FUSED MULTIPLE GRAPHICAL LASSO

SEN YANG[*], ZHAOSONG LU[†], XIAOTONG SHEN[‡], PETER WONKA[*], 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

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 [11, 17, 44]. 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 [5]. 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, 47]. In the framework of 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, 47].

In recent years many research efforts have focused on estimating the precision matrix and the corresponding graphical model (see, for example [2, 10, 16, 17, 22, 23, 25, 26, 29, 30, 33, 47]. Meinshausen and Bühlmann [30] estimated edges for each node in the graph by fitting a lasso problem [36] using the remaining variables as predictors. Yuan and Lin [47] and Banerjee et al. [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 et al. [8] and Lu [25, 26] studied Nesterov's smooth gradient methods [32] for solving this problem or its dual. Banerjee et al. [2] and Friedman et al. [10] proposed block coordinate ascent methods for solving the dual problem. The latter method [10] is

---

[*]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona 85287, USA (senyang@asu.edu, peter.wonka@asu.edu, jieping.ye@asu.edu).

[†]Department of Mathematics, Simon Frasor University, Burnaby, BC, V5A 156, Canada (zhaosong@sfu.ca).

[‡]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA (xshen@umn.edu).

widely referred to as Graphical lasso (GLasso). Mazumder and Hastie [29] proposed a new algorithm called DP-GLasso, each step of which is a box-constrained QP problem. Scheinberg and Rish [35] proposed a coordinate descent method for solving this model in a greedy approach. Yuan [48] and Scheinberg et al. [34] applied alternating direction method of multipliers (ADMM) [4] to this problem. Li and Toh [22] and Yuan and Lin [47] proposed to solve this problem using interior point methods. Wang et al. [40], Hsieh et al. [16], Olsen et al. [33], and Dinh et al. [9] studied 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 et al. [42] and Mazumder and Hastie [28] 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 [7, 11, 12, 13, 19, 20, 31, 49]. Guo et al. [11] proposed a method to jointly estimate multiple graphical models using a hierarchical penalty. However, their model is not convex. Honorio and Samaras [13] proposed a convex formulation to estimate multiple graphical models using the $\ell_{1,\infty}$ regularizer. Hara and Washio [12] introduced a method to learn common substructures among multiple graphical models. Danaher et al. [7] 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 eigen decompositions at each iteration. Mohan et al. [31] proposed to estimate 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 to solve. Recently, a necessary and sufficient condition for multiple graphs to be decomposable was proposed in [7]. However, such necessary and sufficient condition was restricted to two graphs only when the fused penalty is used. It is not clear whether this screening rule can be extended to the more general case with more than two graphs, which is the case in brain network modeling.

There are two types of fused penalties that can be used for estimating multiple (more than two) graphs: (a) pairwise fused or (b) sequential fused [37]. 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 con-

nections, 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 drive the necessary and sufficient condition. Based on this crucial property of FMGL, we develop a screening rule which enables the efficient estimation of large multiple precision matrices for FMGL. The proposed screening rule can be combined with any algorithms to reduce computational cost. We employ a second-order method [16, 21, 38] to solve the fused multiple graphical lasso, where each step is solved by the spectral projected gradient method [27, 43]. In addition, we propose a shrinking 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 $\langle \mathbf{X}, \mathbf{Y} \rangle := \text{tr}(\mathbf{X}\mathbf{Y}^T)$, where $\text{tr}(\cdot)$ denotes the trace of a matrix. $\mathbf{X} \circ \mathbf{Y}$ and $\mathbf{X} \otimes \mathbf{Y}$ means the Hadamard and Kronecker product of $\mathbf{X}$ and $\mathbf{Y}$, respectively. 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 $\mathbf{X} \in \Re^{n \times n}$, $\text{diag}(\mathbf{X})$ denotes the vector formed by the diagonal of $\mathbf{X}$, that is, $\text{diag}(\mathbf{X})_i = \mathbf{X}_{ii}$ for $i = 1, \ldots, n$. $\text{Diag}(\mathbf{X})$ is the diagonal matrix which shares the same diagonal as $\mathbf{X}$. $\text{vec}(\mathbf{X})$ 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 fused multiple graphical lasso formulation 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, $x^{(k)} \in \mathcal{R}^{n_k \times p}$, $k = 1, \ldots, K$ 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 $n_1 = \ldots = n_K = n$. 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

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

Clearly, minimizing (2.1) leads to the maximum likelihood estimate (MLE) $\hat{\mathbf{\Theta}}^{(k)} = (\mathbf{S}^{(k)})^{-1}$. 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, 10, 46]. 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:

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

where

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

$\lambda_1 > 0$ 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 $\mathrm{diag}(\mathbf{S}^{(k)}) > 0, k = 1, \dots, K$. Recall that $\mathbf{S}^{(k)}$ is a sample covariance matrix, and hence $\mathrm{diag}(\mathbf{S}^{(k)}) \geq 0$. The diagonal entries may be not, however, strictly positive. But we can always add a small perturbation (say 10^{-8}) to ensure the above assumption holds.

The following theorem shows that under this assumption the FMGL (2.2) has a unique solution.

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

To prove Theorem 2.1, we first establish a technical lemma which regards the existence of a solution for a standard graphical lasso problem.

LEMMA 2.2. *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*

$$(2.3) \qquad \min_{\mathbf{X} \succ 0} \underbrace{-\log \det(\mathbf{X}) + \mathrm{tr}(\mathbf{S}\mathbf{X}) + \sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}|}_{f(\mathbf{X})}.$$

Then the following statements hold:

    (a) Problem (2.3) has a unique optimal solution;

    (b) The sub-level 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 (2.3).*

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

$$(2.4) \qquad \max_{\mathbf{U} \in \mathcal{U}} \left\{ \log \det(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) : \mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U} \succ 0 \right\}.$$

We first claim that the feasible region of problem (2.4) 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

$$\max_{\mathbf{U} \in \mathcal{U}} \lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) = \max_{t, \mathbf{U} \in \mathcal{U}} \{ t : \mathbf{\Lambda} \circ \mathbf{U} + \mathbf{S} - t\mathbf{I} \succeq 0 \},$$

$$= \min_{\mathbf{X} \succeq 0} \max_{t, \mathbf{U} \in \mathcal{U}} \left\{ t + \mathrm{tr}(\mathbf{X}(\mathbf{\Lambda} \circ \mathbf{U} + \mathbf{S} - t\mathbf{I})) \right\},$$

(2.5)
$$= \min_{\mathbf{X} \succeq 0} \left\{ \mathrm{tr}(\mathbf{S}\mathbf{X}) + \sum_{ij} \mathbf{\Lambda}_{ij} |\mathbf{X}_{ij}| : \ \mathrm{tr}(\mathbf{X}) = 1 \right\},$$

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 $\mathrm{tr}(\mathbf{S}\mathbf{X}) \geq 0$ 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,

(2.6)
$$\mathrm{tr}(\mathbf{S}\mathbf{X}) + \sum_{ij} \mathbf{\Lambda}_{ij} |\mathbf{X}_{ij}| \ > \ 0, \ \forall \mathbf{X} \in \Omega.$$

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 $\mathbf{X} \in \Omega$,

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

Hence, (2.6) again holds. Combining (2.5) with (2.6), one can see that $\max_{\mathbf{U} \in \mathcal{U}} \lambda_{\min}(\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}) > 0$. Therefore, problem (2.4) has at least a feasible solution.

We next show that problem (2.4) has an optimal solution. Let $\bar{\mathbf{U}}$ be a feasible point of (2.4), 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 (2.4) 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 is problem (2.4).

Finally we show that $\mathbf{X}^* = (\mathbf{S} + \mathbf{\Lambda} \circ \mathbf{U}^*)^{-1}$ is the unique optimal solution of (2.3), where $\mathbf{U}^*$ is an optimal solution of (2.4). 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 (2.4) 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 (2.3). 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 (2.3) and moreover it is unique due to the strict convexity of $-\log\det(\cdot)$.

(b) By statement (a), problem (2.3) has a finite optimal value f^*. Hence, the above sub-level set $\mathcal{L}$ is nonempty. We can observe that for any $\mathbf{X} \in \mathcal{L}$,

$$\frac{1}{2}\sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}| = \ f(\mathbf{X}) - \underbrace{[-\log\det(\mathbf{X}) + \mathrm{tr}(\mathbf{SX}) + \frac{1}{2}\sum_{ij} \mathbf{\Lambda}_{ij}|\mathbf{X}_{ij}|]}_{\underline{f}(\mathbf{X})},$$

(2.7) $$\leq \alpha - \underline{f}^*,$$

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 (2.7) 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 $\det(\mathbf{X}) \leq \mathbf{X}_{11}\mathbf{X}_{22}\cdots\mathbf{X}_{pp}$ 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 $\mathbf{X} \in \mathcal{L}$,

$$\sum_i -\log(\mathbf{X}_{ii}) + (\mathbf{S}_{ii} + \Lambda_{ii})\mathbf{X}_{ii} \leq f(\mathbf{X}) - \sum_{i \neq j}(\mathbf{S}_{ij}\mathbf{X}_{ij} + \Lambda_{ij}|\mathbf{X}_{ij}|),$$

(2.8) $$\leq \alpha - \sum_{i \neq j}(\mathbf{S}_{ij}\mathbf{X}_{ij} + \Lambda_{ij}|\mathbf{X}_{ij}|) \ \leq \ \delta$$

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

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

for all i. This relation together with (2.8) implies that for every $\mathbf{X} \in \mathcal{L}$ and all i,

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

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 $\lambda_{\min}(\mathbf{X}) \geq \nu$ for all $\mathbf{X} \in \mathcal{L}$. Hence, one has

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

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. $\square$

We are now ready to prove Theorem 2.1.

Proof. Since $\lambda_1 > 0$ and $\mathrm{diag}(\mathbf{S}^{(k)}) > 0, k = 1,\ldots,K$, it follows from Lemma 2.2 that there exists some δ such that for each $k = 1,\ldots,K$,

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

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

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

for $k = 1, \ldots, K$, where $\bar{\delta} = h(\bar{\boldsymbol{\Theta}}) - (K-1)\delta$. 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

$$
\text{(2.9)} \qquad\qquad \min_{\boldsymbol{\Theta} \in \bar{\Omega}} h(\boldsymbol{\Theta}).
$$

In view of Lemma 2.2, 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 (2.9) has a unique optimal solution and so is problem (2.2). $\square$

3. The screening rule for fused multiple graphical lasso. Due to the presence of the log determinant, it is challenging to solve the formulations involving the penalized log-likelihood efficiently. The existing methods for single graphical lasso are not scalable to the problems with a large amount of features because of the high computational complexity. Recent studies have shown that the graphical model may contain many connected components, which are disjoint with 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 et al. [7] 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 fused multiple graphical lasso 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 arbitrary number of graphs.

Let $C_1, \ldots, C_L$ be a partition of the p features into L non-overlapping sets, with $C_l \cap C_{l'} = \emptyset$, $\forall l \neq l'$ and $\bigcup_{l=1}^{L} C_l = \{1, \ldots, p\}$. We say that the solution $\widehat{\boldsymbol{\Theta}}$ 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

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

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 $\widehat{\boldsymbol{\Theta}}$ 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* $\widehat{\boldsymbol{\Theta}}_l = (\widehat{\boldsymbol{\Theta}}_l^{(1)}, \ldots, \widehat{\boldsymbol{\Theta}}_l^{(K)})$ *for* $l = 1, \ldots, L$. *Then there holds:*

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

where $\boldsymbol{\Theta}_l^{(k)}$ *and* $\mathbf{S}_l^{(k)}$ *are the* $|C_l| \times |C_l|$ *symmetric submatrices of* $\boldsymbol{\Theta}^{(k)}$ *and* $\mathbf{S}^{(k)}$ *corresponding to the* l-*th diagonal block, respectively, for* $k = 1, \ldots, K$, *and* $\boldsymbol{\Theta}_l = (\boldsymbol{\Theta}_l^{(1)}, \ldots, \boldsymbol{\Theta}_l^{(K)})$ *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 FMGL to be block diagonal with L blocks $C_l, l = 1, \ldots, L$, which is a key for developing 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* $\widehat{\boldsymbol{\Theta}}^{(k)}, k = 1, \ldots, K$ *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:*

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

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

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 about 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 cheaper. 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 $I \subseteq \{1, \cdots, n\}$ and $J \subseteq \{1, \cdots, n-1\}$, let $\bar{I}$ and $\bar{J}$ be the complement of I and J with respect to $\{1, \cdots, n\}$ and $\{1, \cdots, n-1\}$, respectively. Define*

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

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*

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

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 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

$$(3.6) \qquad d_I \geq 0,\ d_{\bar{I}} \leq 0,\ d_J - d_{J+1} \geq 0,\ d_{\bar{J}} - d_{\bar{J}+1} \leq 0,$$

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

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

One can immediately observe that

$$(3.8) \qquad d_{m_i} = \cdots = d_{n_i} = 0, \quad d_j - d_{j+1} = 0, \quad m_i \leq j \leq n_i - 1,\ 1 \leq i \leq k.$$

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 the active inequalities given in (3.8), all other possible active inequalities at d are

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

(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 $n - k + 1 \geq n - 1$, which implies $k \leq 2$. 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 a similar argument as in case (a), one can see that the number of independent active inequalities at d is at most

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

Similarly 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 a similar argument as in case (a), we see that the number of independent active inequalities at d is at most

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

Similarly 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 a similar argument as 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 $k \geq 1$ 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 $\text{ext}(P_{I,J}) \subseteq Q$. $\square$

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

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

Proof. It follows from Lemma 3.3 (iii) that

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

We next show that

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

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 d_i's are 0. If $\alpha > 0$, it is not hard to see that $d \in \text{ext}(P_{I,J})$ with $I = \{1, \cdots, n\}$ and $J = \{m_1, \cdots, n-1\}$. Similarly, if $\alpha < 0$, $d \in \text{ext}(P_{I,J})$ with $I = \emptyset$ and J being the complement of $\bar{J} = \{m_1, \cdots, n-1\}$. Hence, $d \in \cup \{\text{ext}(P_{I,J}) : I \subseteq \{1, \cdots, n\}, \ J \subseteq \{1, \cdots, n-1\}\}$. $\square$

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:

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

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

Proof. Let $P_{I,J}$ be defined in (3.4) for any $I \subseteq \{1, \ldots, n\}$ and $J \subseteq \{1, \ldots, n-1\}$. We observe that
  (a) $\Re^n = \cup \{P_{I,J} : \ I \subseteq \{1, \ldots, n\}, \ J \subseteq \{1, \ldots, n-1\}\}$;
  (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 $I \subseteq \{1, \ldots, n\}$ and $J \subseteq \{1, \ldots, n-1\}$;
  (c) $f(y)$ is a linear function of y when restricted to the set $P_{I,J}$ for every $I \subseteq \{1, \ldots, n\}$ and $J \subseteq \{1, \ldots, n-1\}$.
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 $\text{ext}(P_{I,J})$. 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 $d \in \text{ext}(P_{I,J})$. 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

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

which together with the definition of f immediately implies that the conclusion of this lemma holds. $\square$

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

(3.10)
$$\begin{cases} x_1 + \lambda_1 \gamma_1 + \lambda_2 v_1 = 0, \\ x_i + \lambda_1 \gamma_i + \lambda_2 (v_i - v_{i-1}) = 0, \ 2 \leq i \leq n-1, \\ x_n + \lambda_1 \gamma_n - \lambda_2 v_{n-1} = 0, \\ -1 \leq \gamma_i \leq 1, \ i = 1, \dots, n, \\ -1 \leq v_i \leq 1, \ i = 1, \dots, n-1 \end{cases}$$

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

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

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

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

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

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

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

which is equivalent to

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

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. $\square$

We are now ready to prove Theorem 3.2.

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

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

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

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

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

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

Necessity: Suppose that $\widehat{\Theta}^{(k)}, k = 1, \ldots, K$ 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 $\widehat{\Theta}^{(k)}$. Hence, $\widehat{\mathbf{W}}_{ij}^{(k)} = \widehat{\Theta}_{ij}^{(k)} = 0$ 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 $(\gamma_{ij}^{(k)}, v_{ij}^{(k,k+1)}), k = 1, \ldots, K-1$ and $\gamma_{ij}^{(K)}$ such that

$$
\begin{aligned}
& \mathbf{S}_{ij}^{(1)} + \lambda_1 \gamma_{ij}^{(1)} + \lambda_2 v_{ij}^{(1,2)} = 0, \\
& \mathbf{S}_{ij}^{(k)} + \lambda_1 \gamma_{ij}^{(k)} + \lambda_2(-v_{ij}^{(k-1,k)} + v_{ij}^{(k,k+1)}) = 0, \ 2 \leq k \leq K-1, \\
& \mathbf{S}_{ij}^{(K)} + \lambda_1 \gamma_{ij}^{(K)} - \lambda_2 v_{ij}^{(K-1,K)} = 0, \\
& -1 \leq \gamma_{ij}^{(k)} \leq 1, \ 1 \leq k \leq K, \\
& -1 \leq v_{ij}^{(k,k+1)} \leq 1, \ \ 1 \leq k \leq K-1.
\end{aligned}
$$
(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 $(\gamma_{ij}^{(k)}, v_{ij}^{(k,k+1)}), k = 1, \ldots, K-1$ and $\gamma_{ij}^{(K)}$ such that (3.17) holds. Now let $\widehat{\Theta}^{(k)}, k = 1, \ldots, K$ be a block diagonal matrix as defined in (3.1) with $\mathbf{U} = \mathbf{I}$, where $\widehat{\Theta}_l = (\widehat{\Theta}_l^{(1)}, \ldots, \widehat{\Theta}_l^{(K)})$ is given by (3.2) for $l = 1, \ldots, L$. Also, let $\widehat{\mathbf{W}}^{(k)}$ be the inverse of $\widehat{\Theta}^{(k)}$ for $k = 1, \ldots, K$. Since $\widehat{\Theta}_l$ 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 $\widehat{\Theta}_{ij}^{(k)} = \widehat{\mathbf{W}}_{ij}^{(k)} = 0$ 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 $\widehat{\Theta}^{(k)}, k = 1, \ldots, K$ is an optimal solution of problem (2.2). In addition, $\widehat{\Theta}^{(k)}, k = 1, \ldots, K$ is block diagonal with L known blocks $C_l, l = 1, \ldots, L$. The conclusion thus holds. □

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 estimated precision matrix $\Theta_l^{(k)}$ by solving the FMGL (2.2) with $\mathbf{S}^{(k)}$ replaced by $\mathbf{S}_l^{(k)}$. In this section, we discuss 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 $\Theta := (\Theta^{(1)}, \ldots, \Theta^{(K)})$ and use t to denote the Newton iteration index. Let $\Theta_t = (\Theta_t^{(1)}, \ldots, \Theta_t^{(K)})$ be the approximate solution obtained at the t-th Newton iteration.

The optimization problem (2.2) can be rewritten as

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

where

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

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

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

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

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

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

$$(4.3) \qquad \mathbf{D} = \bar{\boldsymbol{\Theta}}_{t+1} - \boldsymbol{\Theta}_t.$$

We shall mention that the subproblem (4.2) can be suitably solved by the non-monotone spectral projected gradient (NSPG) method (see, for example, [43, 27]). It was shown by Lu and Zhang [27] 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 the proximal subproblems in the form of

$$(4.4) \qquad \min_{\boldsymbol{\Theta}} \frac{1}{2} \sum_{k=1}^{K} \|\boldsymbol{\Theta}^{(k)} - \mathbf{G}^{(k)}\|_F^2 + \alpha P(\boldsymbol{\Theta})$$

for some $\mathbf{G} = (\mathbf{G}^{(1)}, \ldots, \mathbf{G}^{(K)})$ and $\alpha > 0$. By the definition of $P(\boldsymbol{\Theta})$, it is not hard to see that problem (4.4) can be decomposed into a set of independent and smaller sized problems

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

for all $i \geq j$, $j = 1, \ldots, p$, where $(\alpha_1, \alpha_2) = \alpha(\lambda_1, \lambda_2)$. The problem (4.5) is known as the fused lasso signal approximator, which can be solved very efficiently and exactly [6, 24]. In addition, they 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 the 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 $\boldsymbol{\Theta}_{t+1}^{(k)} = \boldsymbol{\Theta}_t^{(k)} + \beta\mathbf{D}^{(k)}, k = 1, \ldots, K$. In the context of the standard (single) graphical lasso,

Hsieh et al. [16] 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).

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*

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

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

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

Proof. Let $\tilde{\beta} = 1/\max\{\|(\mathbf{\Theta}_t^{(k)})^{-1}\mathbf{D}^{(k)}\|_2 : k = 1, \ldots, K\}$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix. Since $\mathbf{D} \neq 0$ and $\mathbf{\Theta}_t^{(k)} \succ 0, k = 1, \ldots, K$, we see that $\tilde{\beta} > 0$. Moreover, we have for all $0 < \beta < \tilde{\beta}$ and $k = 1, \ldots, K$,

$$(\mathbf{\Theta}_t^{(k)})^{-\frac{1}{2}}\left(\mathbf{\Theta}_t^{(k)} + \beta \mathbf{D}^{(k)}\right)(\mathbf{\Theta}_t^{(k)})^{-\frac{1}{2}} \quad = \quad \mathbf{I} + \beta (\mathbf{\Theta}_t^{(k)})^{-\frac{1}{2}}\mathbf{D}^{(k)}(\mathbf{\Theta}_t^{(k)})^{-\frac{1}{2}}$$

$$\succeq \quad (1 - \beta \|(\mathbf{\Theta}_t^{(k)})^{-1}\mathbf{D}^{(k)}\|_2)\mathbf{I} \succ 0.$$

By the definition of $\mathbf{D}$ and (4.2), one can easily show that

$$\delta \leq -\sum_{k=1}^{K} \mathrm{tr}(\mathbf{W}_t^{(k)}\mathbf{D}^{(k)}\mathbf{W}_t^{(k)}\mathbf{D}^{(k)}),$$

which together with the fact that $\mathbf{W}_t^{(k)} \succ 0, k = 1, \ldots, K$ and $\mathbf{D} \neq 0$ implies that $\delta < 0$. Using differentiability of f_k, convexity of P, and the definition of δ, we obtain that for all sufficiently small $\beta > 0$,

$$F(\mathbf{\Theta}_t + \beta \mathbf{D}) - F(\mathbf{\Theta}_t) = \sum_{k=1}^{K}(f_k(\mathbf{\Theta}_t^{(k)} + \beta \mathbf{D}^{(k)}) - f_k(\mathbf{\Theta}_t^{(k)})) + P(\mathbf{\Theta}_t + \beta \mathbf{D}) - P(\mathbf{\Theta}_t),$$

$$= \sum_{k=1}^{K} \mathrm{tr}((\mathbf{S}^{(k)} - \mathbf{W}_t^{(k)})\mathbf{D}^{(k)})\beta + o(\beta) + P(\beta(\mathbf{\Theta}_t + \mathbf{D}) + (1 - \beta)\mathbf{\Theta}_t) - P(\mathbf{\Theta}_t),$$

$$\leq \sum_{k=1}^{K} \mathrm{tr}((\mathbf{S}^{(k)} - \mathbf{W}_t^{(k)})\mathbf{D}^{(k)})\beta + o(\beta) + \beta P(\mathbf{\Theta}_t + \mathbf{D}) + (1 - \beta)P(\mathbf{\Theta}_t) - P(\mathbf{\Theta}_t),$$

$$\leq \beta\delta + o(\beta).$$

This inequality together with $\delta < 0$ and $\sigma \in (0, 1)$ implies that there exists $\hat{\beta} > 0$ such that for all $\beta \in (0, \hat{\beta})$, $F(\mathbf{\Theta}_t + \beta \mathbf{D}) - F(\mathbf{\Theta}_t) \leq \sigma \beta \delta$. It then follows that the conclusion of this lemma holds for $\bar{\beta} = \min\{\tilde{\beta}, \hat{\beta}\}$. □

By virtue of Lemma 4.1, we can adopt the well-known Armijo's backtracking line search rule [38] 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 $\mathbf{\Theta}_t^{(k)} + \beta \mathbf{D}^{(k)}, k = 1, \ldots, K$ [16]. 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 $\mathbf{\Theta}_t^{(k)} + \beta \mathbf{D}^{(k)}$.

---

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

4.1. Shrinking scheme. Given the large number of unknown variables in (4.2), it is advantageous to minimize (4.2) in a reduced space. The issue now is how to identify the 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. [16] 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, J_{free} and J_{fixed}, based on the gradient at the start of each Newton iteration, and then the minimization is only performed on the variables in J_{free}. We call this technique "shrinking" in this paper. Due to the sparsity of the precision matrix, the size of J_{free} is usually much smaller than p^2. Moreover, it has been shown in the single graph case that the size of J_{free} will decrease quickly [16]. The shrinking technique can thus improve the computational efficiency. This technique was also successfully used in [18, 33, 45]. We show that shrinking can be extended to the fused multiple graphical lasso based on the results established in Section 3.

Denote the gradient of f_k at t-th 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 fixed set J_{fixed} as*

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

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

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

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

$$(4.8) \qquad \min_{\mathbf{D}} Q_t(\mathbf{\Theta}_t + \mathbf{D}) \text{ such that } \mathbf{D}_{ij}^{(1)} = \ldots = \mathbf{D}_{ij}^{(K)} = 0, (i,j) \notin J_{fixed}.$$

Proof. Consider problem (4.8), which can be reformulated to

$$
(4.9) \quad \begin{aligned}
\min_{\mathbf{D}} \quad & \sum_{k=1}^{K} \left(\tfrac{1}{2} \text{vec}(\mathbf{D}^{(k)})^T \mathbf{H}_t^{(k)} \text{vec}(\mathbf{D}^{(k)}) + \text{vec}(\widetilde{\mathbf{G}}_t^{(k)})^T \text{vec}(\mathbf{D}^{(k)}) \right) \\
& + P(\mathbf{\Theta}_t + \mathbf{D}), \\
\text{s.t.} \quad & \mathbf{D}_{ij}^{(1)} = \ldots = \mathbf{D}_{ij}^{(K)} = 0, (i,j) \notin J_{fixed},
\end{aligned}
$$

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

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

Since $\sum_{k=1}^{K} \text{vec}(\mathbf{D}^{(k)})^T \mathbf{H}_t^{(k)} \text{vec}(\mathbf{D}^{(k)}) \geq 0$, the optimal solution of (4.8) is given by $\mathbf{D}^{(1)} = \ldots = \mathbf{D}^{(K)} = 0$. □

Lemma 4.2 provides a shrinking scheme to partition the variables into the free set J_{free} and the fixed set J_{fixed}. With shrinking, each Newton step of the proposed second-order method falls into a block coordinate gradient descent framework [38]. Lemma 4.2 shows that when the variables in the free set J_{free} are fixed, no update is needed for the variables in the fixed set J_{fixed}. Minimization of (4.2) restricted to the free set can therefore guarantee the convergence to the unique optimal solution [16, 38]. In addition, it has been shown that local quadratic convergence rate can be achieved when the exact Hessian is used (see, for example, [16, 21]).

The resulting second-order method for solving the fused multiple graphical lasso is summarized in Algorithm 1.

Algorithm 1: Proposed second-order method for Fused Multiple Graphical Lasso (FMGL)

Input: $\mathbf{S}^{(k)}, k = 1, \ldots, K, \lambda_1, \lambda_2$
Output: $\mathbf{\Theta}^{(k)}, k = 1, \ldots, K$
Initialization: $\mathbf{\Theta}_0^{(k)} = (\text{diag}(\mathbf{S}^{(k)}))^{-1}$;
while *Not Converged* **do**
> Determine the sets of free and fixed indices J_{free} and J_{fixed} 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 J_{free}.
> 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 $\mathbf{\Theta}^{(k)}, k = 1, \ldots, K$;

---

5. Experimental results. In this section, we evaluate the proposed algorithm and screening rule on synthetic datasets and two real datasets: ADHD-200[2] and FDG-PET images[3]. The experiments are performed on a PC with quad-core Intel 2.67GHz CPU and 9GB memory.

5.1. Simulation.

5.1.1. Efficiency. We conduct experiments to demonstrate the effectiveness of the proposed screening rule and the efficiency of our method FMGL. The following algorithms are included in our comparisons:
- FMGL: the proposed second-order method in Algorithm 1.
- ADMM: ADMM method.
- FMGL-S: FMGL with screening.
- ADMM-S: ADMM with screening.

Both FMGL and ADMM are written in Matlab. Since both methods involve solving (4.4) which involves a double loop, we implement the sub-routine 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 $\boldsymbol{\Theta}^{(k)}$ with L blocks, and each block $\boldsymbol{\Theta}_l^{(k)}$ is of size $(p/L) \times (p/L)$. Each $\boldsymbol{\Theta}_l^{(k)}, l = 1, \ldots, L, k = 1, \ldots, K$ has random sparsity structures. We control the number of nonzeros in each $\boldsymbol{\Theta}_l^{(k)}$ 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 $10Kp$. We terminate the NSPG in the FMGL when the relative error $\frac{\max\{\|\boldsymbol{\Theta}_r^{(k)} - \boldsymbol{\Theta}_{r-1}^{(k)}\|_\infty\}}{\max\{\|\boldsymbol{\Theta}_{r-1}^{(k)}\|_\infty\}} \leq 1e\text{-}6$. The FMGL is terminated when the relative error of the objective value is smaller than $1e\text{-}5$, and ADMM stops until it achieves an objective value equal to or smaller than that of FMGL. The results presented in Table 5.1 show that FMGL is consistently faster than ADMM. FMGL converges much more quickly than 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.

TABLE 5.1

Comparison of the proposed FMGL and ADMM 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. $\|\boldsymbol{\Theta}\|_0$ represents the total number of nonzero entries in ground truth precision matrices $\boldsymbol{\Theta}^{(k)}, k = 1, \ldots, K$, and $\|\boldsymbol{\Theta}^*\|_0$ is the number of nonzeros in the solution.

Data and parameter setting						Computational time (iteration numbers)			
p	K	L	$\|\boldsymbol{\Theta}\|_0$	λ_1	$\|\boldsymbol{\Theta}^*\|_0$	FMGL-S	FMGL	ADMM-S	ADMM
500	2	5	9766	0.08	10228	**0.86**	11.19 (6)	13.78	98.89 (152)
1000			19832	0.088	19322	**6.21**	57.78 (6)	58.75	529.36 (140)
500	5		24494	0.055	23878	**2.62**	33.15 (6)	34.85	256.33 (146)
1000			50836	0.054	44724	**14.70**	197.53 (6)	171.68	1431.91 (150)
500	10		49500	0.051	45756	**6.01**	70.87 (6)	73.42	524.84 (152)
1000			100292	0.046	86774	**30.49**	383.46 (6)	357.34	2991.16 (155)
500	2	10	9528	0.07	9884	**0.81**	16.25 (7)	5.01	109.24 (155)
1000			19658	0.08	20612	**1.65**	75.44 (6)	25.89	560.75 (155)
500	5		23562	0.055	23600	**1.69**	47.32 (7)	11.17	261.00 (153)
1000			49274	0.054	46582	**5.72**	207.97 (6)	75.61	1661.24 (172)
500	10		47364	0.051	48360	**3.70**	103.54 (6)	24.75	552.09 (157)
1000			98650	0.046	96216	**12.16**	409.94 (6)	150.62	3192.02 (168)

5.1.2. Stability. We conduct experiments to demonstrate the effectiveness of FMGL. The synthetic sparse precision matrices are generated in the following way: we set the first precision matrix $\boldsymbol{\Theta}^{(1)}$ as $0.25I_{p \times p}$, where $p = 100$. When adding an edge (i, j) in the graph, we add σ to $\theta_{ii}^{(1)}$ and $\theta_{jj}^{(1)}$, and subtract σ from $\theta_{ij}^{(1)}$ and $\theta_{ji}^{(1)}$ to keep the positive definiteness of $\boldsymbol{\Theta}^{(1)}$, where σ is uniformly drawn from $[0.1, 0.3]$. When deleting an edge (i, j) from the graph, we reverse the above steps with $\sigma = \theta_{ij}^{(1)}$. We randomly assign 200 edges for $\boldsymbol{\Theta}^{(1)}$. $\boldsymbol{\Theta}^{(2)}$ is obtained by adding 25 edges and deleting 25 different edges from $\boldsymbol{\Theta}^{(1)}$. $\boldsymbol{\Theta}^{(3)}$ is obtained from $\boldsymbol{\Theta}^{(2)}$ in the same way. For each precision matrix, we randomly draw n samples from the Gaussian distribution with the corresponding precision matrix, where n varies from 40 to 200 with a step of 20. We perform 500 replications for each n. For each n,

λ_2 is fixed to 0.08, and λ_1 is adjusted to make sure that the edge number is about 200. The accuracy n_d/n_g is used to measure the performance of FMGL and GLasso, where n_d is the number of true edges detected by FGML and GLasso, and n_g is the number of true edges. The results are shown in Figure 5.1. We can see from the figure that FMGL achieves higher accuracies, demonstrating the effectiveness of FMGL for learning multiple graphical models simultaneously.
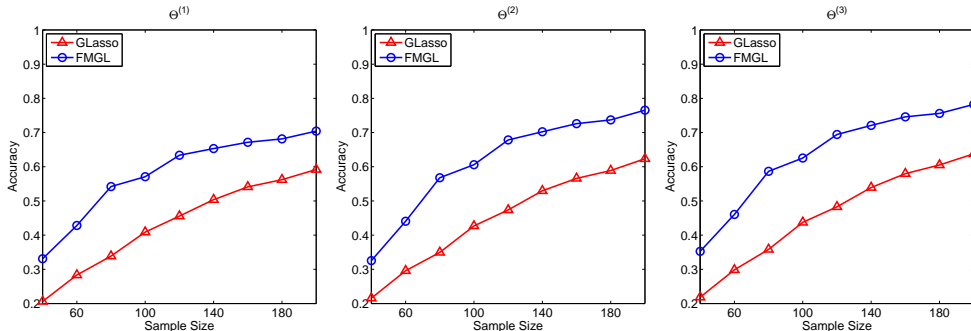

FIG. 5.1. *Comparison of FMGL and GLasso in detecting true edges. Sample size varies from 40 to 200 with a step of 20.*

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 the research on ADHD. The data used in this experiment is preprocessed using the NIAK pipeline, and downloaded from neurobureau[4]. More details about the preprocessing strategy can be found in the same website. The dataset 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 convergence criterion is 1e-5. The comparison of FMGL and ADMM in terms of the objective value curve is shown in Figure 5.2. The result shows that FMGL converges much faster than ADMM. The computational times of FMGL and ADMM are 1557.08 and 8306.35 seconds respectively. However, utilizing the screening, the computational times of FMGL-S and ADMM-S are 18.08 and 119.23 seconds respectively, demonstrating the superiority of the screening rule. The obtained solution has 1443 blocks. The largest one including 634 nodes is shown in Figure 5.3.

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

---

[4]`http://www.nitrc.org/plugins/mwiki/index.php?title=neurobureau:NIAKPipeline/`
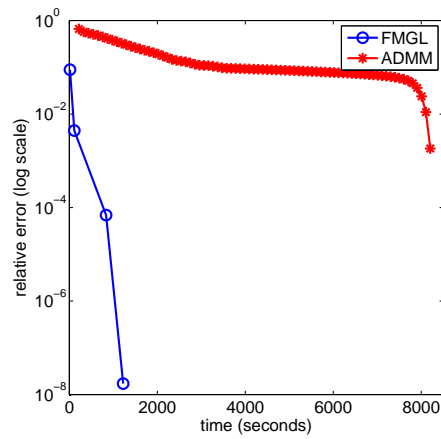
FIG. 5.2. *Comparison of FMGL and ADMM in terms of objective value curve on the ADHD-200 dataset. The dimension p is 2834, and the number of graphs K is 3.*
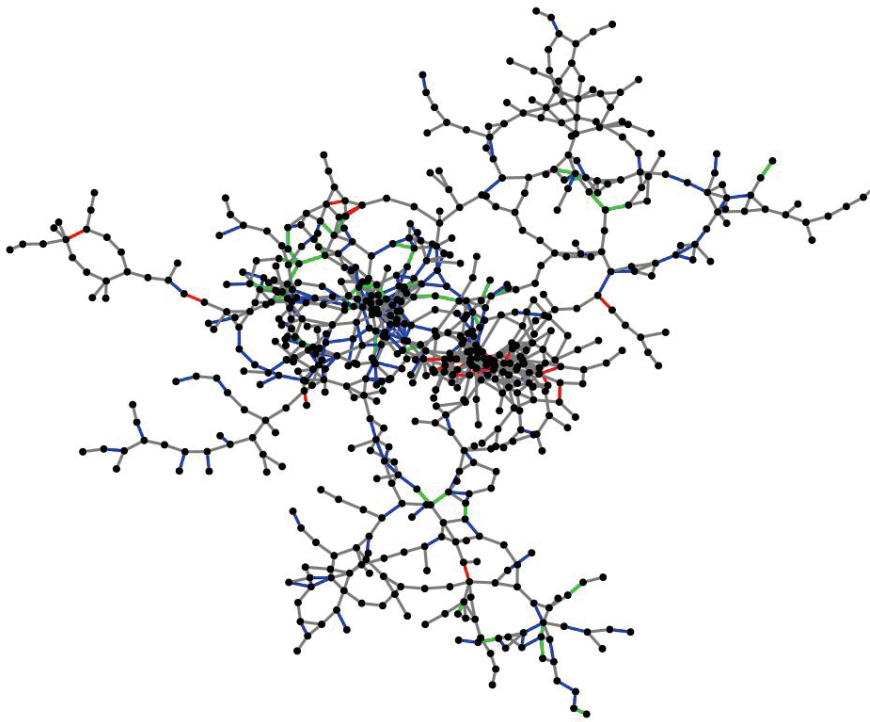

FIG. 5.3. *A subgraph of ADHD-200 identified by FMGL with the proposed screening rule. The grey edges are common edges among the three graphs; the red, green, and blue edges are the specific edges for TDC, ADHD-I, and ADHD-C respectively.*

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 5.4. We can observe that the number of blocks increases, and the size of blocks deceases when the
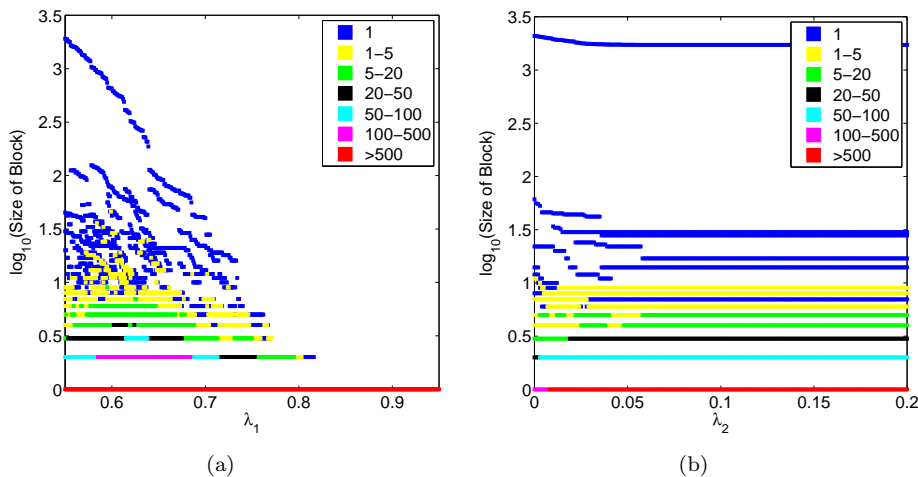
FIG. 5.4. *The size distribution of blocks (in the logarithmic scale) identified by the proposed screening rule. The color represents the number of blocks of a specified size. (a): λ_1 varies from 0.5 to 0.95 with λ_2 fixed to 0.015. (b): λ_2 varies from 0 to 0.2 with λ_1 fixed to 0.55.*

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. The different regions of the whole brain volume can be represented by 116 anatomical volumes of interest (AVOI), defined by Automated Anatomical Labeling (AAL) [39]. 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 10 groups: prefrontal lobe, other parts of the frontal lobe, parietal lobe, occipital lobe, thalamus, insula, temporal lobe, corpus striatum, cerebellum, and vermis. More details about the categories can be found in [39, 41]. 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 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 5.5. 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 5.6. 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 are supported by previous literatures [1, 14]. 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
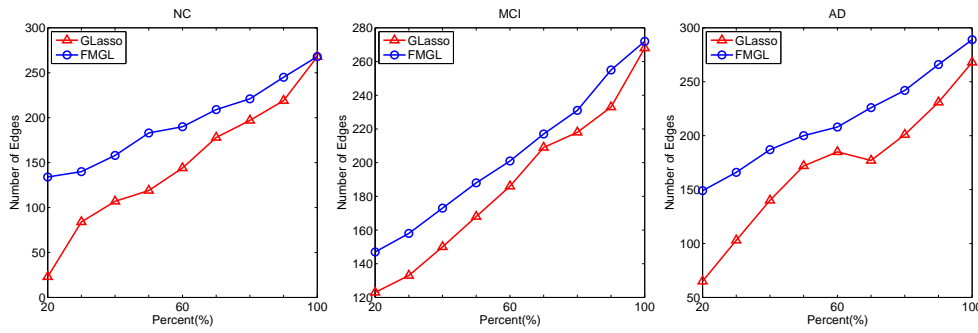
FIG. 5.5. *The average number of stable edges detected by FMGL and GLasso in NC, MCI, and AD of 500 replications. Sample size varies from 20% to 100% with a step of 10%.*
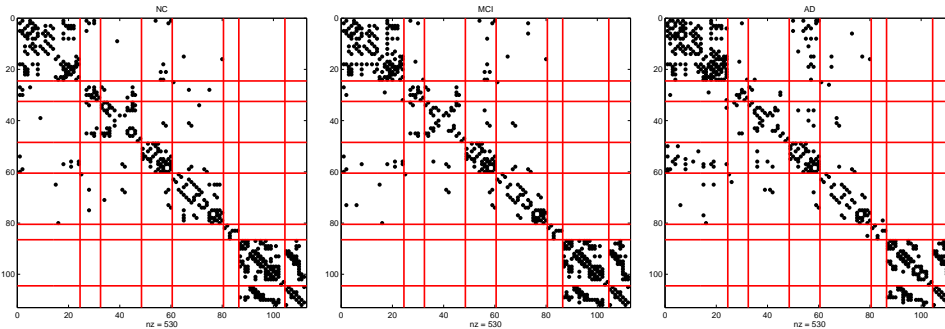

FIG. 5.6. *Brain connection models with 265 edges: NC, MCI, and AD. In each figure, the diagonal blocks are prefrontal lobe, other parts of frontal lobe, parietal lobe, occipital lobe, temporal lobe, corpus striatum, cerebellum, and vermis respectively.*

the meaningful differences.

6. Conclusion. In this paper, we consider 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 fused multiple graphical lasso. The global convergence of the proposed method is guaranteed, and the convergence rate is local quadratic. A shrinking scheme is proposed to identify the variables to be updated during the Newton iterations, thus reduces 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 explore the convergence properties of the second-order method using the inexact Newton direction. Due to the shrinking scheme, the proposed second-order method is suitable for warm-start techniques. A good initial solution can further speedup the computation. As part of the 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 [15].

REFERENCES

[1] NP Azari, SI Rapoport, CL Grady, MB Schapiro, JA 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*, The Journal of Machine Learning Research, 9 (2008), pp. 485–516.

[3] D. Bertsekas and J. N. Tsitsiklis, *Introduction to linear optimization*, (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*, Foundations and Trends® in Machine Learning, 3 (2011), pp. 1–122.

[5] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis*, Molecular systems biology, 3 (2007).

[6] L. Condat, *A direct algorithm for 1d total variation denoising*, Signal Processing Letters, (2013).

[7] P. Danaher, P. Wang, and D.M. Daniela, *The joint graphical lasso for inverse covariance estimation across multiple classes*, Arxiv preprint arXiv:1111.0324v3, (2012).

[8] A. d'Aspremont and L. Banerjee, O.and El Ghaoui, *First-order methods for sparse covariance selection*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 56–66.

[9] Q. Dinh, A. Kyrillidis, and V. Cevher, *A proximal newton framework for composite minimization: Graph learning without cholesky decompositions and matrix inversions*, arXiv preprint arXiv:1301.1459, (2013).

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

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

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

[13] J. Honorio and D. Samaras, *Multi-task learning of gaussian graphical models*, in ICML, 2010.

[14] B. Horwitz, C.L. Grady, NL Schlageter, R. Duara, and SI Rapoport, *Intercorrelations of regional cerebral glucose metabolic rates in alzheimer's disease*, Brain research, 407 (1987), pp. 294–306.

[15] C.J. Hsieh, I. Dhillon, P. Ravikumar, and A. Banerjee, *A divide-and-conquer method for sparse inverse covariance estimation*, in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 2339–2347.

[16] C.J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, *Sparse inverse covariance matrix estimation using quadratic approximation*, Advances in Neural Information Processing Systems (NIPS), 24 (2011).

[17] S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, A. Fleisher, E. Reiman, and J. Ye, *Learning brain connectivity of alzheimers disease from neuroimaging data*, in Advances in Neural Information Processing Systems (NIPS), 2009, pp. 808–816.

[18] T. Joachims, *Making large-scale support vector machine learning practical*, in Advances in kernel methods, MIT Press, 1999, pp. 169–184.

[19] M. Kolar, L. Song, A. Ahmed, and E.P. Xing, *Estimating time-varying networks*, The Annals of Applied Statistics, 4 (2010), pp. 94–123.

[20] M. Kolar and E.P. Xing, *On time varying undirected graphs*, in AISTAT, 2011.

[21] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal newton-type methods for minimizing convex objective functions in composite form*, arXiv preprint arXiv:1206.1623, (2012).

[22] L. Li and K.C. Toh, *An inexact interior point method for l 1-regularized sparse covariance selection*, Mathematical Programming Computation, 2 (2010), pp. 291–315.

[23] H. Liu, K. Roeder, and L. Wasserman, *Stability approach to regularization selection (StARS) for high dimensional graphical models*, in Advances in Neural Information Processing Systems (NIPS), 2011.

[24] J. Liu, L. Yuan, and J. Ye, *An efficient algorithm for a class of fused lasso problems*, in KDD, 2010, pp. 323–332.

[25] Z. Lu, *Smooth optimization approach for sparse covariance selection*, SIAM Journal on Optimization, 19 (2009), pp. 1807–1827.

[26] ——, *Adaptive first-order methods for general sparse inverse covariance selection*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2000–2016.

[27] Z. Lu and Y. Zhang, *An augmented lagrangian approach for sparse principal component analysis*, Mathematical Programming, (2011), pp. 1–45.

[28] R. Mazumder and T. Hastie, *Exact covariance thresholding into connected components for large-scale graphical lasso*, The Journal of Machine Learning Research, 13 (2012), pp. 781–794.

[29] ——, *The graphical lasso: New insights and alternatives*, Electronic Journal of Statistics, 6 (2012), pp. 2125–2149.

[30] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics, 34 (2006), pp. 1436–1462.

[31] K. Mohan, M. Chung, S. Han, D. Witten, S. Lee, and M. Fazel, *Structured learning of gaussian graphical models*, in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 629–637.

[32] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.

[33] P. A. Olsen, F. Oztoprak, J. Nocedal, and S. J. Rennie, *Newton-like methods for sparse inverse covariance estimation*, in Advances in Neural Information Processing Systems (NIPS), 2012.

[34] K. Scheinberg, S. Ma, and D. Goldfarb, *Sparse inverse covariance selection via alternating linearization methods*, in Advances in Neural Information Processing Systems (NIPS), 2010.

[35] K. Scheinberg and I. Rish, *Sinco-a greedy coordinate ascent method for sparse inverse covariance selection problem*, preprint, (2009).

[36] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.

[37] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 91–108.

[38] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.

[39] 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.

[40] C. Wang, D. Sun, and K.C. Toh, *Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm*, SIAM Journal on Optimization, 20 (2010), pp. 2994–3013.

[41] 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.

[42] D.M. Witten, J.H. Friedman, and N. Simon, *New insights and faster computations for the graphical lasso*, Journal of Computational and Graphical Statistics, 20 (2011), pp. 892–900.

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

[44] S. Yang, L. Yuan, Y.C. Lai, X. Shen, P. Wonka, and J. Ye, *Feature grouping and selection over an undirected graph*, in KDD, ACM, 2012, pp. 922–930.

[45] G.X. Yuan, C.H. Ho, and C.J. Lin, *An improved glmnet for l1-regularized logistic regression*, The Journal of Machine Learning Research, 13 (2012), pp. 1999–2030.

[46] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2006), pp. 49–67.

[47] ——, *Model selection and estimation in the gaussian graphical model*, Biometrika, 94 (2007), pp. 19–35.

[48] X. Yuan, *Alternating direction method for covariance selection models*, Journal of Scientific Computing, 51 (2012), pp. 261–273.

[49] S. Zhou, J. Lafferty, and L. Wasserman, *Time varying undirected graphs*, COLT, (2008).