras, *Multi-task learning of Gaussian graphical models*, in Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, 2010, pp. 447–454.

[15] B. Horwitz, C. L. Grady, N. L. Schlageter, R. Duara, and S. I. Rapoport, *Intercorrelations of regional cerebral glucose metabolic rates in Alzheimer's disease*, Brain Res., 407 (1987), pp. 294–306.

[16] C. J. Hsieh, I. Dhillon, P. Ravikumar, and A. Banerjee, *A divide-and-conquer method for sparse inverse covariance estimation*, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, 2012, pp. 2339–2347.

[17] C. J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, *Sparse inverse covariance matrix estimation using quadratic approximation*, in Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS), Grenada, 2011, pp. 2330–2338.

[18] S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, A. Fleisher, E. Reiman, and J. Ye, *Learning brain connectivity of Alzheimer's disease from neuroimaging data*, in Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), 2009, pp. 808–816.

[19] K. Jiang, D. Sun, and K.-C. Toh, *An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP*, SIAM J. Optim., 22 (2012), pp. 1042–1064.
[20] T. Joachims, *Making large-scale support vector machine learning practical*, in Advances in Kernel Methods, MIT Press, Cambridge, MA, 1999, pp. 169–184.
[21] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, *Estimating time-varying networks*, Ann. Appl. Statist., 4 (2010), pp. 94–123.
[22] M. Kolar and E. P. Xing, *On time varying undirected graphs*, in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTAT), Ft. Lauderdale, 2011, pp. 407–415.
[23] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM J. Optim., 24 (2014), pp. 1420–1443.
[24] L. Li and K. C. Toh, *An inexact interior point method for l1-regularized sparse covariance selection*, Math, Program, Comput., 2 (2010), pp. 291–315.
[25] H. Liu, K. Roeder, and L. Wasserman, *Stability approach to regularization selection (StARS) for high dimensional graphical models*, in Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS), Grenada, 2011, pp. 1432–1440.
[26] J. Liu and J. Ye, *Moreau-Yosida regularization for grouped tree structure learning*, in Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), 2010, pp. 1459–1467.
[27] J. Liu, L. Yuan, and J. Ye, *An efficient algorithm for a class of fused lasso problems*, in Proceedings of the 16th ACM SIGDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2010, pp. 323–332.
[28] Z. Lu, *Smooth optimization approach for sparse covariance selection*, SIAM J. Optim., 19 (2009), pp. 1807–1827.
[29] Z. Lu, *Adaptive first-order methods for general sparse inverse covariance selection*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2000–2016.
[30] Z. Lu and Y. Zhang, *An augmented Lagrangian approach for sparse principal component analysis*, Math. Program., 135 (2012), pp. 149–193.
[31] R. Mazumder and T. Hastie, *Exact covariance thresholding into connected components for large-scale graphical lasso*, J. Mach. Learning Res., 13 (2012), pp. 781–794.
[32] R. Mazumder and T. Hastie, *The graphical lasso: New insights and alternatives*, Electron. J. Statist., 6 (2012), pp. 2125–2149.
[33] M. P. Milham, *The ADHD-*200 *Sample*, fMRI dataset, available online from http://fcon_1000.projects.nitrc.org/indi/adhd200/.
[34] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist., 34 (2006), pp. 1436–1462.
[35] K. Mohan, M. Chung, S. Han, D. Witten, S. Lee, and M. Fazel, *Structured learning of Gaussian graphical models*, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, 2012, pp. 629–637.
[36] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[37] C. Oberlin and S. J. Wright, *Active set identification in nonlinear programming*, SIAM J. Optim., 17 (2006), pp. 577–605.
[38] P. A. Olsen, F. Oztoprak, J. Nocedal, and S. J. Rennie, *Newton-like methods for sparse inverse covariance estimation*, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, 2012, pp. 755–763.
[39] K. Scheinberg, S. Ma, and D. Goldfarb, *Sparse inverse covariance selection via alternating linearization methods*, in Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), 2010, pp. 2102–2109.
[40] K. Scheinberg and I. Rish, *Sinco—A Greedy Coordinate Ascent Method for Sparse Inverse Covariance Selection Problem*, Technical Report IBM RC24837, 2009.
[41] K. Scheinberg and X. Tang, *Practical inexact proximal quasi-Newton method with global complexity analysis*, arXiv:1311.6547v3, 2014.
[42] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[43] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, J. Roy. Statist. Soc. Ser. B, 67 (2005), pp. 91–108.
[44] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.
[45] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, *Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain*, Neuroimage, 15 (2002), pp. 273–289.

[46] C. Wang, D. Sun, and K.-C. Toh, *Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm*, SIAM J. Optim., 20 (2010), pp. 2994–3013.

[47] K. Wang, M. Liang, L. Wang, L. Tian, X. Zhang, K. Li, and T. Jiang, *Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study*, Human Brain Mapping, 28 (2007), pp. 967–978.

[48] M. W. Weinter et al., *Alzheimer's Disease Neuroimaging Initiative*, joint project and data-sharing site, http://adni.loni.ucla.edu/.

[49] D. M. Witten, J. H. Friedman, and N. Simon, *New insights and faster computations for the graphical lasso*, J. Comput. Graph. Statist., 20 (2011), pp. 892–900.

[50] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.

[51] S. Yang, L. Yuan, Y. C. Lai, X. Shen, P. Wonka, and J. Ye, *Feature grouping and selection over an undirected graph*, in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2012, pp. 922–930.

[52] G. X. Yuan, C. H. Ho, and C. J. Lin, *An improved GLMNET for $l1$-regularized logistic regression*, J. Mach. Learning Res., 13 (2012), pp. 1999–2030.

[53] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Roy. Statist. Soc. Ser. B , 68 (2006), pp. 49–67.

[54] M. Yuan and Y. Lin, *Model selection and estimation in the Gaussian graphical model*, Biometrika, 94 (2007), pp. 19–35.

[55] X. Yuan, *Alternating direction method for covariance selection models*, J. Sci. Comput., 51 (2012), pp. 261–273.

[56] S. Zhou, J. Lafferty, and L. Wasserman, *Time varying undirected graphs*, in Proceedings of the 21st Annual Conference on Learning Theory (COLT), Helsinki, 2008, pp. 295–319.