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. This work was supported in part by research grants from the NIH (R01 LM010730) and NSF (IIS-0953662, HH-1421057, and HH-1421100).


†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 $\ell_1$ 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,
block coordinate ascent methods for solving the dual problem. The latter method
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 multi-
ple (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 $\ell_1$ and sequential fused regularization. The $\ell_1$ 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, $\mathbb{R}$ stands for the set of all real numbers, $\mathbb{R}^n$ denotes the $n$-dimensional Euclidean space, and the set of all $m \times n$ matrices with real entries is denoted by $\mathbb{R}^{m \times n}$. All matrices are presented in bold format. The space of symmetric matrices is denoted by $\mathbb{S}^n$. If $X \in \mathbb{S}^n$ is positive semidefinite (resp., definite), we write $X \succeq 0$ (resp., $X \succ 0$). Also, we write $X \succeq Y$ to mean $X - Y \succeq 0$. The cone of positive semidefinite matrices in $\mathbb{S}^n$ is denoted by $\mathbb{S}^n_+$. Given matrices $X$ and $Y$ in $\mathbb{R}^{m \times n}$, the standard inner product is defined by $\langle X, Y \rangle := \text{tr}(XY^T)$, where $\text{tr}(\cdot)$ denotes the trace of a matrix. $X \odot Y$ and $X \otimes Y$ mean the Hadamard and Kronecker product, respectively, of $X$ and $Y$. We denote the identity matrix by $I$, whose dimension should be clear from the context. The determinant and the minimal eigenvalue of a real symmetric matrix $X$ are denoted by $\det(X)$ and $\lambda_{\min}(X)$, respectively. Given a matrix $X \in \mathbb{R}^{m \times n}$, $\text{diag}(X)$ denotes the vector formed by the diagonal of $X$; that is, $\text{diag}(X)_i = X_{ii}$ for $i = 1, \ldots, n$. $\text{Diag}(X)$ is the diagonal matrix which has the same diagonal as $X$. $\text{vec}(X)$ is the vectorization of $X$. In addition, $X > 0$ means that all entries of $X$ 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, $x^{(k)} \in \mathbb{R}^{n_k \times p}$, $k = 1, \ldots, K$, 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 $\Sigma^{(k)}$, and there are many conditionally independent pairs of features; i.e., the precision matrix $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ 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 $S^{(k)}$ with $S^{(k)} = \frac{1}{n} (x^{(k)})^T x^{(k)}$, and $\Theta = (\Theta^{(1)}, \ldots, \Theta^{(K)})$. Then the negative log likelihood for the data takes the form of

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

Clearly, minimizing (2.1) leads to the maximum likelihood estimate (MLE), $\widehat{\Theta}^{(k)} = (S^{(k)})^{-1}$. However, the MLE fails when $S^{(k)}$ is singular. Furthermore, the MLE is usually dense. The $\ell_1$ regularization has been employed to induce sparsity, resulting in the sparse inverse covariance estimation [2, 11, 53]. In this paper, we employ both the $\ell_1$ regularization and the fused regularization for simultaneously estimating multiple graphs. The $\ell_1$ regularization leads to a sparse solution, and the fused penalty encourages $\Theta^{(k)}$ to be similar to its neighbors. Mathematically, we solve the following formulation:

$$
(2.2) \quad \min_{\Theta^{(k)} > 0, k = 1, \ldots, K} \sum_{k=1}^{K} \left( - \log \det (\Theta^{(k)}) + \text{tr}(S^{(k)} \Theta^{(k)}) \right) + P(\Theta),
$$

where

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

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 $\text{diag}(S^{(k)}) > 0$, $k = 1, \ldots, K$. Recall that $S^{(k)}$ is a sample covariance matrix, and hence $\text{diag}(S^{(k)}) \geq 0$. 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 $\text{diag}(S^{(k)}) > 0$, $k = 1, \ldots, K$, 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 \( C_i \cap C_i' = \emptyset \) for all \( i \neq i' \) and \( \bigcup_{i=1}^{L} C_i = \{1, \ldots, p\} \). We say that the solution \( \hat{\Theta} \) 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 \( U \in \mathbb{R}^{p \times p} \) such that each estimation precision matrix takes the form of

\[
(3.1) \quad \hat{\Theta}^{(k)} = U \left( \begin{array}{ccc} \hat{\Theta}_1^{(k)} & \cdots & \hat{\Theta}_L^{(k)} \end{array} \right)^T, \quad k = 1, \ldots, K.
\]

For simplicity of presentation, we assume throughout this paper that \( U = I \).

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

**Theorem 3.1.** Suppose that the solution \( \hat{\Theta} \) 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 \( U = I \). Let \( \hat{\Theta}_l = (\hat{\Theta}_l^{(1)}, \ldots, \hat{\Theta}_l^{(K)}) \) for \( l = 1, \ldots, L \). Then we have

\[
(3.2) \quad \hat{\Theta}_l = \arg \min_{\Theta_l \geq 0} \sum_{k=1}^{K} \left( -\log \det(\Theta_l^{(k)}) + \text{tr}(S_l^{(k)} \Theta_l^{(k)}) \right) + P(\Theta_l), \quad l = 1, \ldots, L,
\]

where \( \Theta_l^{(k)} \) and \( S_l^{(k)} \) are the \( |C_l| \times |C_l| \) symmetric submatrices of \( \Theta^{(k)} \) and \( S^{(k)} \), respectively, corresponding to the \( l \)th diagonal block, for \( k = 1, \ldots, K \), and \( \Theta_l = (\Theta_l^{(1)}, \ldots, \Theta_l^{(K)}) \) 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 \( \hat{\Theta}^{(k)}, k = 1, \ldots, K \), with \( L \) known blocks \( C_l, l = 1, \ldots, L \), if and only if \( S^{(k)}, k = 1, \ldots, K \), satisfy the following inequalities:

\[
(3.3) \quad \begin{cases} |\sum_{k=1}^{t} S_{ij}^{(k)}| \leq t\lambda_1 + \lambda_2, \\ |\sum_{k=0}^{t-1} S_{ij}^{(r+k)}| \leq t\lambda_1 + 2\lambda_2, & 2 \leq r \leq K - t, \\ |\sum_{k=1}^{t} S_{ij}^{(K-t+k)}| \leq t\lambda_1 + \lambda_2, \\ |\sum_{k=1}^{K} S_{ij}^{(k)}| \leq K\lambda_1 \end{cases}
\]
for $t = 1, \ldots, K - 1, i \in C_t, j \in C_t', l \neq l'$.

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 $E = I_{p \times p}$. Set $E_{ij} = E_{ji} = 0$ if $S^{(k)}$, $k = 1, \ldots, K$, satisfy the conditions (3.3). Otherwise, set $E_{ij} = E_{ji} = 1$.
2. Identify the connected components of the adjacency matrix $E$ (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 severaltechnical 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 $I$ and $J$ be the complement of $I$ and $J$ with respect to $\{1, \ldots, n\}$ and $\{1, \ldots, n - 1\}$, respectively. Define

\[
(3.4) \quad P_{l, J} = \{y \in \mathbb{R}^n : y_i \geq 0, y_j \leq 0, y_j - y_{j+1} \geq 0, y_j - y_{j+1} \leq 0\},
\]

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 $P_{l, J} = \{0\}$ or $P_{l, J}$ is unbounded.

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

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

\[
(3.5) \quad Q := \{\alpha(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^T \in \mathbb{R}^n : \alpha \neq 0, m \geq 0, 1 \leq l \leq n\}.
\]

**Proof.** (i) We observe that $0 \in P_{l, J}$. If $P_{l, J} \neq \{0\}$, then there exists $0 \neq y \in P_{l, J}$.

Hence, $\{\alpha y : \alpha \geq 0\} \subseteq P_{l, J}$, which implies that $P_{l, J}$ is unbounded.

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

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

\[
(3.6) \quad d_l \geq 0, \quad d_l \leq 0, \quad d_j - d_{j+1} \geq 0, \quad d_j - d_{j+1} \leq 0,
\]

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 $d_j - d_{j+1} = 0$ (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

\[
(3.7) \quad \{i : d_i = 0\} = \{m_1, \ldots, n_1\} \cup \{m_2, \ldots, n_2\} \cup \cdots \cup \{m_k, \ldots, n_k\}.
\]
One can immediately observe that
\begin{equation}
(3.8) \quad d_m = \cdots = d_n = 0, \quad d_j - d_{j+1} = 0, \quad m_i \leq j \leq n_i - 1, \quad 1 \leq i \leq k.
\end{equation}

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 \( d_{m_1} - d_{m_i} \neq 0 \) and \( d_{n_i} - d_{n_i+1} \neq 0 \) 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
\begin{equation}
(3.9) \quad d_j - d_{j+1} = 0, \quad n_{i-1} < j < m_i - 1, \quad 2 \leq i \leq k
\end{equation}
(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 \( d_{m_1} - d_{m_i} \neq 0 \) 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 \( d_{m_i} - d_{m_i} \neq 0 \) 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 \( d_{m_i} - d_{m_i} \neq 0 \) 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 \( k \geq 1 \) 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 $\text{ext}(P_{I,J}) \subseteq Q$. 

**Lemma 3.4.** Let $P_{I,J}$ and $Q$ be defined in (3.4) and (3.5), respectively. Then,

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

**Proof.** It follows from Lemma 3.3(iii) that

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

We next show that

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

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 $d \in \text{ext}(P_{I,J})$ with $I = \{1,\ldots,n\}$ and $J = \{m_1,\ldots,n-1\}$. Similarly, if $\alpha < 0$, $d \in \text{ext}(P_{I,J})$ with $I = \emptyset$ and $J$ being the complement of $J = \{m_1,\ldots,n-1\}$. Hence, $d \in \bigcup \{\text{ext}(P_{I,J}) : I \subseteq \{1,\ldots,n\}, \ J \subseteq \{1,\ldots,n-1\} \}$. 

**Lemma 3.5.** Let $x \in \mathbb{R}^n$, $\lambda_1, \lambda_2 \geq 0$ 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 $y \in \mathbb{R}^n$ 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 $k = 1,\ldots,n-1$.

**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 the following:

(a) $\mathbb{R}^n = \bigcup \{P_{I,J} : I \subseteq \{1,\ldots,n\}, \ J \subseteq \{1,\ldots,n-1\} \}$.

(b) $f(y) \leq 0$ for all $y \in \mathbb{R}^n$ 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 $y \in \mathbb{R}^n$ 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 $y \in \mathbb{R}^n$ if and only if $f(d) \leq 0$ for all $d \in \{\pm (0,\ldots,0,1,\ldots,1,0,\ldots,0)^T \in \mathbb{R}^n : m \geq 0, 1 \leq l \leq n \}$.
which together with the definition of \( f \) immediately implies that the conclusion of this lemma holds.

**Lemma 3.6.** Let \( x \in \mathbb{R}^n \), \( \lambda_1 \), \( \lambda_2 \geq 0 \) be given. The linear system

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

(3.10)

has a solution \((\gamma, v)\) if and only if the linear program

\[
\min_{\gamma, v} \left\{ 0^T \gamma + 0^T v : (\gamma, v) \text{ satisfies (3.10)} \right\}
\]

(3.11)

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

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

(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 \( y \in \mathbb{R}^n \). 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 \( \tilde{\Theta}^{(k)} \) as \( \tilde{W}^{(k)} \) for \( k = 1, \ldots, K \). By the first-order optimality conditions, we observe that \( \tilde{\Theta}^{(k)} > 0 \), \( k = 1, \ldots, K \), is the optimal solution of problem (2.2) if and only if it satisfies

\[
\begin{align*}
-\tilde{W}_i^{(k)} + S_i^{(k)} &= 0, & 1 \leq k \leq K, \\
-\tilde{W}_i^{(1)} + S_i^{(1)} + \lambda_1 \gamma_i^{(1)} + \lambda_2 v_i^{(1,2)} &= 0, \\
-\tilde{W}_i^{(k)} + S_i^{(k)} + \lambda_1 \gamma_i^{(k)} + \lambda_2 (-v_i^{(k-1,k)} + v_i^{(k,k+1)}) &= 0, & 2 \leq k \leq K - 1, \\
-\tilde{W}_i^{(K)} + S_i^{(K)} + \lambda_1 \gamma_i^{(K)} - \lambda_2 v_i^{(K-1,K)} &= 0
\end{align*}
\]

(3.13) - (3.16)
for all \( i, j = 1, \ldots, p, \ i \neq j \), where \( \gamma^{(k)}_{ij} \) is a subgradient of \(|\Theta^{(k)}_{ij}|\) at \( \Theta^{(k)}_{ij} = \hat{\Theta}^{(k)}_{ij} \) and \( v^{(k,k+1)}_{ij} \) is a subgradient of \(|\Theta^{(k)}_{ij} - \Theta^{(k+1)}_{ij}|\) with respect to \( \Theta^{(k)}_{ij} \) at \((\Theta^{(k)}_{ij}, \Theta^{(k+1)}_{ij}) = (\hat{\Theta}^{(k)}_{ij}, \hat{\Theta}^{(k+1)}_{ij})\); that is, \( v^{(k,k+1)}_{ij} = 1 \) if \( \hat{\Theta}^{(k)}_{ij} > \hat{\Theta}^{(k+1)}_{ij} \), \( v^{(k,k+1)}_{ij} = -1 \) if \( \hat{\Theta}^{(k)}_{ij} < \hat{\Theta}^{(k+1)}_{ij} \), and \( v^{(k,k+1)}_{ij} \in [-1, 1] \) if \( \hat{\Theta}^{(k)}_{ij} = \hat{\Theta}^{(k+1)}_{ij} \).

Necessity. Suppose that \( \hat{\Theta}^{(k)}_{ij}, k = 1, \ldots, K \), is a block diagonal optimal solution of problem (2.2) with \( L \) known blocks \( C_l, l = 1, \ldots, L \). Note that \( \hat{W}^{(k)} \) has the same block diagonal structure as \( \hat{\Theta}^{(k)} \). Hence, \( \hat{W}^{(k)}_{ij} = \hat{\Theta}^{(k)}_{ij} = 0 \) 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 \((\gamma^{(k)}_{ij}, v^{(k,k+1)}_{ij})\), \( k = 1, \ldots, K - 1 \), and \( \gamma^{(k)}_{ij} \) such that

\[
\begin{align*}
S^{(1)}_{ij} + \lambda_1 \gamma^{(1)}_{ij} + \lambda_2 v^{(1,2)}_{ij} &= 0, \\
S^{(k)}_{ij} + \lambda_1 \gamma^{(k)}_{ij} + \lambda_2 (v^{(k-1,k),i} - v^{(k,k+1)}_{ij}) &= 0, \ 2 \leq k \leq K - 1, \\
S^{(K)}_{ij} + \lambda_1 \gamma^{(K)}_{ij} - \lambda_2 v^{(K-1,K)}_{ij} &= 0, \\
-1 &\leq \gamma^{(k)}_{ij} \leq 1, \ 1 \leq k \leq K, \\
-1 &\leq v^{(k,k+1)}_{ij} \leq 1, \ 1 \leq k \leq K - 1.
\end{align*}
\]

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 \((\gamma^{(k)}_{ij}, v^{(k,k+1)}_{ij})\), \( k = 1, \ldots, K - 1 \), and \( \gamma^{(k)}_{ij} \) such that (3.17) holds. Now let \( \hat{\Theta}^{(k)}_{ij}, k = 1, \ldots, K \), be a block diagonal matrix as defined in (3.1) with \( U = I \), where \( \hat{\Theta} = (\hat{\Theta}^{(1)}_{ij}, \ldots, \hat{\Theta}^{(K)}_{ij}) \) is given by (3.2) for \( l = 1, \ldots, L \). Also, let \( \hat{W}^{(k)}_{ij} \) be the inverse of \( \hat{\Theta}^{(k)}_{ij} \) for \( k = 1, \ldots, K \). Since \( \hat{\Theta} \) 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 \( \hat{\Theta}^{(k)}_{ij} = \hat{W}^{(k)}_{ij} = 0 \) 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 \( \hat{\Theta}^{(k)}_{ij}, k = 1, \ldots, K \), is an optimal solution of problem (2.2). In addition, \( \hat{\Theta}^{(k)}_{ij}, k = 1, \ldots, K \), 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

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

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 $G = \{G_1, G_2, \ldots, G_M\}$ 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 $G \in G$. Let $V = \{V_1, V_2, \ldots, V_M\}$, 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

$$P_{I,J} = \{ y \in \mathbb{R}^n : y_I \geq 0, y_J \leq 0, y_u - y_v \geq 0 \ \forall (u, v) \in E_J, y_u - y_v \leq 0 \ \forall (u, v) \in E_I \},$$

where $E_I$ and $E_J$ 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, $\emptyset \neq \text{ext}(P_{I,J}) \subseteq Q$, where

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

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

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 $|E_{\setminus V_m}|$ 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

$$\sum_{k=1}^{|V_m|} s^{(u_k)}_{ij} \leq |V_m| \lambda_1 + |E_{\setminus V_m}| \lambda_2, \quad u_k \in V_m, \ \forall V_m \in V.$$ 

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:

$$\min_{\Theta^{(k)}, k = 1, \ldots, K} \sum_{k=1}^K \left( -\log \det(\Theta^{(k)}) + \text{tr}(S^{(k)} \Theta^{(k)}) \right) + \sum_{i \neq j} \phi(\Theta_{ij}),$$

where $\Theta_{ij} = (\Theta_{ij}^{(1)}, \ldots, \Theta_{ij}^{(K)})^{T} \in \mathbb{R}^K$ 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:

$$\phi(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{g=1}^g \|x_{G_g}\|_2,$$
where \( G_i, i = 1, \ldots, g \), are \( g \) groups such that \( \bigcup_{i=1}^{g} G_i = \{1, \ldots, K\} \). Different groups may overlap.

- Tree structured group regularization:

\[
\phi(x) = \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_{ij} \|x_{G_{ij}}\|_2^2,
\]

where \( w_{ij} \) is a positive weight and the groups \( G_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, d \), exhibit a tree structure \[26\].

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

\[
(3.23) \quad \min_{x} \frac{1}{2} \|x + S_{ij}\|_2^2 + \phi(x)
\]

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 \( \Theta_l^{(k)} \) by solving the FMGL \((2.2)\) with \( S^{(k)} \) replaced by \( S_l^{(k)} \). 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 \( \Theta := (\Theta_1^{(1)}, \ldots, \Theta_K^{(K)}) \) 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

\[
(4.1) \quad \min_{\Theta \succ 0} F(\Theta) := \sum_{k=1}^{K} f_k(\Theta^{(k)}) + P(\Theta),
\]

where

\[
f_k(\Theta^{(k)}) = -\log \det(\Theta^{(k)}) + \text{tr}(S^{(k)} \Theta^{(k)}).
\]

In the second-order method, we approximate the objective function \( F(\Theta) \) at the current iterate \( \Theta_t \) by a “quadratic” model \( Q_t(\Theta) \):

\[
(4.2) \quad \min_{\Theta} Q_t(\Theta) := \sum_{k=1}^{K} q_k(\Theta^{(k)}) + P(\Theta),
\]
where \( q_k \) is the quadratic approximation of \( f_k \) at \( \Theta^{(k)}_t \); that is,
\[
q_k(\Theta^{(k)}) = \frac{1}{2} \text{tr}(W^{(k)}_t D^{(k)} W^{(k)}_t) + \text{tr}((S^{(k)} - W^{(k)}_t)D^{(k)}) + f_k(\Theta^{(k)}_t)
\]
with \( W^{(k)}_t = (\Theta^{(k)}_t)^{-1} \) and \( D^{(k)} = \Theta^{(k)}_t - \Theta^{(k)}_t^{(k)} \). Suppose that \( \Theta^{(k)}_t \) is the optimal solution of (4.2). Then we obtain the Newton search direction
\[
D = \Theta^{(k)}_{t+1} - \Theta^{(k)}_t.
\]

We shall mention that the subproblem (4.2) can be suitably solved by the nonmonotone 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
\[
\text{prox}_\alpha P(Z_r) := \arg \min_{\Theta} \frac{1}{2} \|\Theta - Z_r\|_F^2 + \alpha P(\Theta),
\]
where \( r \) represents the \( r \)-th iteration in NSPG, \( \|\Theta - Z_r\|_F^2 = \sum_{k=1}^K \|\Theta^{(k)}_r - Z^{(k)}_r\|_F^2 \), \( Z_r = \Theta_r - \alpha G_r \), and \( G^{(k)}_r = S^{(k)} - 2W^{(k)}_t + W^{(k)}_t \Theta^{(k)}_t W^{(k)}_t \). Denote \( R = \Theta_r - \Theta_{r-1} \) and \( \bar{\alpha} = \sum_{k=1}^K \text{tr}(R^{(k)}_r W^{(k)}_t R^{(k)}_r W^{(k)}_t)/\sum_{k=1}^K \|R^{(k)}_r\|^2_F \). Then \( \alpha \) 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 \( P(\Theta) \), it is not hard to see that problem (4.4) can be decomposed into a set of independent and smaller sized problems,
\[
\min_{\Theta^{(k)}_{ij}, k=1, ..., K} \frac{1}{2} \sum_{k=1}^K (\Theta^{(k)}_{ij} - Z^{(k)}_{r,ij})^2 + \alpha_1 \sum_{k=1}^K |\Theta^{(k)}_{ij}| + \alpha_2 \sum_{k=1}^{K-1} |\Theta^{(k)}_{ij} - \Theta^{(k+1)}_{ij}| \tag{4.5}
\]
for all \( i > j \), \( (\alpha_1, \alpha_2) = (\lambda_1, \lambda_2) \), 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 \( D = (D^{(1)}, \ldots, D^{(K)}) \) 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 \( \Theta^{(k)}_{t+1} = \Theta^{(k)}_t + \beta D^{(k)} \), \( k = 1, \ldots, K \). 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 proof is similar to that in [17] and is thus omitted).

**Lemma 4.1.** Let \( \Theta_t = (\Theta^{(1)}_t, \ldots, \Theta^{(K)}_t) \) be such that \( \Theta^{(k)}_t > 0 \) for \( k = 1, \ldots, K \), and let \( D = (D^{(1)}, \ldots, D^{(K)}) \) be the associated Newton search direction computed according to (4.2). Suppose \( D \neq 0 \).

Then there exists a \( \beta > 0 \) such that \( \Theta_t + \beta D^{(k)} > 0 \) and the sufficient reduction condition
\[
F(\Theta_t + \beta D) \leq F(\Theta_t) + \sigma \beta \delta_t
\]

\(^{1}\text{It is well known that if } D = 0, \text{ then } \Theta_t \text{ is the optimal solution of problem (2.2).} \)
holds for all $0 < \beta < \bar{\beta}$, where $\sigma \in (0, 1/2)$ is a given constant and
\begin{equation}
\delta_t = \sum_{k=1}^{K} \text{tr}((S^{(k)} - W_t^{(k)})D^{(k)}) + P(\Theta_t + D) - P(\Theta_t).
\end{equation}

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 $\Theta_t^{(k)} + \beta D^{(k)} \succ 0$ and (4.6) holds. In particular, we choose $\beta$ to be the largest number of the sequence \{1, 1/2, \ldots, 1/2, \ldots\} that satisfies these requirements. We can use the Cholesky factorization to check the positive definiteness of $\Theta_t^{(k)} + \beta D^{(k)}$, $k = 1, \ldots, K$. In addition, the associated terms $\log \det(\Theta_t^{(k)} + \beta D^{(k)})$ and $(\Theta_t^{(k)} + \beta D^{(k)})^{-1}$ can be efficiently computed as a byproduct of the Cholesky decomposition of $\Theta_t^{(k)} + \beta D^{(k)}$.

### 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 $F$ and an active set $A$, based on the gradient at the start of each Newton iteration, and then the minimization is performed only on the variables in $F$. We call this technique “active set identification” in this paper. Due to the sparsity of the precision matrix, the size of $F$ is usually much smaller than $p^2$. Moreover, it has been shown in the single graph case that the size of $F$ 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 $t$th iteration by $\bar{G}_{t,ij}^{(k)} = S^{(k)} - W_t^{(k)}$, and its $(i,j)$th element by $\bar{G}_{t,ij}^{(k)}$. Then we have the following result.

**Lemma 4.2.** For $\Theta_t$ in the $t$th iteration, define the active set $A$ as

\[ A = \{(i,j) | \Theta_t^{(1)} = \cdots = \Theta_t^{(K)} = 0 \text{ and } \bar{G}_{t,ij}^{(1)} = \cdots = \bar{G}_{t,ij}^{(K)} \text{ satisfy the inequalities below} \} : \]

\begin{equation}
\begin{cases}
|\sum_{k=1}^{u} \bar{G}_{t,ij}^{(k)}| < u\lambda_1 + \lambda_2, \\
|\sum_{k=0}^{u} \bar{G}_{t,ij}^{(r+k)}| < u\lambda_1 + 2\lambda_2, & 2 \leq r \leq K - u, \\
|\sum_{k=1}^{u} \bar{G}_{t,ij}^{(K-w+k)}| < u\lambda_1 + \lambda_2, \\
|\sum_{k=1}^{K} \bar{G}_{t,ij}^{(k)}| < K\lambda_1
\end{cases}
\end{equation}

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

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

\begin{equation}
\min_{D} Q_t(\Theta_t + D) \quad \text{such that } D^{(1)} = \cdots = D^{(K)} = 0, \quad (i,j) \notin A.
\end{equation}

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

\begin{equation}
\min_{D} \sum_{k=1}^{K} \left( \frac{1}{2} \text{vec}(D^{(k)})^T H_t^{(k)} \text{vec}(D^{(k)}) + \text{vec}(\bar{G}_{t,ij}^{(k)})^T \text{vec}(D^{(k)}) \right) + P(\Theta_t + D)
\end{equation}

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

$$\min_{D_A} \sum_{k=1}^{K} \text{vec}(\tilde{G}_{t,A}^{(k)})^T \text{vec}(D_A^{(k)}) + P(D_A).$$

Since $\sum_{k=1}^{K} \text{vec}(D^{(k)})^T H_t^{(k)} \text{vec}(D^{(k)}) \geq 0$, the optimal solution of (4.9) is given by $D^{(1)} = \cdots = D^{(K)} = 0$. \hfill \Box

Lemma 4.2 provides an active set identification scheme to partition the variables into the free set $F$ and the active set $A$. Lemma 4.2 shows that when the variables in the free set $F$ are fixed, no update is needed for the variables in the active set $A$. 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:** $S^{(k)}$, $k = 1, \ldots, K$, $\lambda_1, \lambda_2$

**Output:** $\Theta^{(k)}$, $k = 1, \ldots, K$

Initialization: $\Theta_0^{(k)} = (\text{Diag}(S^{(k)}))^{-1}$

while Not Converged do

- Determine the sets of free and fixed indices $F$ and $A$ using Lemma 4.2.
- Compute the Newton direction $D^{(k)}$, $k = 1, \ldots, K$, by solving (4.2) and (4.3) over the free variables $F$.
- Choose $\Theta_t^{(k)}$ by performing the Armijo backtracking line search along $\Theta_t^{(k)} + \beta D^{(k)}$ for $k = 1, \ldots, K$.

end

return $\Theta^{(k)}$, $k = 1, \ldots, K$;

**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

$$H_t^{(k)} = \begin{pmatrix} H_t^{(k)}_{t,F} \\ H_t^{(k)}_{t,A} \end{pmatrix},$$

where $H_t^{(k)}_{t,F}$ is the submatrix of the exact Hessian $H_t^{(k)}$ with variables in $F$. Using $H_t^{(k)}$ instead, the subproblem (4.2) can be decomposed into the following two problems:

$$\min_{D_{J}} \sum_{k=1}^{K} \left( \frac{1}{2} \text{vec}(D_{J}^{(k)})^T H_t^{(k)}_{t,J} \text{vec}(D_{J}^{(k)}) + \text{vec}(\tilde{G}_{t,J}^{(k)})^T \text{vec}(D_{J}^{(k)}) \right) + P(\Theta_{t,J} + D_{J}), \quad J = F, A.$$
Consider the problem with respect to the variables in $\mathcal{A}$:

$$
\min_{D_{\mathcal{A}}} \sum_{k=1}^{K} \left( \frac{1}{2} \text{vec}(D_{\mathcal{A}}^{(k)})^T H_{t,A}^{(k)} \text{vec}(D_{\mathcal{A}}^{(k)}) + \text{vec}(G_{t,A}^{(k)})^T \text{vec}(D_{\mathcal{A}}^{(k)}) + P(\Theta_{t,A} + D_{\mathcal{A}}) \right),
$$

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 $D_{\mathcal{A}}^{(k)} = 0$, $k = 1, \ldots, K$. 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:

$$
\|M_{\bar{q}}(\bar{\Theta})\| \leq \eta_t \|M_{\bar{f}}(\bar{\Theta}_t)\|, \quad Q_t^H(\bar{\Theta}) - Q_t^H(\bar{\Theta}_t) \leq \zeta (L_t(\bar{\Theta}) - L_t(\bar{\Theta}_t)),
$$

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

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

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

$$
M_{\bar{f}}(\bar{\Theta}) = \frac{1}{\tau} (\bar{\Theta} - \text{prox}_{\tau P}(\bar{\Theta} - \tau \nabla \bar{f}(\bar{\Theta}))).
$$

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

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

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., $D_{\mathcal{A}} = 0$ and $\Theta_{t,A} = 0$), 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 $L_t(\Theta_{t} + \beta D_t) - L_t(\Theta_{t})$ 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 $\Theta^{(k)}$ with $L$ blocks, and each block $\Theta^{(k)}_l$ is of size $(p/L) \times (p/L)$. Each $\Theta^{(k)}_l$, $l = 1, \ldots, L$, $k = 1, \ldots, K$, has random sparsity structures. We control the number of nonzeros in each $\Theta^{(k)}_l$ 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 $\lambda_2$ is fixed to 0.1, and the $\ell_1$ regularization parameter $\lambda_1$ 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 $\frac{\max\{\|\Theta^{(k)} - \Theta^{(k)}_1\|_\infty\}}{\max\{\|\Theta^{(k)}_1\|_\infty\}} \leq 1e-6$.
- Exact: the subproblems are solved accurately as in [23]. (More precisely, NSPG stops when $\frac{\max\{\|\Theta^{(k)} - \Theta^{(k)}_1\|_\infty\}}{\max\{\|\Theta^{(k)}_1\|_\infty\}} \leq 1e-12$).
- Adaptive: stop when the adaptive stopping criterion (4.13) is satisfied. The forcing term $\eta_k$ 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

---

2http://www.yelab.net/software/MGL/
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. More details about the preprocessing strategy can be found in the same website. The dataset

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 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 $\lambda_1$ is the $l_1$ regularization parameter. The fused penalty parameter $\lambda_2$ is fixed to 0.1. $\|\Theta\|_0$ represents the total number of nonzero entries in ground truth precision matrices $\Theta^{(k)}$, $k = 1, \ldots, K$, and $\|\Theta^*\|_0$ is the number of nonzeros in the solution.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$K$</th>
<th>$L$</th>
<th>$|\Theta|_0$</th>
<th>$\lambda_1$</th>
<th>$|\Theta^*|_0$</th>
<th>FMGL-S</th>
<th>FMGL</th>
<th>ADMM-S</th>
<th>ADMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2</td>
<td>5</td>
<td>9848</td>
<td>0.08</td>
<td>9810</td>
<td>0.44</td>
<td>4.13</td>
<td>13.30</td>
<td>100.79</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>10</td>
<td>24866</td>
<td>0.055</td>
<td>23304</td>
<td>0.97</td>
<td>12.23</td>
<td>32.40</td>
<td>286.98</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>5</td>
<td>50598</td>
<td>0.054</td>
<td>44030</td>
<td>5.16</td>
<td>50.95</td>
<td>174.91</td>
<td>1595.91</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>10</td>
<td>49092</td>
<td>0.051</td>
<td>45474</td>
<td>2.33</td>
<td>24.35</td>
<td>63.75</td>
<td>458.51</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>10</td>
<td>100804</td>
<td>0.046</td>
<td>84310</td>
<td>10.27</td>
<td>111.78</td>
<td>302.86</td>
<td>2966.72</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>10</td>
<td>9348</td>
<td>0.07</td>
<td>9386</td>
<td>0.32</td>
<td>4.87</td>
<td>6.82</td>
<td>105.01</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>10</td>
<td>19750</td>
<td>0.08</td>
<td>20198</td>
<td>0.76</td>
<td>17.93</td>
<td>25.62</td>
<td>674.28</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>10</td>
<td>23538</td>
<td>0.055</td>
<td>22900</td>
<td>0.77</td>
<td>14.96</td>
<td>15.09</td>
<td>256.33</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>10</td>
<td>49184</td>
<td>0.054</td>
<td>45766</td>
<td>1.92</td>
<td>53.96</td>
<td>64.31</td>
<td>1314.18</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>10</td>
<td>47184</td>
<td>0.051</td>
<td>47814</td>
<td>1.66</td>
<td>52.32</td>
<td>29.86</td>
<td>455.43</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>10</td>
<td>98564</td>
<td>0.046</td>
<td>94566</td>
<td>4.44</td>
<td>126.26</td>
<td>128.52</td>
<td>2654.24</td>
</tr>
</tbody>
</table>

Fig. 1. Convergence behavior of FMGL with three stopping criteria (exact, adaptive, and 1E-6) and ADMM.
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. \( \lambda_1 \) and \( \lambda_2 \) 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 \( \lambda_1 \) and \( \lambda_2 \). 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.

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) $\lambda_1$ varies from 0.5 to 0.95 with $\lambda_2$ fixed to 0.015. (b) $\lambda_2$ varies from 0 to 0.2 with $\lambda_1$ 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
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%.

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 \( g \), \( \lambda_2 \) is fixed to 0.1, and \( \lambda_1 \) 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 $S \in S^p_+$ and $\Lambda \in S^p$ be such that $\text{Diag}(S) + \Lambda > 0$ and $\text{diag}(\Lambda) \geq 0$. Consider the problem

\[
\min_{X \succ 0} \log \det(X) + \text{tr}(SX) + \sum_{ij} \Lambda_{ij} |X_{ij}|.
\]

Then the following statements hold:

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

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

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

\[
\max_{U \in \mathcal{U}} \{\log \det(S + \Lambda \circ U) : S + \Lambda \circ U \succ 0\}.
\]

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

\[
\max_{U \in \mathcal{U}} \lambda_{\min}(S + \Lambda \circ U) = \max_{t, U \in \mathcal{U}} \{t : \Lambda \circ U + S - tI \succeq 0\}
\]

\[
= \min_{X \succeq 0} \max_{t, U \in \mathcal{U}} \{t + \text{tr}(X(\Lambda \circ U + S - tI))\}
\]

\[
= \min_{X \succeq 0} \left\{\text{tr}(SX) + \sum_{ij} \Lambda_{ij} |X_{ij}| : \text{tr}(X) = 1\right\},
\]

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

$$\text{(A.4)} \quad \text{tr}(SX) + \sum_{ij} \Lambda_{ij} |X_{ij}| > 0 \quad \forall X \in \Omega.$$ 

Otherwise, one has $X_{ij} = 0$ for all $i \neq j$, which, together with the facts that $S_{ii} + \Lambda_{ii} > 0$ for all $i$ and $\text{tr}(X) = 1$, implies that for all $X \in \Omega$,

$$\text{tr}(SX) + \sum_{ij} \Lambda_{ij} |X_{ij}| = \sum_i (S_{ii} + \Lambda_{ii}) X_{ii} \geq \text{tr}(X) \min_i (S_{ii} + \Lambda_{ii}) > 0.$$ 

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

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

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

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

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

which implies that

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

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

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

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

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

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

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

$$-(X^*)^{-1} + S + \Lambda \circ X^* \in \partial f(X^*),$$
which, together with \( X^* = (S + \Lambda \circ U^*)^{-1} \), implies that \( 0 \in \partial f(X^*) \). Hence, \( X^* \) 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 \( L \) is nonempty. We can observe that for any \( X \in L \),

\[
\frac{1}{2} \sum_{ij} \Lambda_{ij} |X_{ij}| = f(X) - \left[ -\log \det(X) + \text{tr}(SX) + \frac{1}{2} \sum_{ij} \Lambda_{ij} |X_{ij}| \right]
\]

\( \leq \alpha - f^* \),

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

\[
\sum_i -\log(X_{ii}) + (S_{ii} + \Lambda_{ii})X_{ii} \leq f(X) - \sum_{i \neq j} (S_{ij}X_{ij} + \Lambda_{ij}|X_{ij}|)
\]

\( \leq \alpha - \sum_{i \neq j} (S_{ij}X_{ij} + \Lambda_{ij}|X_{ij}|) \leq \delta 
\)

for some \( \delta > 0 \). In addition, notice from the assumption that \( S_{ii} + \Lambda_{ii} > 0 \) for all \( i \), and hence

\[
-\log(X_{ii}) + (S_{ii} + \Lambda_{ii})X_{ii} \geq 1 + \min_k \log(S_{kk} + \Lambda_{kk}) =: \sigma
\]

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

\[
-\log(X_{ii}) + (S_{ii} + \Lambda_{ii})X_{ii} \leq \delta - (p-1)\sigma,
\]

and hence \( X_{ii} \) is bounded for all \( i \) and \( X \in L \). We thus conclude that \( L \) 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 \( \lambda_{\min}(X) \geq \nu \) for all \( X \in L \). Hence, one has

\[
L = \{ X \succeq \nu I : f(X) \leq \alpha \}.
\]

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

We are now ready to prove Theorem 2.1.

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

\[
-\log \det(\Theta^{(k)}) + \text{tr}(S^{(k)} \Theta^{(k)}) + \lambda_1 \sum_{i \neq j} |\Theta_{ij}^{(k)}| \geq \delta \quad \forall \Theta^{(k)} > 0.
\]

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

\[
\Omega = \{ \Theta = (\Theta^{(1)}, \ldots, \Theta^{(K)}) : h(\Theta) \leq h(\hat{\Theta}), \Theta^{(k)} > 0, k = 1, \ldots, K \},
\]

\[
\Omega_k = \left\{ \Theta^{(k)} > 0 : -\log \det(\Theta^{(k)}) + \text{tr}(S^{(k)} \Theta^{(k)}) + \lambda_1 \sum_{i \neq j} |\Theta_{ij}^{(k)}| \leq \delta \right\}
\]
for $k = 1, \ldots, K$, where $\delta = h(\Theta) - (K - 1)\delta$. 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

$$\min_{\Theta \in \bar{\Omega}} h(\Theta).$$

(A.7)

In view of Lemma A.1, we know that $\Omega_k$ 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).


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 $\text{ext}(P_{l,j}) \neq \emptyset$. Next we show that $\cup \{\text{ext}(P_{l,j}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\} \} \subseteq Q$. Denote $G_J$ and $G_j$ as the subgraphs with edges only in $E_J$ and $E_j$, respectively. Accordingly, $G_J$ represents the set of all possible subgraphs with only one connected component in $G_J$, and $V_J$ denotes the corresponding node sets of $G_J$. Then we have $V_J \cup V_J \subseteq V$. Moreover, $\cup \{V_J \cup V_J, J \subseteq \{1, \ldots, |E|\}\} = V$.

Let $d \in \cup \{\text{ext}(P_{l,j}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\} \}$. 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 $V_m \in V$ 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 $V$. Therefore, $\cup \{\text{ext}(P_{l,j}) : I \subseteq \{1, \ldots, n\}, J \subseteq \{1, \ldots, |E|\} \} \subseteq Q$.

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

A.2.1. Proof of Theorem 3.8.

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

$$\begin{align*}
-\bar{\mathbf{W}}_{ii}^{(k)} + S_{ii}^{(k)} &= 0, \quad 1 \leq k \leq K, \\
-\bar{\mathbf{W}}_{ij} + S_{ij} + \partial \phi_{ij} &= 0,
\end{align*}$$

(A.8)

(A.9)

for all $i, j = 1, \ldots, p$, $i \neq j$, where $\bar{\mathbf{W}}_{ij} = (\bar{\mathbf{W}}_{ij}^{(1)}, \ldots, \bar{\mathbf{W}}_{ij}^{(K)})^T$, $S_{ij} = (S_{ij}^{(1)}, \ldots, S_{ij}^{(K)})^T$, and $\partial \phi_{ij}$ is a subgradient of $\phi(\Theta_{ij})$ at $\Theta_{ij} = \hat{\Theta}_{ij}$. 
Suppose that $\hat{\Theta}^{(k)}$, $k = 1, \ldots, K$, is a block diagonal optimal solution of problem (3.22) with $L$ known blocks $C_l$, $l = 1, \ldots, L$. $\hat{W}^{(k)}_{ij} = \hat{\Theta}^{(k)}_{ij} = 0$ 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

$$S_{ij} + \partial \phi_{ij} = 0,$$

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. □

REFERENCES


