Grouping penalties and its applications to high-dimensional models

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yunzhang Zhu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Xiaotong Shen

June, 2014

Acknowledgements

Foremost, I would like to express my deepest gratitude to Prof. Xiaotong Shen for being a truly amazing and perfect advisor. He is someone who keeps me going when times were tough, asks insightful questions, and offers invaluable advices. Throughout my graduate study, he has always been supportive in all aspects of my professional life, as well as personal development. I could not have imagined having a better advisor and mentor, and I feel really lucky and thankful to have Xiaotong as my PhD advisor.

I would also like to thank the rest of my thesis committee: Prof. Charles Geyer, Prof. Galin Jones, and Prof. Wei Pan, for their guidance and encouragement. In particular, I want to thank Prof. Wei Pan and Prof. Galin Jones for writing me several important recommendation letters to support my job applications and Doctoral Dissertation Fellowship.

I am also grateful to many friends who have made my life in graduate school full of joy. Thanks go to Yiping Yuan, Sen Yuan, Chen Xing, and Danning Li as being truly wonderful friends, and for our numerous enjoyable discussions about research and life. Thanks go to Zhihua Su, Lingzhou Xue, Fan Yang, Bo Peng, Gang Chen, Qi Yan, Jie Ren, Shanshan Ding, Wei Qian, and many others as generous and supportive friends, and all the graduate students in the statistics department and other friends at University of Minnesota for making my life here so enjoyable.

My special thanks goes to my parents, who brought me up with unconditional love. In particular, I want to thank my father, who instill me in the value of hard work and a love for mathematics from an early age. I also want to thank my mother for visiting me twice during my graduate study.

I also want thank my beautiful wife, Jing Yang, for her constant support and patience throughout the past two years, especially during the time I was job searching, for always

being there for me through the ups and downs, and for all the sacrifices she has made.

Dedication

To my mother Zhijuan Wang and my father Jianzi Zhu.

Abstract

Part I:

In high-dimensional regression, grouping pursuit and feature selection have their own merits while complementing each other in battling the curse of dimensionality. To seek a parsimonious model, we perform simultaneous grouping pursuit and feature selection over an arbitrary undirected graph with each node corresponding to one predictor. When the corresponding nodes are reachable from each other over the graph, regression coefficients can be grouped, whose absolute values are the same or close. This is motivated from gene network analysis, where genes tend to work in groups according to their biological functionalities. Through a nonconvex penalty, we develop a computational strategy and analyze the proposed method. Theoretical analysis indicates that the proposed method reconstructs the oracle estimator, that is, the unbiased least squares estimator given the true grouping, leading to consistent reconstruction of grouping structures and informative features, as well as to optimal parameter estimation. Simulation studies suggest that the method combines the benefit of grouping pursuit with that of feature selection, and compares favorably against its competitors in selection accuracy and predictive performance. An application to eQTL data is used to illustrate the methodology, where a network is incorporated into analysis through an undirected graph.

Part II:

Gaussian graphical models are useful to analyze and visualize conditional dependence relationships between interacting units. Motivated from network analysis under different experimental conditions, such as gene networks for disparate cancer subtypes, we model structural changes over multiple networks with possible heterogeneities. In particular, we estimate multiple precision matrices describing dependencies among interacting units through maximum penalized likelihood. Of particular interest are homogeneous groups of similar entries across and zero-entries of these matrices, referred to as clustering and sparseness structures, respectively. A non-convex method is proposed to seek a sparse representation for each matrix and identify clusters of the entries across the matrices. Computationally, we develop an efficient method on the basis of difference

convex programming, the augmented Lagrangian method and the block-wise coordinate descent method, which is scalable to hundreds of graphs of thousands nodes through a simple necessary and sufficient partition rule, which divides nodes into smaller disjoint subproblems excluding zero-coefficients nodes for arbitrary graphs with convex relaxation. Theoretically, a finite-sample error bound is derived for the proposed method to reconstruct the clustering and sparseness structures. This leads to consistent reconstruction of these two structures simultaneously, permitting the number of unknown parameters to be exponential in the sample size, and yielding the optimal performance of the oracle estimator as if the true structures were given *a priori*. Simulation studies suggest that the method enjoys the benefit of pursuing these two disparate kinds of structures, and compares favorably against its convex counterpart in the accuracy of structure pursuit and parameter estimation.

Contents

List of Tables

List of Figures

Chapter 1

High-dimensional linear regression with grouping penalties

1.1 Introduction and background

For high-dimensional structured data, the dimension of parameters of interest is usually high. This occurs, for instance, in a study of identifying disease-causing genes for Parkinson's disease, where expression profiles of 22283 genes are collected from 105 patients with 55 disease versus 50 control cases; see [7] for more details. In such a situation, the number of candidate genes $p = 22283$ is much higher than the sample size $n = 105$. To battle the "curse of dimensionality", one must exploit additional dependency structures from gene interactions, grouping and causal relationships. In other words, low-dimensional structures must be identified and integrated with present biological knowledge for data analysis. The central issue is simultaneous estimation of grouping and sparseness structures, called simultaneous grouping pursuit and feature selection, for structured data over a given undirected graph.

In linear regression, we consider structured data, where dependencies among predictors are loosely modeled by connectivity of an undirected graph. Grouping is only possible when predictors are connected through paths over the graph, representing prior

biological information. In this setting, we identify homogeneous subgroups of regression coefficients in absolute values, including the zero-coefficient group (feature selection). This investigation is motivated from the foregoing study, where simultaneous grouping pursuit and feature selection becomes essential over a network describing biological functionalities of genes.

Grouping pursuit has not received much attention in the literature. There is a paucity of literature for guiding practice. Two types of grouping have been investigated so far, identifying coefficients of the same values and absolute values, called Types I and II, respectively. For Type I grouping, the Fused Lasso of [8] introduces a L_1-regularization method for estimating homogeneous subgroups in a certain serial order; [9] proposes a nonconvex method for all possible homogeneous subgroups; [10] studies parameter estimation of the Fused Lasso. For Type II grouping, the OSCAR [4] suggests pairwise L_∞-penalties, and [11] employs a weighted L_γ-regularization over a graph, and [2] uses a Type I grouping method involving the pairwise sample correlations. It is Type II grouping that we shall study here. Yet, simultaneous grouping pursuit and feature selection over an arbitrary undirected graph remains under-studied. In particular, neither the interrelation between grouping pursuit and feature selection nor the impact of graph on grouping is known.

One major issue in feature selection is that highly correlated predictors impose a challenge, that is, if some predictors are included in a model then predictors that are highly correlated with them tend to be excluded in the model. This results in inaccurate feature selection. To resolve this issue, several attempts have been made. Adaptive model selection corrects the selection bias through data-driven penalty [12], and Elastic Net [3] encourages highly correlated predictors to stay together by imposing an additional ridge penalty. Relevant works can be founded in [1, 13, 14]. Despite progress, this issue remains unsettled.

Embedding feature selection into the framework of grouping pursuit, we study simultaneous grouping pursuit and feature selection through a nonconvex method. As to be seen, the method, combining the benefit of grouping pursuit with that of feature selection, outperforms either alone in predictive performance as well as accuracy of both grouping pursuit and feature selection.

We establish three main results. First, grouping pursuit and feature selection are

complementary through the proposed method. On one hand, grouping pursuit guides feature selection to yield more accurate selection than that without it. This resolves the aforementioned issue of feature selection, because highly correlated predictors can be set to be informative as an entire group when they are grouped together through grouping pursuit. On the other hand, accuracy of grouping pursuit is enhanced through feature selection by removing the group of redundant predictors. Second, simultaneous grouping pursuit and feature selection is an integrated process, improving a model's predictive performance by reducing estimation variance while maintaining roughly the same amount of bias. Third, a graph plays a critical role in the process of grouping pursuit and feature selection. A "sufficiently precise" graph, to be defined in Definition 2, enables the proposed method to handle the least favorable situation in which informative or non-informative predictors are perfectly correlated, which is impossible for other feature selection methods.

Technically, we derive a finite-sample error bound for accuracy of grouping pursuit and feature selection of the proposed method, based on which we prove that the method consistently reconstructs the unbiased least squares estimator given the true grouping, called the *oracle estimator* in what follows, as $n, p \to \infty$. This permits roughly exponentially many predictors in $p = \exp\left(n\frac{C_{\min}}{20\sigma^2 p_0}\right)$, for grouping pursuit consistency and feature selection consistency, where σ^2 is the noise variance and C_{min} a quantity to be introduced later in (1.6). In addition, the optimal performance of the oracle estimator is recovered by the proposed method in parameter estimation. Most strikingly, if the graph provides a sufficient amount of information regarding grouping, then the proposed method continues to do so even when informative or non-informative predictors are perfectly correlated, whereas feature selection alone is inconsistent without grouping pursuit [15].

To demonstrate utility of the proposed method, we analyze a dataset consisting of 210 unrelated individuals in [16], where the DNA single nucleotide polymorphisms (SNPs) data are obtained from the International HapMap Project, together with the expression data from lymphoblastoid cell lines with the Illumina Sentrix Human-6 Expression BeadChip. Then we identify some SNP locations that map *cis*-acting DNA variants for a representative gene, GLT1D1.

Chapter 1 is organized in six sections. Section 2 introduces the proposed method,

followed by computational developments in Section 3. Section 4 is devoted to a theoretical analysis of the proposed method for oracle properties. Section 5 performs some simulations and demonstrates, in simulations, that the proposed method compares favorably against some competitors. An application to analysis of SNPs data is presented as well. Section 6 contains technical proofs.

1.2 Proposed method

Consider a linear model in which responses Y_i depends on a vector of p predictors:

$$\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} = \sum_{i=1}^{p} \beta_i^0 \boldsymbol{x}_i + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{p\times p}), \tag{1.1}$$

where $\boldsymbol{\beta}^0 = (\beta_1^0, \cdots, \beta_p^0)^T$ is a vector of regression coefficients, and $\boldsymbol{X}$ is independent of random error $\boldsymbol{\varepsilon}$. In (1.1), our goal is to estimate homogeneous subgroups of components of $\boldsymbol{\beta}$ in sizes, including the zero-coefficient group of $\boldsymbol{\beta}$, particularly when p greatly exceeds n.

In (1.1), each predictor corresponds to one node over a given undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, describing prior knowledge concerning grouping, where $\mathcal{N} = \{1, \cdots, p\}$ is a set of nodes, and $\mathcal{E}$ consists of edges connecting nodes. If nodes i and j are reachable from each other, then predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be grouped; otherwise, they are impossible.

For simultaneous grouping pursuit and feature selection, we propose a nonconvex regularization cost function to minimize through pairwise comparisons over $\mathcal{G}$:

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^p} g(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big(\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 \boldsymbol{p}_1(\boldsymbol{\beta}) + \lambda_2 \boldsymbol{p}_2(\boldsymbol{\beta}) \Big), \\ &\text{where } \boldsymbol{p}_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_\tau(|\beta_j|), \quad \boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} J_\tau\big(\big||\beta_j| - |\beta_{j'}|\big|\big), \end{aligned} \tag{1.2}$$

where $J_\tau(x) = \min(\frac{x}{\tau}, 1)$ is a surrogate of the L_0-function [17]; and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ and τ are nonnegative tuning parameters. In (1.2), grouping penalty $\boldsymbol{p}_2(\boldsymbol{\beta})$ controls only magnitudes of differences or sums of coefficients ignoring their signs over $\mathcal{G}$. Through $\boldsymbol{p}_j(\boldsymbol{\beta})$; $j = 1, 2$, simultaneous grouping pursuit and feature selection is performed by adaptive shrinkage toward unknown locations and the origin jointly, where only large coefficients and pairwise differences are shrunken.

In (1.2), the proposed method is designed to outperform grouping pursuit alone and feature selection alone, through tuning two regularizers. Moreover, the method is positively impacted by the prior information specified by the given graph. These aspects will be confirmed by our theoretical analysis in Section 5.

To understand the role that $\boldsymbol{p}_2(\boldsymbol{\beta})$ plays, we now examine alternative forms of penalties for grouping. Five forms of $\boldsymbol{p}_2(\boldsymbol{\beta})$ have been proposed, including Elastic Net with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 = \frac{1}{2(p-1)} \sum_{j<j'} \big((\beta_j - \beta_{j'})^2 + (\beta_j + \beta_{j'})^2\big)$, a graph version of Elastic Net [1] with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} \big(\frac{\beta_j}{\sqrt{d_j}} - \frac{\beta_{j'}}{\sqrt{d_{j'}}}\big)^2$ with d_i being the number of direct neighbors of node x_i in $\mathcal{G}$, the OSCAR with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{j<j'} \max(|\beta_j|, |\beta_{j'}|)$, and a weighted penalty [11] with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} 2^{1/\gamma'} \big(\frac{|\beta_j|^\gamma}{w_j} + \frac{|\beta_{j'}|^\gamma}{w_{j'}}\big)^{1/\gamma}$, $\frac{1}{\gamma} + \frac{1}{\gamma'} = 1$ and weight factor $\boldsymbol{w}$, and [2] proposes $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} |\beta_j - sign(\hat{\rho}_{jj'})\beta_{j'}|)$, where $sign(\hat{\rho}_{jj'})$ is the sign of the sample correlation between predictors $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$. Although these grouping penalties and their variants can improve accuracy of feature selection, additional estimation bias may occur due to strict convexity of $\boldsymbol{p}_2(\boldsymbol{\beta})$ as in the Lasso case [18] or due to possible graph misspecification. For instance, additional bias may be introduced by the grouping penalty in [2], when $\hat{\rho}_{jj'}$ wrongly estimates the sign of $\hat{\beta}_j\hat{\beta}_{j'}$. Despite good empirical performance, statistical properties of these methods have not been studied, regarding grouping pursuit as well as its impact on feature selection.

The proposed nonconvex grouping penalty resolves aforementioned issues of convex grouping penalties through adaptive shrinkage, because it shrinks small differences in absolute values, as opposed to large ones. As a result, estimation bias is reduced as compared to a convex penalty. This phenomenon has been noted in feature selection, where there is a trade-off between estimation bias and feature selection consistency [19]. Most critically, as to be shown later by both theoretical results and numerical examples, the nonconvex method continues to perform well even when the graph is wrongly specified, which is unlike a convex method.

1.3 Computation

This section develops a computational method for nonconvex minimization in (1.2) through difference convex (DC) programming [20]. One key idea to DC programming is decomposing the objective $g(\boldsymbol{\beta})$ into a difference of two convex functions $g(\boldsymbol{\beta}) =$

$g_1(\boldsymbol{\beta}) - g_2(\boldsymbol{\beta})$, where

$$g_1(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\lambda_1}{\tau}\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\mathcal{E}}\big(|\beta_j + \beta_{j'}| + |\beta_j - \beta_{j'}|\big),$$

$$g_2(\boldsymbol{\beta}) = \frac{\lambda_1}{\tau}\sum_{j=1}^{p}\max(|\beta_j| - \tau, 0) + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\mathcal{E}}\max(2|\beta_j| - \tau, 2|\beta_{j'}| - \tau, |\beta_j| + |\beta_{j'}|).$$

Our unconstrained DC method is then summarized as follows.

Algorithm 1:

Step 1. (Initialization) Supply an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, for instance, $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. Specify precision tolerance level $\epsilon > 0$.

Step 2. (Iteration) At iteration $k+1$, compute $\hat{\boldsymbol{\beta}}^{(k+1)}$ by solving subproblem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big(g_1(\boldsymbol{\beta}) - \langle\boldsymbol{\beta}, \nabla g_2(\hat{\boldsymbol{\beta}}^{(k)}), \rangle\Big) \tag{1.3}$$

where $\nabla g_2(\hat{\boldsymbol{\beta}}^{(k)})$ is a gradient vector of $g_2(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k)}$ and $\langle\cdot,\cdot\rangle$ is the inner product. (Perturbation) For each j, if $|\beta_j| = \tau$ or there exists j' such that $(j,j') \in \mathcal{E}$ and $\big||\beta_j| - |\beta_{j'}|\big| = \tau$, we perturb β_j by $\beta_j \pm \epsilon^*$ to strictly decrease the cost function.

Step 3. (Stopping rule) Terminate when $g(\hat{\boldsymbol{\beta}}^{(k+1)}) - g(\hat{\boldsymbol{\beta}}^{(k)}) \le \epsilon$.

Next we present some computational properties of **Algorithm 1**.

Theorem 1 *For any $\boldsymbol{\beta}$, if $|\beta_j| = \tau$ for some j; $1 \le j \le p$, or $\big||\beta_j| - |\beta_{j'}|\big| = \tau$ for some (j,j'); $j' \ne j$, then we can perturb the β_j to strictly decrease the value of $g(\boldsymbol{\beta})$ in (1.2). Moreover,* **Algorithm 1** *converges exactly in finite iteration steps from any initial value.*

The finite convergence property of **Algorithm 1** is unique, due primarily to piecewise linearity of $\boldsymbol{p}_j(\boldsymbol{\beta})$; $j = 1, 2$. However, other smooth non convex (differentiable) penalties may not possess this computationally attractive feature.

1.4 Theory

This section considers a constrained L_0-version of (1.2) for theoretical investigation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\ S(\boldsymbol{\beta}) = \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2, \text{ subject to}$$
$$\sum_{j=1}^{p}\mathbb{I}(|\beta_j| \ne 0) \le C_1,\ \sum_{(j,j')\in\mathcal{E}}\mathbb{I}\big(\big||\beta_j| - |\beta_{j'}|\big| \ne 0\big) \le C_2. \tag{1.4}$$

Moreover, we study a constrained computational surrogate of the L_0-version (1.4):

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \; S(\boldsymbol{\beta}) = \frac{1}{2}\big\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\big\|^2, \text{ subject to}$$
$$\sum_{j=1}^{p} J_\tau(|\beta_j|) \le C_1, \sum_{(j,j')\in\mathcal{E}} J_\tau\big(\big||\beta_j|-|\beta_{j'}|\big|\big) \le C_2, \tag{1.5}$$

where the three non-negative tuning parameters (C_1, C_2, τ) control two-level adaptive shrinkage toward unknown locations and the origin. As discussed in Section 3, the DC method described in **Algorithm 1** targets at a local minimizer of (1.2), which can be viewed a convex relaxation of (1.4) or (1.5).

With regard to simultaneous grouping pursuit and feature selection, we will prove that global minimizers of (1.4) and (1.5) reconstruct the ideal "oracle estimator" as if the true grouping were available in advance. As a result of the reconstruction, key properties of the oracle estimator are simultaneously achieved by the proposed method.

1.4.1 The oracle estimator

Throughout this section, we write the $n \times p$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is the ith column of $\boldsymbol{X}$. Denote by $\lambda_{min}\big(\boldsymbol{A}\big)$ the smallest eigenvalue of a square matrix $\boldsymbol{A}$. For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$, rewrite $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_{\mathcal{I}_0}, \boldsymbol{\beta}_{\mathcal{I}_1}, \cdots, \boldsymbol{\beta}_{\mathcal{I}_K})$, where $\boldsymbol{\beta}_{\mathcal{I}_0} = \boldsymbol{0}$, $\boldsymbol{\beta}_{\mathcal{I}_j} = (\alpha_j \boldsymbol{1}_{\mathcal{I}_{j1}}, -\alpha_j \boldsymbol{1}_{\mathcal{I}_{j2}})^T$; $j = 1, \cdots, K$, is a vector of length $|\mathcal{I}_j|$, with $\mathcal{I}_j = \mathcal{I}_{j1} \cup \mathcal{I}_{j2}$ and $\mathcal{I}_{j1} \cap \mathcal{I}_{j2} = \emptyset$, consisting of two disjoint subgroups with coefficients being opposite signs, where $|\mathcal{I}_{j1}| = 0$ or $|\mathcal{I}_{j2}| = 0$ is permitted. Given $\boldsymbol{\beta}$, let $\mathcal{G} = \big(\mathcal{I}_0, \mathcal{I}_1, \cdots, \mathcal{I}_K\big)$ with $\mathcal{I}_j = \mathcal{I}_{j_1} \cup \mathcal{I}_{j_2}$, which partitions $\mathcal{I} = \{1, \cdots, p\}$. Given $\mathcal{G}$, define $\boldsymbol{X}_\mathcal{G}$ as $\Big(\sum_{k\in\mathcal{I}_{11}} \boldsymbol{x}_k - \sum_{k\in\mathcal{I}_{12}} \boldsymbol{x}_k, \cdots, \sum_{k\in\mathcal{I}_{K1}} \boldsymbol{x}_k - \sum_{k\in\mathcal{I}_{K2}} \boldsymbol{x}_k\Big)$ to be a collapsed matrix by collapsing columns of $\boldsymbol{X}$ according to $\mathcal{G}$. Given $B = \{i_1, \cdots, i_{|B|}\} \in \mathcal{I}$, where $i_1 < \cdots < i_{|B|}$, define $\boldsymbol{X}_B$ as $(\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_{|B|}})$ to be a submatrix of $\boldsymbol{X}$; and $\boldsymbol{\beta}_B$ to be vector $(\beta_{i_1}, \cdots, \beta_{i_{|B|}})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$.

Definition 1 (Oracle estimator) *Given the true grouping* $\mathcal{G}^0 = \big(\mathcal{I}_0^0, \mathcal{I}_1^0, \cdots, \mathcal{I}_{K_0}^0\big)$ *with* $\mathcal{I}_j^0 = \mathcal{I}_{j1}^0 \cup \mathcal{I}_{j2}^0, j = 1, \cdots, K_0$, *the oracle estimator* $\hat{\boldsymbol{\beta}}^{ol} = (\hat{\beta}_1^{ol}, \cdots, \hat{\beta}_p^{ol})^T$ *is* $\hat{\beta}_k^{ol} = \hat{\alpha}_j$ *if* $k \in \mathcal{I}_{j1}^0$, $\hat{\beta}_k^{ol} = -\hat{\alpha}_j$ *if* $k \in \mathcal{I}_{j2}^0$; $j = 1, \cdots, K_0$, *and* $\hat{\beta}_k^{ol} = 0$ *if* $k \in \mathcal{I}_0^0$, *where* $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{K_0}) = \mathrm{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^{K_0}} \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}_{\mathcal{G}^0}\boldsymbol{\alpha}\big\|^2$.

The *oracle estimator* is the unbiased least squares estimate given the true grouping $\mathcal{G}^0$. It reduces to the oracle estimator for feature selection alone when no homogeneous groups exist for informative predictors.

1.4.2 Non-asymptotic probability error bounds

This section derives a non-asymptotic probability error bound for simultaneous grouping pursuit and feature selection, based on which we prove that (1.4) and (1.5) reconstruct the oracle estimator. This implies grouping pursuit consistency as well as feature selection consistency, under one simple assumption, what we call the degree-of-separation condition.

Let $\mathcal{S} = \big\{\mathcal{G} \neq \mathcal{G}^0 : C_1(\mathcal{G}) \leq p_0; C_2(\mathcal{G}, \mathcal{E}) \leq c_0\big\}$ be a constrained set defined in (1.4), with $C_1(\mathcal{G}) = \sum_{j=1}^{p} \mathbb{I}(|\beta_j| \neq 0) = |\mathcal{I} \setminus \mathcal{I}_0|$ and $C_2(\mathcal{G}, \mathcal{E}) = \sum_{(j,j')\in\mathcal{E}} \mathbb{I}\big(\big||\beta_j| - |\beta_{j'}|\big| \neq 0\big) = \sum_{0\leq i<i'\leq K} \sum_{j\in\mathcal{I}_i, j'\in\mathcal{I}_{i'}} \mathbb{I}\big((j,j') \in \mathcal{E}\big)$, $p_0 = C_1(\mathcal{G}^0)$ and $c_0 = C_2(\mathcal{G}^0, \mathcal{E})$.

Let $A \subset \{1, \cdots, p\}$, and $A_0 = \mathcal{I} \setminus \mathcal{I}_0$ whose size $|A_0| \equiv p_0$. Define $\mathcal{S}_A = \big\{\mathcal{G} \in \mathcal{S} : \mathcal{I} \setminus \mathcal{I}_0 = A\big\}$ to be a set of groupings indexed by set A of nonzero coefficients. Let $S_i^* \equiv \max_{A:|A_0\setminus A|=i} |\mathcal{S}_A|$ be the maximal of $\mathcal{S}_A$ satisfying $|A_0 \setminus A| = i$ and further let $S^* = \exp\big(\max_{1\leq i\leq p_0} \frac{\log S_i^*}{i}\big)$. Finally, let $K_i^* \equiv \max_{\mathcal{G}\in\mathcal{S}:|A_0\setminus A|=i} K(\mathcal{G})$, with $K^* = \max_{1\leq i\leq p_0} \frac{K_i^*}{i}$.

The degree-of-separation condition is stated as follows.

$$C_{\min} \geq d_0 \frac{2\log p + K^* + 2\log S^*}{n} \sigma^2, \tag{1.6}$$

where $d_0 > 10$ is a constant, $C_{\min} \equiv \min_{\mathcal{G}\in\mathcal{S}} \frac{\|(I - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}^0\|^2}{|A_0\setminus A|n}$, and $\boldsymbol{P}_{\mathcal{G}}$ is a projection onto the linear space spanned by columns of the collapsed design matrix $\boldsymbol{X}_{\mathcal{G}}$. Here $C_{\min}$ describes the least favorable situation for simultaneous grouping pursuit and feature selection, and characterizes the level of difficulty of the underlying problem.

In (1.6), the graph specification may have an impact on C_{min}. We introduce the notion of "consistent" graph in Definition 2. A "consistent" graph is a minimal requirement for reconstruction of the oracle estimator, where there exists a path in $\mathcal{E}$ connecting any two predictors in the same true group.

Definition 2 (*"Consistent"* graph) *An undirected graph $\big(\mathcal{N}, \mathcal{E}\big)$ is consistent with respect to the true grouping $\mathcal{G}^0 = \big(\mathcal{I}_0^0, \cdots, \mathcal{I}_{K_0}^0\big)$, if for any $j = 1, \cdots, K_0$, $\mathcal{E}|_{\mathcal{I}_j^0}$, the subgraph restricted on node set $\mathcal{I}_j^0$, is connected.*

We now present our non-asymptotic probability error bounds for global minimizers of (1.4) and (1.5) in terms of $(C_{min}, n, p, p_0, \sigma^2)$, where p_0, p may depend on n.

Theorem 2 *(L_0 method) If $\mathcal{E}$ is consistent with respect to $\mathcal{G}^0$, then for a global minimizer of (1.4) $\hat{\boldsymbol{\beta}}^{l_0}$ with estimated grouping $\hat{\mathcal{G}}^{l_0}$ at $(C_1, C_2) = (p_0, c_0)$,*

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\leq\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 10\sigma^2\frac{\log S^*}{n}\big)\Big). \quad (1.7)$$

Under (1.6), $\mathbb{P}\Big(\hat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\Big) \leq \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \to 0$, and $\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{l_0} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1+o(1)\big)\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{ol} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1+o(1)\big)\frac{K_0}{n}$, as $n, p \to \infty$.

Theorem 3 *(Surrogate method) If $\mathcal{E}$ is consistent with respect to $\mathcal{G}^0$, then for a global minimizer of (1.5) $\hat{\boldsymbol{\beta}}^{g}$ with estimated grouping $\hat{\mathcal{G}}^{g}$ when $(C_1, C_2) = (p_0, c_0)$ and $\tau \leq 2\sigma\sqrt{\frac{\log p}{2np^3\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)}}$,*

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\leq\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 20\sigma^2\frac{\log S^*}{n}\big)\Big). \quad (1.8)$$

Under (1.6), $\mathbb{P}\Big(\hat{\mathcal{G}}^{g} \neq \mathcal{G}^0\Big) \leq \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \to 0$, and $\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{g} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1+o(1)\big)\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{ol} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1+o(1)\big)\frac{K_0}{n}$, as $n, p \to \infty$.

In Theorems 2 and 3, K^* and S^* need to be computed. Next we present some bounds for (K^*, S^*).

Corollary 1 *If $\mathcal{E}$ is a fused graph, that is $\mathcal{E} = \{(i, i+1) : i = 1, \cdots, p-1\}$, then*

$$S^* \leq \sum_{i=1}^{K_0}\binom{p_0}{i} \leq p_0^{K_0+1}, \text{ and } K^* \leq K_i^* \leq K_0; i = 1, \cdots, K_0 - 1. \quad (1.9)$$

As a result, (2) and (3) reduce to

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\leq\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10K_0\sigma^2\frac{\log p_0}{n}\big)\Big), \quad (1.10)$$

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\leq\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 20K_0\sigma^2\frac{\log p_0}{n}\big)\Big). \quad (1.11)$$

For the purpose of comparing simultaneous grouping pursuit and feature selection with feature selection alone without grouping pursuit, we present (1.7) and (1.8) in a parallel manner as that in [15] for feature selection alone, where the degree of separation

for feature selection alone is $C_{min}^T = \inf_{A\neq A_0,|A|\leq p_0}\Big(\big(|A_0\setminus A|n\big)^{-1}\|(I-\boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}^0\|^2\Big)$, which is in contrast to C_{min} in (1.6). Specifically, the feature selection estimators in [15] correspond to that in (1.4) and (1.5) with $(C_1, C_2) = (p_0, +\infty)$. By the necessary condition in Theorem 1 of [15], the necessary condition for feature selection alone requires that

$$C_{\min}^T \geq d_1 \frac{\log p}{n}\sigma^2, \text{ as } n, p \to +\infty \tag{1.12}$$

for some $d_1 > 0$. Note that the lower bound of C_{min} in (1.6) can be larger than that of C_{min}^T in (1.12). This generally means that, in terms of complexity, the problem of recovering *oracle estimator* in the sense of simultaneous grouping pursuit and feature selection is more difficult than that of feature selection alone.

To study the impact of a graph on simultaneous grouping pursuit and feature selection, we introduce another notion "sufficiently preciseness" in Definitions 3. A sufficiently precise graph is consistent, and the number of correctly connected edges for each true group is two times higher than that of wrongly connected ones, where within group connections refer to correct connections whereas between group connections are defined to be wrongly corrected.

Definition 3 (*"Sufficiently precise"* graph) *For any index sets $\mathcal{I}_j$; $j = 1, 2$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$, we define $d_{\mathcal{E}}(\mathcal{I}_1, \mathcal{I}_2) = \sum_{i\in\mathcal{I}_1; j\in\mathcal{I}_2}\mathbb{I}\big((i,j)\in\mathcal{E}\big)$ to be the number of connections between them over $\mathcal{E}$. A graph is sufficiently precise with respect to $\mathcal{G}^0$, if it is a consistent graph and satisfies: for any $j = 0, \cdots, K_0$, the number of within-group connections exceeds two times that of between-group connections for $\mathcal{I}_j^0$, that is, $d_{\mathcal{E}}(E, \mathcal{I}_j^0 \setminus E) > 2d_{\mathcal{E}}(E, \cup_{i\neq j}\mathcal{I}_i^0)$, for any $E \subset \mathcal{I}_j^0$.*

Lemma 1 below establishes a connection between C_{min} and C_{min}^T, and describes their behaviors in presence of perfectly correlated predictors.

Lemma 1 *(Level of difficulty) For any consistent graph,*

$$C_{min} \geq \eta^2 c_{min}, \quad C_{min}^T \geq \gamma^2 c_{min}, \text{ and } \gamma \geq \eta, \tag{1.13}$$

where

$$c_{min} = \min_{|B|\leq 2|\mathcal{I}\setminus\mathcal{I}_0^0|, \mathcal{I}\setminus\mathcal{I}_0^0\subseteq B}\lambda_{min}\Big(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B\Big),$$

$$\eta^2 = \min\big(\min_{(j,j'):j\sim j',|\beta_j^0|\neq|\beta_{j'}^0|}\frac{1}{2}(|\beta_j^0|-|\beta_{j'}^0|)^2,\gamma^2\big),$$

and $\gamma = \min_{j\in A_0}|\beta_j^0|$. *If the graph is sufficiently precise, and* $\mathcal{I}_i^0$ *can be further partitioned into perfectly correlated subgroups* $\mathcal{I}_i^0 = \{A_{i1},\cdots,A_{in_i}\}$; $i = 1,\cdots,K_0$, *then*

$$C_{min} \geq c_{min}^G \min_{\boldsymbol{\alpha},\boldsymbol{A}}\|\boldsymbol{\gamma}-\boldsymbol{A}\boldsymbol{\alpha}\| > 0, \text{ and } C_{min}^T = 0, \tag{1.14}$$

where $\boldsymbol{A} = (a_{ns})$ *is a* $N_0\times(K_0-1)$ *matrix with* $a_{ns}\in\mathbb{Z}$, $N_0 = \sum_{i=1}^{K_0} n_i$, $\sum_{s=1}^{K_0-1}|a_{ns}| \leq |A_{im}|$, $\boldsymbol{\gamma} = (\gamma_1,\cdots,\gamma_{N_0})$ *with* $\gamma_i = |A_{im}|\beta_i^0$; $n = \sum_{j=1}^{i-1} n_k + m$; $m = 1,\cdots,n_i, i = 1,\cdots,K_0$, *and*

$$c_{\min}^G = \min_{B:|B\cap(\mathcal{I}\setminus\mathcal{I}_0^0)|\leq p_0,|B\cap\mathcal{I}_0^0|\leq p_0,|B\cap A_{im}|\leq 1,i=1,\cdots,K_0,m=1,\cdots,n_i}\lambda_{min}\Big(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B\Big).$$

Here $c_{\min}^G = c_{\min}$ *in absence of perfectly correlated predictors, and* $c_{\min}^G \geq c_{\min}$ *otherwise.*

Lemma 1 says that simultaneous grouping pursuit and feature selection is generally more difficult than feature selection alone, as described by the degree-of-separation condition for $C_{\min}$ and $C_{\min}^T$ in (1.6) and (1.12). Importantly, the impact of grouping pursuit on feature selection is evident in situations where some informative features are perfectly correlated. When a graph is sufficiently precise, simultaneous grouping and feature selection continues to work when $C_{\min} > 0$ by Lemma 1. However, any feature selection method breaks down because of non-identifiable models when $C_{\min}^T = 0$, leading to inconsistent selection in view of the necessary condition in Theorem 1 of [15]. In other words, simultaneous grouping and feature selection overcomes the difficulty of highly correlated features in feature selection.

Lemma 2 *The results in Theorems 2 and 3 continue to hold for fixed* p *with* $n\to+\infty$ *with (1.6) replaced by* $\lim_{n\to+\infty} nC_{min} = +\infty$.

1.5 Numerical examples

1.5.1 Simulations

This section examines operating characteristics of the proposed method and compares it against some competitors, through simulations, with regard to accuracy of grouping

pursuit as well as feature selection, in addition to accuracy of parameter estimation. The competitors are OSCAR [4], *GFlasso* [2] and aGrace [1].

To measure accuracy of grouping pursuit and feature selection, we introduce four separate metrics. For the accuracy of feature selection, we use false and negative positives for feature selection, denoted by $VFP = \frac{\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0, \beta_j^0 = 0)}{p - p_0}$ and $VFN = \frac{\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j = 0, \beta_j^0 \neq 0)}{p_0}$. For grouping pursuit, we consider false and negative positives for feature selection, that is, $GFP = \frac{\sum_{(j,j') \in \mathcal{E}^0} \mathbb{I}(\hat{\beta}_j sign(\beta_j^0) \neq \hat{\beta}_{j'} sign(\beta_{j'}^0))}{|\mathcal{E}^0|}$ and $GFN = \frac{\sum_{(j,j') \notin \mathcal{E}^0} \mathbb{I}(|\hat{\beta}_j| = |\hat{\beta}_{j'}|)}{p(p-2)/2 - |\mathcal{E}^0|}$. Clearly, VFP, VFN, GFP and GFN are between $[0, 1]$, with a small value indicating high accuracy for variable selection and grouping pursuit.

To measure the performance of parameter estimation for $\hat{\boldsymbol{\beta}}$, we use predictive mean squared error $PMSE(\hat{\boldsymbol{\beta}}) = \frac{\|\boldsymbol{Y}^{\text{test}} - \boldsymbol{X}^{\text{test}} \hat{\boldsymbol{\beta}}\|^2}{n^{\text{test}}}$, where $\boldsymbol{Y}^{\text{test}}, \boldsymbol{X}^{\text{test}}$ are test data and and n^{test} is the sample size of the test data. In simulations, the values of PMSE are reported, as well as values of (VFP, VFN, GFP, GFN).

Example 1 (Gene network: Large p but small n). Consider a regulatory gene network example in [1], where an entire network consists of 200 subnetworks, each with one transcription factor (TF) and its 10 regulatory target genes; see [1] for a display of the network. For this network, each predictor is generated according to $\mathcal{N}(0, 1)$. To mimic a regulatory relationship, the predictor of each target gene and the TF had a bivariate normal distribution with correlation $\rho = .2, .5, .9$; conditional on the TF, the target genes are independent. In addition, $\varepsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ with $\sigma_e^2 = \frac{\sum_j^p (\beta_j^0)^2}{4}$. The true regression coefficients are:

$$\boldsymbol{\beta}^0 = \big(2, \underbrace{2/\sqrt{10}, \ldots, 2/\sqrt{10}}_{10}, -2, \underbrace{-2/\sqrt{10}, \ldots, -2/\sqrt{10}}_{10},$$
$$4, \underbrace{4/\sqrt{10}, \ldots, 4/\sqrt{10}}_{10}, -4, \underbrace{-4/\sqrt{10}, \ldots, -4/\sqrt{10}}_{10}, \underbrace{0, \ldots, 0}_{p-44}\big)^T, \quad p = 2200$$

As suggested by Table 1.1, the proposed method compares favorably against its competitors across all the situations, in terms of parameter estimation and accuracy of grouping pursuit and feature selection. Interestingly, *GFlasso* and aGrace perform similarly. Furthermore, all the graph-based methods performs reasonably well except Elastic Net where it does not exploit the informative graph information.

Table 1.1: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, based on 100 simulation replications in Example 1, for our proposed method (Grouping), adaptive Grace (aGrace) [1], *GFlasso* [2], Elastic-Net (Enet) [3] and Oscar [4]

Correlation	Method	PMSE	VFP	VFN	GFP	GFN
$Cor = .9$	Ours	20.6(2.2)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)
	GFlasso	22.6(2.4)	11.3%(4.93%)	.15%(.66%)	20.8%(8.50%)	.14%(.59%)
	Oscar	22.7(2.5)	63.2%(10.6%)	.00%(.00%)	83.6%(8.20%)	.00%(.00%)
	Enet	45.7(4.9)	18.2%(22.6%)	6.29%(5.84%)	22.4%(3.51%)	6.95%(6.30%)
	aGrace	22.5(2.4)	39.1%(43.0%)	.00%(.00%)	43.1%(37.5%)	.00%(.00%)
$Cor = .5$	Ours	20.5(2.1)	.17%(.56%)	.25%(.84%)	.33%(1.07%)	.24%(.84%)
	GFlasso	22.6(2.5)	15.10%(6.30%)	.00%(.00%)	27.15%(10.4%)	.00%(.00%)
	Oscar	24.3(2.8)	72.7%(6.23%)	.00%(.00%)	89.9%(3.90%)	.00%(.00%)
	Enet	40.8(4.7)	40.6%(42.8%)	3.43%(3.80%)	2.12%(8.85%)	6.35%(5.15%)
	aGrace	22.4(2.5)	36.2%(41.4%)	.00%(.00%)	41.2%(36.8%)	.00%(.00%)
$Cor = .2$	Ours	20.8(2.1)	.04%(.18%)	.84%(3.38%)	.09%(.36%)	.83%(3.35%)
	GFlasso	22.6(2.5)	19.7%(7.73%)	.00%(.00%)	34.7%(12.3%)	.00%(.00%)
	Oscar	26.7(3.3)	68.3%(9.80%)	.00%(.00%)	86.7%(8.49%)	.00%(.00%)
	Enet	47.1(6.7)	10.7%(12.8%)	16.1%(7.45%)	17.5%(4.83%)	14.8%(6.54%)
	aGrace	23.9(3.3)	35.7%(34.7%)	.13%(.54%)	45.8%(30.3%)	.10%(.42%)

To see the impact of grouping pursuit on feature selection and vice versa, we compare the proposed method with (λ_1, λ_2) jointly against feature selection alone with $(\lambda_1, \lambda_2 = 0)$, and grouping pursuit alone with $(\lambda_1 = 0, \lambda_2)$. As indicated in Table 1.2, simultaneous grouping pursuit and feature selection outperforms either, as expected. The improvement in accuracy of feature selection is large, as measured by VFP, VFN, where nearly perfect reconstruction is evident. This is in contrast to accuracy of feature selection alone, where the false negative rate is high for either, in the presence of highly correlated predictors with the TF-gene correlation .9. This confirms our foregoing discussion about the impact of grouping pursuit on feature selection. Meanwhile, feature selection also enhances grouping pursuit as evident from an improvement over grouping pursuit alone.

Example 2 (Impact of erroneous edges) To understand the impact of specification of prior knowledge on a method's performance, we consider the network in **Example 1** with a varying fraction of erroneous edges adding into the network, involving different correlation structures among predictors. In set-up 1, we set the TF-gene

Table 1.2: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, based on 100 simulation replications in Example 1, for feature selection alone with $\lambda_2 = 0$ in (1.2) (TLP), grouping pursuit alone with $\lambda_1 = 0$ in (1.2) (Grouping), and simultaneous grouping pursuit and feature selection (Both).

Correlation	Method	PMSE	VFP	VFN	GFP	GFN
$Cor = .9$	Both	20.55(2.23)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)
	TLP	24.54(2.43)	.01%(.03%)	45.7%(9.44%)	.02%(.06%)	45.5%(9.44%)
	Grouping	372(218)	100%(.00%)	.00%(.00%)	82.8%(22.8%)	18.9%(24.6%)
$Cor = .5$	Both	20.54(2.12)	.17%(.56%)	.25%(.84%)	.33%(1.07%)	.24%(.84%)
	TLP	31.86(3.49)	.09%(.13%)	42.8%(9.01%)	.19%(.26%)	42.6%(9.02%)
	Grouping	462(47.8)	100%(.00%)	.00%(.00%)	49.3%(.96%)	59.4%(11.9%)
$Cor = .2$	Both	20.75(2.12)	.04%(.18%)	.84%(3.38%)	.08%(.36%)	.83%(3.35%)
	TLP	41.66(5.57)	.42%(.59%)	50.1%(13.1%)	.84%(1.17%)	49.7%(13.2%)
	Grouping	287(29.1)	100%(.00%)	.00%(.00%)	50.6%(.71%)	69.2%(12.4%)

correlation to be .9 with independent TF's. For set-up 2, the TF-TF correlation is set to be .5 so that the correlation between the informative and noisy TF's is .5. For both the set-ups, we randomly add $k = 0, 10, 100$ edges between each active and other inactive nodes. As a result, the network has $p_0 k$ more edges than that in the previous example, where p_0 is the number of active nodes. In this case, the true regression coefficients are

$$\boldsymbol{\beta}^0 = \big(\underbrace{2,\ldots,2}_{11},\underbrace{-2,\ldots,-2}_{11},\underbrace{4,\ldots,4}_{11},\underbrace{-4,\ldots,-4}_{11},\underbrace{0,\ldots,0}_{p-44}\big)^T, \quad p = 2200,$$

with $\sigma_e^2 = 1$. Moreover, we use the "oracle recovery rate", defined as the percentage of times that the oracle estimator is reconstructed over 100 simulation replications. The total number of erroneous edges is 0, 440 and 4400. Results of Example 1 in presence of erroneous edges are also reported in Table 1.4 with correlation .9 and the average number of erroneous edges $0, 2, 10$.

As suggested by Table 1.3, the proposed method performs best in terms of parameter estimation and reconstruction of the oracle estimator across all the set-ups. As a result, it yields accurate identification of grouping structures, as evident by nearly zero false positives and negatives for grouping and feature selection VFP, VFN, GFP and

Table 1.3: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, as well as %Oracle, the percentage of time that our method reconstructs the oracle estimator, based on 100 simulation replications in Example 2, for our proposed method (Our), adaptive Grace (aGrace) [1], *GFlasso* [2], Elastic-Net (Enet) [3] and Oscar [4]. Setups have the TF-TF correlation of 0 and .5; k is the average number of erroneous edges.

Setup 1	Method	PMSE	VFP	VFN	GFP	GFN	%Oracle
$k = 0$	Ours	1.02(.02)	.00%(.00%)	.00%(.00%)	.00%(.00%)	.00%(.00%)	100%
	GFlasso	1.12(.05)	8.41%(1.66%)	.00%(.00%)	16.0%(3.03%)	.00%(.00%)	0%
	Oscar	1.20(.07)	85.0%(5.78%)	.00%(.00%)	96.4%(2.60%)	.00%(.00%)	0%
	Enet	1.51(.13)	1.42%(.44%)	.00%(.00%)	2.84%(.87%)	.00%(.00%)	0%
	aGrace	1.11(.05)	9.37%(3.31%)	.00%(.00%)	17.7%(5.93)	.00%(.00%)	0%
$k = 10$	Ours	1.02(.02)	.00%(.01%)	.00%(.00%)	.00%(.01%)	.00%(.00%)	90%
	GFlasso	1.12(.05)	11.3%(3.31%)	.00%(.00%)	20.7%(5.73%)	.00%(.00%)	0%
	Oscar	1.49(.14)	100%(.00%)	.00%(.00%)	88.7%(3.29%)	.00%(.00%)	0%
	Enet	1.53(.13)	1.39%(.45%)	.00%(.00%)	2.77%(.90%)	.00%(.00%)	0%
	aGrace	1.45(.10)	100%(.00%)	.00%(.00%)	96.7%(.61%)	.00%(.00%)	0%
$k = 100$	Ours	1.02(.02)	.00%(.01%)	.00%(.00%)	.00%(.01%)	.00%(.00%)	85%
	GFlasso	1.16(.06)	100%(.00%)	.00%(.00%)	89.2%(2.71%)	.00%(.00%)	0%
	Oscar	1.49(.12)	100%(.00%)	.00%(.00%)	90.6%(2.81%)	.00%(.00%)	0%
	Enet	1.52(.13)	1.38%(.45%)	.00%(.00%)	2.75%(.88%)	.00%(.00%)	0%
	aGrace	1.45(.11)	100%(.00%)	.00%(.00%)	96.0%(.81%)	.00%(.01%)	0%
Setup 2	Method	PMSE	VFP	VFN	GFP	GFN	%Oracle
$k = 0$	Ours	1.02(.02)	.18%(.54%)	.00%(.00%)	.36%(1.07%)	.00%(.00%)	%75
	GFlasso	1.12(.05)	12.4%(4.47%)	.00%(.00%)	23.1%(7.60%)	.00%(.00%)	0%
	Oscar	1.25(.08)	38.9%(7.93%)	.00%(.00%)	61.3%(9.11%)	.00%(.00%)	0%
	Enet	1.59(.15)	1.92%(.29%)	.00%(.00%)	3.81%(.57%)	.00%(.00%)	0%
	aGrace	1.11(.05)	11.5%(4.16%)	.00%(.00%)	21.6%(7.32%)	.00%(.00%)	0%
$k = 10$	Ours	1.02(.02)	.01%(.01%)	.00%(.00%)	.01%(.02%)	.00%(.00%)	78%
	GFlasso	1.36(.11)	3.57%(1.74%)	.00%(.00%)	6.71%(3.23%)	.00%(.00%)	0%
	Oscar	1.54(.15)	100%(.00%)	.00%(.00%)	83.9(4.96%)	.00%(.00%)	0%
	Enet	1.61(.16)	1.45%(.46%)	.00%(.23%)	2.9%(.91%)	.02%(.22%)	0%
	aGrace	1.51(.12)	100%(.00%)	.00%(.00%)	94.3%(1.3%)	.00%(.00%)	0%
$k = 100$	Ours	1.02(.02)	.01%(.02%)	.00%(.00%)	.01%(.03%)	.00%(.00%)	73%
	GFlasso	1.54(.14)	100%(.00%)	.00%(.00%)	87.6%(3.93%)	.00%(.00%)	0%
	Oscar	1.54(.14)	100%(.00%)	.00%(.00%)	87.8%(3.78%)	.00%(.00%)	0%
	Enet	1.61(.16)	1.45%(.46%)	.00%(.23%)	2.9%(.91%)	.02%(.22%)	0%
	aGrace	1.51(.13)	100%(.00%)	.00%(.00%)	95.1%(.84%)	.00%(.02%)	0%

Table 1.4: Performance of our methods after adding k ($k = 0, 2, 10$) erroneous edges for each informative predictors in Example 2.

Eroneous edges	PMSE	VFP	VFN	GFP	GFN	%Oracle
0	20.55(2.23)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)	67%
2	20.63(2.16)	.02%(.05%)	.93%(2.29%)	.03%(.10%)	.93%(2.27%)	58%
10	20.64(2.17)	.01%(.05%)	3.00%(4.78%)	.03%(.11%)	2.98%(4.74%)	35%

GFN. Interestingly, our algorithm gives a high percentage of reconstructing the oracle estimator across all the situations, indicating that it has a high chance to produce a global minimizer that is the oracle estimator with a high probability as suggested by Theorem 3. In fact, our method has a recovery rate between 100% and 85% in set-up 1, whereas it has a rate from 78% to 73% in set-up 2. Note that the recovery percentage depends on the design matrix. Overall, the level of difficulty for set-up 2 is higher, because of stronger correlations between informative and noisy predictors.

Compared to other methods, *GFlasso* and aGrace perform slightly worse in parameter estimation but much worse in terms of oracle reconstruction. These methods seem sensitive to erroneous edges in the graph, especially in setup 2 where correlation between informative and noise variables incur bias to *GFlasso*. Finally, neither OSCAR nor Elastic Net performs well, because OSCAR is heavily biased and Elastic Net has not utilized the informative knowledge specified by the graph.

Next we investigate sensitivity of erroneous edges of the specified graphs on performance of a method. As suggested by Table 1.4, the oracle recovery rate dips from 67% to 35% as the average number of erroneous edges increases from 0 to 10 in Example 1. However, in Example 2, the proposed method does not seem sensitive, giving nearly unchanged PMSEs and small differences in the oracle recovery rate, where the error variance is much smaller with $\sigma_e^2 = 1$ compared to $\sigma_e^2 = 20$ in Example 1. The performance of aGrace and *GFlasso* deteriorate significantly, as the number of erroneous edges increases from 10 to 100 for each informative node. For aGrace, it has an elevated PMSE value from 1.11 to 1.45 and 1.55 in set-ups 1 and 2. This is expected because aGrace incurs additional bias through erroneous edges. For *GFlasso*, its PMSE values

increase from 1.12 to 1.16 for $k = 100$ in set-up 1, but from 1.12 to 1.36 for $k = 10$ and to 1.54 for $k = 100$. This is also expected because $GFlasso$ uses the correlations among variables as weights to alleviate bias, which can be affected by erroneous edges between correlated predictors.

Finally, based on Theorem 2, Corollary 1 and our numerical experience, in addition to the graph specification, the oracle recovery probability depends on error variance σ^2, the level of difficulty η^2, sample size n and the number of predictors p. Our numerical results suggest that our "sufficiently precise" condition for oracle recovery may be a bit conservative but is still qualitatively correct in that given the rest are the same, the less erroneous edges one have in the graph, the better chance one can recover the oracle.

Example 3 (Illustration of Corollary 1) The error bound in Corollary 1 suggests that the recovery rate depends on the number of groups K_0 and the level of difficulty η^2. We now perform a simulation study to confirm. Consider two scenarios

$$\beta^0 = \big(\underbrace{1,\ldots,1}_{p_0/K_0},\underbrace{2,\ldots,2}_{p_0/K_0},\cdots,\underbrace{K_0,\ldots,K_0}_{p_0/K_0},\underbrace{0,\ldots,0}_{p-p_0}\big)^T, \quad \eta^2 = 1/2.$$

$$\beta^0 = \big(\underbrace{3,\ldots,3}_{p_0/K_0},\underbrace{6,\ldots,6}_{p_0/K_0},\cdots,\underbrace{3K_0,\ldots,3K_0}_{p_0/K_0},\underbrace{0,\ldots,0}_{p-p_0}\big)^T, \quad \eta^2 = 9/2$$

with $p_0 = 100$, $p = 1000$ and $K_0 = 2, 5, 10, 20$. The correlation structure remains the same as in Example 1 but has within-group correlation .9 with $n = 200$.

Table 1.5: Performance of our methods with different numbers of groups and different levels of difficulty in Example 3.

	$\gamma_{min} = 1$		$\gamma_{min} = 3$	
# groups	PMSE	%Oracle	PMSE	%Oracle
2	1.01(.02)	97%	1.01(.02)	91%
5	1.04(.03)	30%	1.03(.02)	76%
10	1.10(.05)	4%	1.06(.03)	79%
20	1.28(.09)	0%	1.15(.09)	32%

As suggested by Table 1.5, the oracle recovery rate deteriorates dramatically in both scenarios as K_0 increases from 2 to 20, as well as PMSE. Moreover, the recovery rate in

the second scenario is higher with a smaller PMSE. This is in agreement with Corollary 1.

In conclusion, the proposed method performs well against its competitors in terms of parameter estimation and identifying grouping structures. In addition, it is less sensitive to the imprecise graph knowledge.

1.5.2 Data analysis: eQTL data

To study genetic variation, one important approach is identifying DNA sequence elements controlling gene expressions. By treating a gene's expression as a quantitative trait, one can identify DNA loci regulating the gene expression, called eQTL, which bridges the gap between genetic variants and clinical outcomes, providing biological insights into molecular mechanisms underlying complex disease missed by genome-wide association studies. Furthermore, there is increasing evidence showing that eQTLs are more likely to be disease risk loci, or can be used to boost statistical power to detect disease loci [21, 22]. Such a genome-scale study utilizes DNA single nucleotide polymorphisms (SNPs) and gene expression data. The current practice of eQTL analysis is limited to simple single gene-single SNP analysis, which ignores joint effects of multiple SNPs. Here we apply the proposed method for a single gene-multiple SNP analysis.

Table 1.6: Mean prediction error (PE), number of non-zero regression coefficient estimates, percentage of grouping s, for four competing methods, in the eQTL analysis for gene GLT1D1 in Section 5.2.

	Tuning			Final Model	
Method	PE	# non-zeros	% grouping s	# non-zeros	% grouping s
Lasso	0.93(0.07)	6.67(2.08)	0(0)	3	0
OSCAR	0.90 (0.07)	42.67(17.90)	0.26(0.17)	16	0.01
TLP	0.87(0.01)	1.33(0.58)	0(0)	1	0
Fuse	0.87(0.01)	1.33(0.58)	0(0)	1	0
Ours	0.85(0.04)	1.66(0.58)	0.67(0.58)	2	1

Our focus here is mapping *cis*-acting DNA variants for a representative gene, GLT1D1. As in [16], we pre-process the data, and select SNPs lying within 500kb upstream of the transcription start site (TSS) and 500kb downstream of the transcription end site

(TES) of gene GLT1D1. After monomorphic SNPs are removed, 1782 SNPs remain. As discussed in [16], the standard approach uses a univariate (or marginal) least squares (U-OLS) by regressing the expression level of GLT1D1 on each of the SNPs, coded as 0, 1 and 2, representing the count of the minor allele for the SNP. It is known that the standard approach has some potential drawbacks for data of this type. First, physically nearby SNPs tend to be correlated due to linkage disequilibrium. As a result, a true causal SNP may introduce spurious associations of its nearby SNPs with gene expressions, leading to false positives. Second, most of the genes are regulated by multiple factors or loci. This means that a univariate analysis considering only one SNP a time can be inefficient. To overcome these issues, we consider high-dimensional linear regression with the expression of gene GLT1D1 as our response and 1782 SNPs as our predictors, where simultaneous grouping pursuit and feature selection is performed, and a graph is constructed based on pairwise sample correlations exceeding a cut-off 0.6; see Figure 1 for display a subnetwork. Although this cut-off is somewhat arbitrary, it has been used to construct co-expression networks [23].

For our SNPs data, the number of SNPs $p = 1782$ is much larger than n, but biologically only a few SNPs are expected to be relevant and the correlation structure of physically nearby SNPs needs to be considered. This makes a compelling case for simultaneous grouping pursuit and feature selection to build a simpler model with higher predictive accuracy. To capture the correlation structure induced by physical locations of SNPs, a graph is constructed based on pairwise sample correlations, with a correlation stronger than 0.6 being connected; see Figure 1 for a display of the graph. Also considered is a fused type of graph, defined by a consecutive series order as in the Fused Lasso. For a comparison, we also examine the Lasso, TLP and OSCAR, where the first two perform feature selection alone and the last one does grouping pursuit and feature selection. For each method, the tuning parameter selection is achieved by randomly dividing the samples into two subset, one training set consisting of 140 samples, one tuning set consisting of 70 samples. Then, by applying the cross-validated model to the whole data set, the prediction error (PE)'s are computed, as well as the numbers of nonzero regression coefficients and homogeneous groups, based on the tuning set for the expressions of GLT1D1.

As suggested in Table 1.6, the proposed method not only yields a parsimonious

Table 1.7: Parameter estimation for the final model in Section 5.2, where only nonzero coefficients are displayed.

	Estimates		
Method	$\hat{\beta}_{787}$	$\hat{\beta}_{790}$	$\hat{\beta}_{1667}$
Lasso	0.064	2.451	-0.101
OSCAR	1.439	1.439	-0.347
TLP	0	5.090	0
Our-Fuse	0	5.090	0
Ours	2.874	2.874	0

model with the smallest mean PE but also includes one pair of physically nearby SNPs. To confirm our analysis, note that the proposed method and TLP, the proposed method with $\lambda_2 = 0$, both tend to include a subset of those SNPs having significant p-values in the marginal analysis of [16]. In contrast, the Lasso and TLP identify no grouping structure, and OSCAR is less parsimonious, including many more SNPs with less significant marginal p-values.

Our final model contains one pair of physically nearby SNPs, locations 787 and 790; see Table 1.7. Interestingly, adjacent locations 788 and 789 are not included in the model, because of their small pairwise sample correlations with the other nearby locations. By comparison, the fused type of graph does not seem promising, and other methods include more isolated locations. Our statistical result can be cross-validated biologically through a confirmative experiment focusing on the SNP regions near locations 787-790.

1.6 Discussion

Chapter 1 proposes a method for high-dimensional least square regression, performing simultaneous grouping pursuit and feature selection over an undirected graph describing grouping information *a priori*. Our theoretical analysis indicates that the proposed method as well as its computational surrogate reconstructs the *oracle estimator* even in difficult situations involving highly-correlated predictors when the graph is precise enough. Our numerical analysis suggests that the proposed method outperforms its

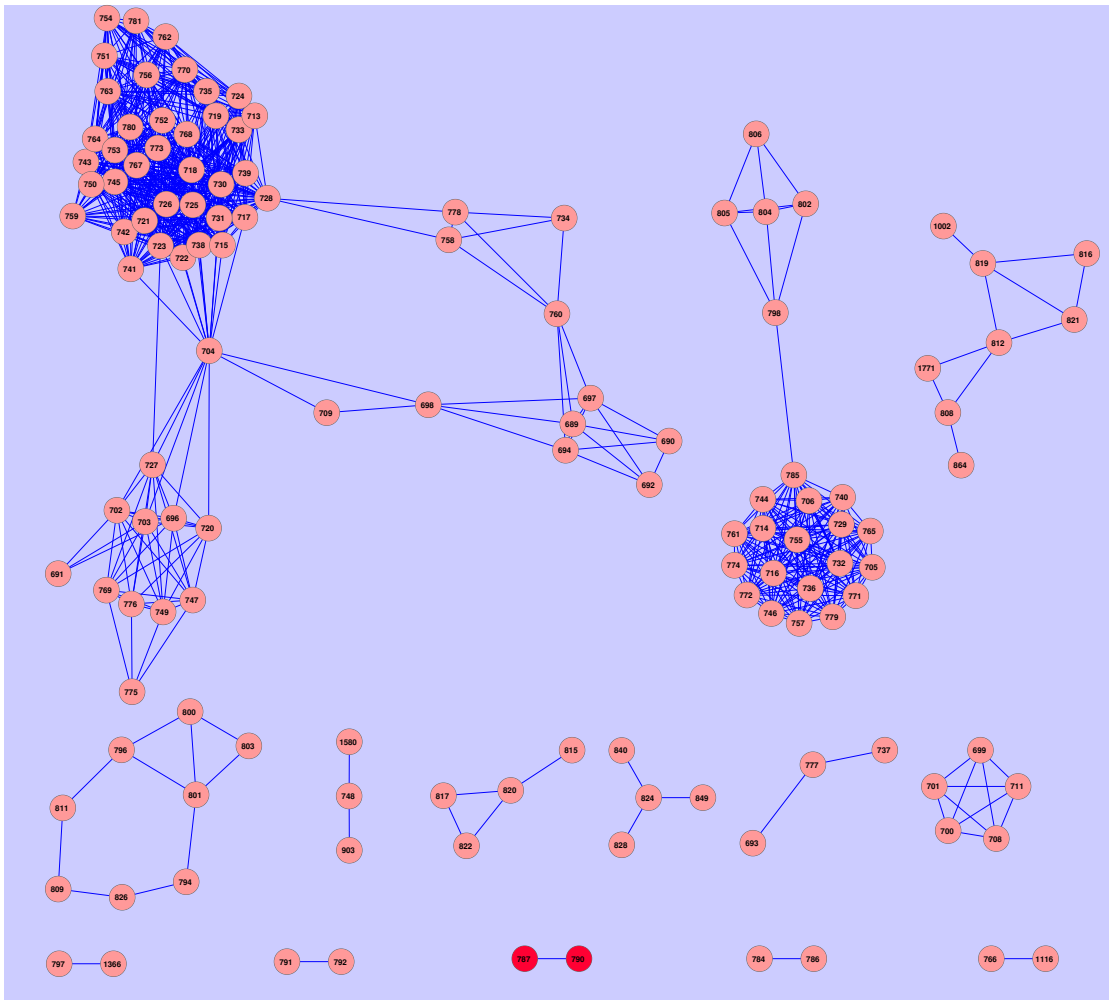


Figure 1.1: Subnetwork consisting of SNPs around informative locations, defined by correlation stronger than .6. Here SNP's locations are numbered with adjacent numbers indicating nearby locations.

competitors in accuracy of selection in addition to estimation. In particular, we have illustrated the application of our method to a single gene-multilocus eQTL analysis; its natural extension is to multiple gene-multilocus eQTL analysis, as advocated by [24, 25], though our method differs from the former two in that ours is built in a general framework of penalized regression.

In order for the proposed method to be useful, further investigation is necessary to understand the interplay between grouping pursuit and feature selection.

Chapter 2

Multiple Gaussian graphical models with grouping penalties

2.1 Introductions

Graphical models are widely used to describe relationships among interacting units. Major components of the models are nodes that represent random variables, and edges encoding conditional dependencies between the nodes. Of great current interest is the identification of certain lower-dimensional structures for undirected graphs. The central topic of this chapter is maximum penalized likelihood estimation of multiple Gaussian graphical models for simultaneously pursuing two disparate kinds of structures– sparseness and clustering.

In the literature on Gaussian graphical models, the current research effort has concentrated on reconstruction of a *single* sparse graph. Methods to exploit matrix sparsity include [26, 27, 28, 29, 30, 31, 13], among others. For *multiple* Gaussian graphical models, existing approaches mainly focus on either exploring temporal smoothing structure [32, 5] or encouraging common sparsity across the networks [33, 34]. In this chapter, we focus on pursuing both clustering and sparseness structures over multiple graphs, including temporal clustering as a special case while allowing for abrupt changes of structures over graphs. For multiple graphs without a temporal ordering, our method enables to identify possible element-wise heterogeneity among undirected graphs. This

is motivated by heterogeneous gene regulatory networks corresponding to disparate cancer subtypes [35, 36]. In such a situation, the overall associations among genes remain similar for each network, whereas specific pathways and certain critical nodes (genes) may be differentiated under disparate conditions.

For multiple Gaussian graphical models, estimation is challenging due to enormous candidate graphs of order 2^{Lp^2}, where p is the total number of nodes and L is total number of graphs. To battle the curse of dimensionality, we explore two dissimilar types of structures simultaneously: (1) sparseness within each graph and (2) element-wise clustering across graphs. The benefit of this exploration is three-fold. First, it goes beyond sparseness pursuit alone for each graph, which is usually inadequate given a large number of unknown parameters relative to the sample size, as demonstrated in four numerical examples in Section 5. Second, borrowing information across graphs enables us to detect the changes of sparseness and clustering structures over the multiple graphs. Third, pursuit of these two structures at the same time is suited for our problem, which seeks both similarities and differences among the multiple graphs.

To this end, we propose a regularized/constrained maximum likelihood method for simultaneous pursuit of sparseness and clustering structures. Computationally, we develop a strategy to convert the optimization involving matrices to a sequence of much simpler quadratic problems. Most critically, we derive a necessary and sufficient partition rule to partition the nodes into disjoint subproblems excluding zero-coefficient nodes for multiple arbitrary graphs with convex relaxation, where the rule is applied before computation is performed. Similar rule has been used in [37] for convex estimation of a single matrix. For multiple precision matrices estimation, [38] derived a similar result, but their proof strategies seem to be quite different to that we used here. This partition rule makes efficient computation possible for multiple large graphical models, which otherwise is rather difficult if not impossible. Theoretically, we develop a novel theory for the proposed method, and show that it enables to reconstruct the oracle estimator as if the true sparseness and element-wise clustering structures were given *a priori*, which leads to reconstruction of the two types of structures consistently. This occurs roughly when the size of L matrices p^2L is of order $\exp(An)$, where p is the dimension of the matrices and A is related to the Hessian matrices of the negative

log-determinant of the true precision matrices and the resolution level for simultaneous pursuit of sparseness and element-wise clustering, c.f., Corollary 3. Moreover, we quantify the improvement due to structural pursuit beyond that of sparsity.

The rest of this dissertation is organized as follows. Section 2 introduces the proposed method. Section 3 is devoted to estimation of partial correlations across multiple graphical models, and develops computational tools for efficient computation. Section 4 presents a theory concerning the accuracy of structural pursuit and parameter estimation, followed by some numerical examples in Section 5 and an application to signaling network inference in Section 6. Section 7 discusses various issues in modeling. Finally, the appendix contains proofs.

2.2 Statistical methodology

This chapter introduces the proposed statistical methodology for estimating multiple precision matrices. Consider the $L-$sample problem with the $l-$th sample $\boldsymbol{X}_1^{(l)}, \cdots, \boldsymbol{X}_{n_l}^{(l)}$ from $\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$; $l = 1, \cdots, L$, we estimate $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_L)$, where $\boldsymbol{\Omega}_l = \boldsymbol{\Sigma}_l^{-1}$ is the $p \times p$ inverse covariance matrix and positive definite, denoted by $\boldsymbol{\Omega}_l \succ 0$, $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are the corresponding mean vector and covariance matrix, and the sample size $n = \sum_{l=1}^{L} n_l$.

For maximum likelihood estimation, the profile likelihood for $\boldsymbol{\Omega}$, after $\mu_1, \cdots, \mu_L$ are maximized out, is proportional to

$$\sum_{l=1}^{L} n_l \big(\log\det(\boldsymbol{\Omega}_l) - tr(\boldsymbol{S}_l \boldsymbol{\Omega}_l) \big), \tag{2.1}$$

where $\bar{\boldsymbol{X}}_l = n_l^{-1} \sum_{i=1}^{n_l} \boldsymbol{X}_i^{(l)}$ and $\boldsymbol{S}_l = n_l^{-1} \sum_{i=1}^{n_l} (\boldsymbol{X}_i^{(l)} - \bar{\boldsymbol{X}}_l)(\boldsymbol{X}_i^{(l)} - \bar{\boldsymbol{X}}_l)^T$ are the corresponding sample mean and covariance matrix, det and tr denote the determinant and trace. In (2.1), the number of unknown parameters in $\boldsymbol{\Omega}$ can greatly exceed the sample size n.

2.2.1 General penalized multiple precision matrices estimation

To avoid non-identifiability in (2.1) and encourage low dimensional structures, we propose a regularized maximum likelihood approach through penalty functions $J_{jk}(\cdot)$:

$$\text{maximize}_{\mathbf{\Omega}\succ 0}\ S(\mathbf{\Omega}) = \sum_{l=1}^{L} n_l\big(\log\det(\mathbf{\Omega}_l) - tr(\boldsymbol{S}_l\mathbf{\Omega}_l)\big) - \sum_{j\neq k} J_{jk}(\omega_{jk1},\cdots,\omega_{jkL}), \quad (2.2)$$

where $\mathbf{\Omega} = (\mathbf{\Omega}_1,\cdots,\mathbf{\Omega}_L)$ and only the off-diagonals $\{\omega_{jkl}\}$ of $\mathbf{\Omega}_l$ are regularized. Note that $J_{jk}(\cdot)$ could be any function that penalizes jk-th entries across $\mathbf{\Omega}_l$'s. This encompasses many existing penalty-based approaches for multiple Gaussian graphical models [33, 39] as special cases.

In general, the maximization problem (2.2) involving L matrices is computationally difficult. To meet the computational challenges, we develop a general block-wise coordinate descent strategy to reduce (2.2) to an iterative procedure involving much easier subproblems. Before proceeding, we introduce some notations. Let the jth row (or column) of $\mathbf{\Omega}_l$ be $\boldsymbol{\omega}_{jl}$, let $\boldsymbol{\omega}_{-jl} = (\omega_{j1l},\ldots,\omega_{j(j-1)l},\omega_{j(j+1)l},\ldots,\omega_{jpl})$ be a $(p-1)$-dimensional vector, excluding the jth component of $\boldsymbol{\omega}_{jl}$, and $\mathbf{\Omega}_{-jl}$ be the sub-matrix without the jth row and column of $\mathbf{\Omega}_l$, and $\mathbf{\Omega}_{-jl}^{-1}$ be the inverse of $\mathbf{\Omega}_{-jl}$.

Our proposed method maximizes (2.2) by sweeping each row (or column) of $\mathbf{\Omega}$ across $l = 1,\cdots,L$. Using the property that $\det(\mathbf{\Omega}_l) = \det\big(\mathbf{\Omega}_{-jl}(\boldsymbol{\omega}_{jjl} - \boldsymbol{\omega}_{-jl}^T\mathbf{\Omega}_{-jl}^{-1}\boldsymbol{\omega}_{-jl})\big)$ with ^{T} indicating the transpose, we rewrite (2.2), after ignoring constant terms, as a function of each row (or column) $(\boldsymbol{\omega}_{j1},\cdots,\boldsymbol{\omega}_{jl})$ across l; $j = 1,\cdots,p$,

$$\sum_{l=1}^{L} n_l\big(\log\big(\omega_{jjl} - \boldsymbol{\omega}_{-jl}^T\mathbf{\Omega}_{-jl}^{-1}\boldsymbol{\omega}_{-jl}\big)\big) - s_{jjl}\omega_{jjl} - 2\boldsymbol{s}_{-jl}^T\boldsymbol{\omega}_{-jl}\big) - \sum_{k\neq j} J_{jk}(\omega_{jk1},\cdots,\omega_{jkL}) \quad (2.3)$$

First, for each fixed row (or column) of $\mathbf{\Omega}$ across $l = 1,\cdots,L$, we maximize (2.3) over the diagonals $(\omega_{jj1},\cdots,\omega_{jjl})$ given the corresponding off-diagonals $(\boldsymbol{\omega}_{-j1},\cdots,\boldsymbol{\omega}_{-jl})$. Setting the partial derivatives of (2.3) in the diagonals to be zero yields the profile maximizer of (2.2)

$$\hat{\omega}_{jjl} \;=\; 1/s_{jjl} + \boldsymbol{\omega}_{-jl}^T\mathbf{\Omega}_{-jl}^{-1}\boldsymbol{\omega}_{-jl}, \quad l = 1,\cdots,L. \quad (2.4)$$

Second, substituting (2.4) into (2.3) yields the negative profile likelihood of (2.2) for

$(\boldsymbol{\omega}_{-jl}, \cdots, \boldsymbol{\omega}_{-jl})$

$$\sum_{l=1}^{L} n_l \big(s_{jjl} \boldsymbol{\omega}_{-jl}^T \boldsymbol{\Omega}_{-jl}^{-1} \boldsymbol{\omega}_{-jl} + 2 \boldsymbol{s}_{-jl}^T \boldsymbol{\omega}_{-jl} \big) + \sum_{k \neq j} J_{jk}(\omega_{jk1}, \cdots, \omega_{jkL}). \tag{2.5}$$

Third, the aforementioned process is repeated for each rows (or columns) of $\boldsymbol{\Omega}$ until a certain stopping criterion is satisfied. By Theorem 1, profiling is equivalent to the original problem for separable convex penalty functions summarized as follows.

Theorem 4 *Iteratively minimizing* (2.5) *over the off-diagonals* $(\boldsymbol{\omega}_{-j1}, \cdots, \boldsymbol{\omega}_{-jL})$ *and updating diagonals* ω_{jjl} *by* (2.4)*;* $j = 1, \cdots, p, l = 1, \cdots, L$ *converges to a local maximizer of* (2.2)*. Moreover, if* $J_{jk}(\cdot)$ *are convex, it converges to a global maximizer.*

Theorem 1 reduces (2.2) to iteratively solving (2.5) which is quadratic in its argument. On this ground we design efficient methods for solving (2.2) with a specific choice of $J_{jk}(\cdot)$ next.

2.2.2 Pursuit of sparseness and clustering structures

A zero element in $\boldsymbol{\Omega}_l$ corresponds to conditional independence between two components of $Y^{(l)}$ given its other components [40]. Thus, within each precision matrix $\boldsymbol{\Omega}_l$, estimating its elements reconstructs its graph structure, where a zero-element of $\boldsymbol{\Omega}_l$ corresponds to no edges between the two nodes, encoding conditional independence. In addition, the nodes connecting many other nodes are identified, called network hubs. On the other hand, over multiple precision matrices, estimating element-wise clustering structure can reveal the change of sparseness and clustering structures.

To detect clustering structures, consider element-wise clustering of entries of $\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_L$ based on possible prior knowledge. The prior knowledge is specified loosely in an undirected graph $\mathcal{U}$ with each node corresponding to a triplet (j, k, l); $1 \leq j < k \leq p$, $1 \leq l \leq L$. That is, an edge between node (j, k, l) and (j, k, l') means that the (j, k)th entry of $\boldsymbol{\Omega}_l$ and the (j, k)th entry of $\boldsymbol{\Omega}_{l'}$ tend to be similar *a priori* and thus can be pushed to share the same value. Specifically, let $\mathcal{E}_{jk}$ denote a set of edges between two distinct nodes $(j, k, l) \neq (j, k, l')$ of $\mathcal{U}$, where $(l, l') \in \mathcal{E}_{jk}$ indicates a connection between the two nodes $(j, k, l), (j, k, l')$. To identify homogeneous subgroups of off-diagonals $\{\omega_{jkl}\}$

of $\mathbf{\Omega}_l$ across $l = 1, \cdots, L$ over $\mathcal{U}$, including the group of zero-elements, we propose a non-convex penalty of the form

$$J_{jk}(\omega_{jk1}, \cdots, \omega_{jkL}) = \lambda_1 \sum_{l=1}^{L} J_\tau(|\omega_{jkl}|) + \lambda_2 \sum_{(l,l') \in \mathcal{E}_{jk}} J_\tau(|\omega_{jkl} - \omega_{jkl'}|), \tag{2.6}$$

to regularize (2.2), where λ_1 and λ_2 are nonnegative tuning parameters controlling the degrees of sparseness and clustering, $J_\tau(z) = \min(|z|, \tau)$ is the truncated L_1-penalty of [17], called TLP in what follows, which, after rescaled by $\frac{1}{\tau}$, approximates the L_0-function when tuning parameter $\tau > 0$ tends to 0^+.

Note that our approach is applicable to a variety of applications by specifying the graph $\mathcal{U}$. For time varying graphs, our method can be used to detect the change of clustering structure, where $\mathcal{E}_{jk}$ is a serial graph as in the fused Lasso [8], and a serial temporal relation is defined only for elements in adjacent matrices. One key difference between our method and the smoothing method [5, 32] is that it enables to accommodate abrupt changes of structures over networks. For multiple graphs without a serial ordering, the proposed method enables to identify possible element-wise heterogeneity among undirected graphs, such as gene regulatory networks corresponding to disparate cancer subtypes [35, 36]. Heterogeneity of this type can be dealt with by specifying a complete graph for each $\mathcal{E}_{jk}$.

2.3 Computational methods

This chapter proposes a relaxation method to treat non-convex penalties in (2.6). For large-scale problems, a partition rule may be useful, which breaks large matrices into many small ones to process separately. A novel necessary and sufficient partition rule is derived for our non-convex penalization method as well as its convex counterpart, generalizing the results for single precision matrix estimation [37, 39].

2.3.1 Non-convex optimization

For the non-convex minimization (2.2) with (2.6), we develop a relaxation method by solving a sequence of convex problems. This method integrates difference convex (DC)

programming with block-wise coordinate descent method based on the foregoing strategy.

For DC programming, we first decompose $S(\mathbf{\Omega})$ into a difference of two convex functions: $S(\mathbf{\Omega}) = S_1(\mathbf{\Omega}) - S_2(\mathbf{\Omega})$, with

$$\begin{aligned} S_1(\mathbf{\Omega}) &= \sum_{l=1}^{L} n_l\big(\log\det(\mathbf{\Omega}_l) - tr(\boldsymbol{S}_l\mathbf{\Omega}_l)\big) + \lambda_1 \sum_{(j,k,l):j\neq k} |\omega_{jkl}| \\ &\quad + \lambda_2 \sum_{1\le j\neq k\le p} \sum_{(l,l')\in\mathcal{E}_{jk}} |\omega_{jkl} - \omega_{j'k'l'}|, \\ S_2(\mathbf{\Omega}) &= \sum_{j\neq k} \lambda_1 \sum_{l=1}^{L} \max(|\omega_{jkl}| - \tau, 0) \\ &\quad + \lambda_2 \sum_{1\le j\neq k\le p} \sum_{(l,l')\in\mathcal{E}_{jk}} \max(|\omega_{jkl} - \omega_{j'k'l'}| - \tau, 0), \end{aligned} \tag{2.7}$$

where a DC decomposition of $J_\tau(|z|) = |z| - \max(|z| - \tau, 0)$ is used. Then the trailing convex function $S_2(\mathbf{\Omega})$ is iteratively approximated by its minorization, say at iteration m, $\lambda_1 \sum_{l=1}^{L}\sum_{j\neq k}(\mathbb{I}(|\hat{\omega}_{jkl}^{(m)}| \le \tau)|\omega_{jkl}| + \lambda_2 \sum_{1\le j\neq k\le p}\sum_{(l,l')\in\mathcal{E}} \mathbb{I}(|\hat{\omega}_{jkl}^{(m)} - \hat{\omega}_{jkl'}^{(m)}| \le \tau)|\omega_{jkl} - \omega_{jkl'}|$. This is obtained through minorization $|z^{(m)}| + \zeta(|z^{(m)}|)(|z| - |z^{(m)}|$ of $\max(|z| - \tau, 0)$ at $|\hat{\omega}_{jkl}^{(m)}|$, which is the solution at iteration $m-1$, where $\zeta(|z^{(m)}|)$ is the gradient of $\max(|z| - \tau, 0)$ at $|z^{(m)}|$; see [17] for more discussions about minorization of this type. At iteration m, the cost function to minimize is

$$\begin{aligned} -\sum_{l=1}^{L} n_l\Big(\log\det(\mathbf{\Omega}_l) - tr(\boldsymbol{S}_l\mathbf{\Omega}_l)\Big) &+ \lambda_1 \sum_{(j,k,l)\in E^{(m)}} |\omega_{jkl}| \\ &+ \lambda_2 \sum_{\{(j,k,l),(j,k,l')\}\in F^{(m)}} |\omega_{jkl} - \omega_{jkl'}| \end{aligned} \tag{2.8}$$

subject to $\mathbf{\Omega}_l \succ 0$; $l = 1, \cdots, L$, where $E^{(m)} = \{(j,k,l) : |\hat{\omega}_{jkl}^{(m)}| \le \tau, j \neq k\}$; $F^{(m)} = \{\{(j,k,l),(j,k,l')\} : (l,l') \in \mathcal{E}_{jk}, |\hat{\omega}_{jkl}^{(m)} - \hat{\omega}_{jkl'}^{(m)}| \le \tau\}$.

To solve (2.8), we apply Theorem 1 to iteratively minimize:

$$\begin{aligned} \sum_{l=1}^{L} n_l\big(s_{jjl}\boldsymbol{\omega}_{-jl}^T\mathbf{\Omega}_{-jl}^{-1}\boldsymbol{\omega}_{-jl} + 2\boldsymbol{s}_{-jl}^T\boldsymbol{\omega}_{-jl}\big) &+ \lambda_1 \sum_{(j,k,l)\in E^{(m)}} |\omega_{jkl}| \\ &+ \lambda_2 \sum_{\{(j,k,l),(j,k,l')\}\in F^{(m)}} |\omega_{jkl} - \omega_{jkl'}|, \end{aligned} \tag{2.9}$$

and update diagonal elements using (2.4). This quadratic problem can then be efficiently solved using augmented Lagrangian methods as in [41]. Unlike the coordinate descent method updating one component at a time, we update one component of $\boldsymbol{\zeta}$ and two components $(\omega_{jkl}, \omega_{jjl})$ for $\boldsymbol{\Omega}_l$ at the same time.

In (2.9), computation of $\boldsymbol{\Omega}_{-jl}^{-1}$ by directly inverting $\boldsymbol{\Omega}_{-jl}$ has a complexity of $O(p^3)$ operations for each (j, l). For efficient computation, we utilize the special property of our sweeping operator in that the $(p-1)^2$ elements of $\boldsymbol{\Omega}_l$ are unchanged except one row and one column are swept, in addition to the rank one property for updating the formula. In (2.9), we derive an analytic formula through block-wise inversion and the Neumann formula of a square matrix, to compute $(\boldsymbol{\Omega}_{-jl})^{-1}$ from $(\omega_{jjl}, \boldsymbol{\omega}_{-jl}, \boldsymbol{\Omega}_l^{-1})$ and $\boldsymbol{\Omega}_l^{-1}$ from $(\omega_{jjl}, \boldsymbol{\omega}_{-jl}, (\boldsymbol{\Omega}_{-l}^{-1})_{-j})$ for each (j, l). That is,

$$(\boldsymbol{\Omega}_{-jl})^{-1} = (\boldsymbol{\Omega}_l^{-1})_{-j} - \frac{(\boldsymbol{\Omega}_l^{-1})_j (\boldsymbol{\Omega}_l^{-1})_j^T}{(\boldsymbol{\Omega}_l^{-1})_{jj}}, \tag{2.10}$$

$$\boldsymbol{\Omega}_l^{-1} = \begin{pmatrix} (\boldsymbol{\Omega}_l^{-1})_{-j} + b\boldsymbol{a}\boldsymbol{a}^T & -b\boldsymbol{a} \\ -b\boldsymbol{a}^T & b \end{pmatrix}, \tag{2.11}$$

where $\boldsymbol{a} = (\boldsymbol{\Omega}_l^{-1})_{-j}\boldsymbol{\omega}_{-jl}, b = (\omega_{jjl} - \boldsymbol{a}^T\boldsymbol{\omega}_{-jl})^{-1}$. This amounts to $O(p^2)$ operations.

The foregoing discussion leads to our DC block-wise coordinate descent algorithm through sweeping operations over $p(p-1)$ off-diagonals of $(\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_L)$, with each operation involving the L corresponding off-diagonals.

Algorithm 1:

Step 1. (Initialization) Set $\hat{\boldsymbol{\Omega}}_l^{(0)} = I$; $l = 1, \cdots, L$, $E^{(0)} = \{(j,k,l) : 1 \le j \ne k \le p, 1 \le l \le L\}$, $F^{(0)} = \mathcal{E}$, $m = 0$ and precision tolerance $\epsilon = 10^{-5}$ for **Step 2**.

Step 2. (Iteration) At current iteration m, initialize $\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}}^{(m)}$. Then solve (2.8) applying the block-wise coordinate descent algorithm to update $\boldsymbol{\Omega}$ to yield $\hat{\boldsymbol{\Omega}}^{(m+1)}$. And set $E^{(m+1)} = \{(j,k,l) : |\hat{\omega}_{jkl}^{(m+1)}| \le \tau, j \ne k\}$; $F^{(m+1)} = \{\{(j,k,l),(j',k',l')\} : (l,l') \in \mathcal{E}_{jk}, |\hat{\omega}_{jkl}^{(m+1)} - \hat{\omega}_{j'k'l'}^{(m+1)}| \le \tau\}$. Specifically,

a) For each row (column) index $j = 1, \cdots, p$, compute $\boldsymbol{\Omega}_{-jl}^{-1}$ using (2.10); $l = 1, \cdots, L$. Solve (2.9) to obtain $\hat{\boldsymbol{\omega}}_{-jl}^{(m)}$, and then compute $\hat{\omega}_{jjl}^{(m)}$ through (2.4); $l = 1, \cdots, L$. Update $\boldsymbol{\Omega}_l^{(m)}$ with its jth row replaced by $(\hat{\omega}_{jjl}^{(m)}, \hat{\boldsymbol{\omega}}_{-jl}^{(m)})$ and its jth column by symmetry. Finally update $(\boldsymbol{\Omega}_l^{(m)})^{-1}$ using (2.11). Go to next iteration $j+1$ until all rows of $\boldsymbol{\Omega}_l^{(m)}$ have been swept.

b) Repeat a) until the decrement of the objective function is less than ϵ. After convergence, update $\boldsymbol{\Omega}$ to yield $\hat{\boldsymbol{\Omega}}^{(m+1)} = \boldsymbol{\Omega}$ based on a).

Step 3. (Stopping criterion) Terminate when $E^{(m+1)} = E^{(m)}$ and $F^{(m+1)} = F^{(m)}$, otherwise, repeat **Step 2** with $m = m + 1$.

The overall complexity of Algorithm 1 is of order $O(p^3L^2)$. And real computational time of our algorithm depends highly on values of λ_1, λ_2 and the number of iterations. In Example 1, it takes about 30 seconds for one simulation run with $(p, L) = (200, 4)$ over 100 grids on a 8-core computer with Intel(R) Core(TM) i7-3770 processors and 16GB of RAM.

2.3.2 Partition rule for large-scale problems

This section establishes a necessary and sufficient partition rule for our non-convex penalization method and its convex counterpart using the sample covariances, permitting fast computation for large-scale problems by partitioning nodes into disjoint subsets excluding the zero-coefficient subset then applying the proposed method to each nonzero subset. Such a result exists only for a single matrix or a special case of multiple matrices, c.f., [37, 39].

In what follows, we only consider the case where $\mathcal{E}_{jk} = \mathcal{E}$ are identical. Given this graph $\mathcal{G} = (V, \mathcal{E})$, with $(V = \{1, \cdots, L\}, \mathcal{E} = \mathcal{E}_{jk}, 1 \le j < k \le p)$ denoting the node and edge sets, we write $l \sim l'$ if $(l, l') \in \mathcal{E}$, or two nodes are connected. First consider the convex grouping penalty over $\mathcal{G}$, followed by a general case, where the penalized log-likelihood is

$$\begin{aligned}\sum_{l=1}^{L}\Big(n_l\big(-\log\det(\boldsymbol{\Omega}_l) + tr(\boldsymbol{\Omega}_l\boldsymbol{S}_l)\big)\Big) \quad &+ \quad \lambda_1\sum_{l=1}^{L}\|\boldsymbol{\Omega}_{l,\text{off}}\|_1 \\ &+ \quad \lambda_2\sum_{l\sim l'}\|\boldsymbol{\Omega}_{l,\text{off}} - \boldsymbol{\Omega}_{l',\text{off}}\|_1, \qquad (2.12)\end{aligned}$$

where $\boldsymbol{\Omega}_{l,\text{off}}$ denotes the off-diagonal elements of $\boldsymbol{\Omega}_l$ and $\boldsymbol{S}_l = (s_{jkl})_{1\le j,k\le p}$ are the sample covariance matrices, $l = 1, \cdots, L$.

The next theorem derives a necessary and sufficient condition for the jkth element of $\hat{\boldsymbol{\Omega}}_l$ $\hat{\Omega}_{jkl} = 0$ across $l = 1, \cdots, L$, for $j \in \mathcal{J}$, $k \in \mathcal{J}^c$, where $(\hat{\boldsymbol{\Omega}}_1, \cdots, \hat{\boldsymbol{\Omega}}_L)$ is the minimizer of (2.12), and $\mathcal{J} \subset \{1, \cdots, p\}$ is any subset. This partitions the node set into disjoint subsets of connected nodes, with no connections between these subsets.

Theorem 5 *(Partition rule for (2.12))* $\hat{\Omega}_{jkl} = 0$ *for all* $j \in \mathcal{J}$; $k \in \mathcal{J}^c$ *and* $l = 1, \cdots, L$, *if and only if* $(s_{jk1}, \cdots, s_{jkL}) \in \mathcal{S}$, *for all* $j \in \mathcal{J}, k \in \mathcal{J}^c$, *where* $\mathcal{S} = \big\{\boldsymbol{s} = (s_1, \cdots, s_L) : |\sum_{l\in\mathcal{I}} n_l s_l| \le \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I}, \mathcal{I}^c), \forall \mathcal{I} \subseteq V\big\}$ *with* $d(\mathcal{I}, \mathcal{I}^c) = \sum_{l\in\mathcal{I}, l'\in\mathcal{I}^c} \mathbb{I}\big(l \sim l'\big)$ *denoting the number of edges between the nodes in* $\mathcal{I}$ *and the remaining nodes in* $\mathcal{I}^c$.

Similar results hold for the proposed non-convex regularized estimators.

Theorem 6 *(Partition rule for non-convex regularization) Denote by* $\hat{\boldsymbol{\Omega}}^{dc}$ *the solution obtained from* **Algorithm 1** *for (2.2). Similarly, given any* $\mathcal{J}$, $\hat{\Omega}^{dc}_{jkl} = 0$ *for all* $j \in \mathcal{J}$; $k \in \mathcal{J}^c$; $l = 1, \cdots, L$, *if and only if* $(s_{jk1}, \cdots, s_{jkL}) \in \mathcal{S}$, *where* $\mathcal{S} = \big\{\boldsymbol{s} = (s_1, \cdots, s_L) : |\sum_{l\in\mathcal{I}} n_l s_l| \le \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I}, \mathcal{I}^c), \forall \mathcal{I} \subseteq V\big\}$.

Corollary 2 simplifies the expression of $\mathcal{S}$ for specific graphs.

Corollary 2 *In the cases of the fused graph and the complete graph, we have*

$$\begin{aligned}
\mathcal{S} &= \Big\{\boldsymbol{s} : \Big|\sum_{i=1}^{l} n_i s_i\Big| \le l\lambda_1 + \lambda_2, \quad \Big|\sum_{i=L-l+1}^{L} n_i s_i\Big| \le l\lambda_1 + \lambda_2, l = 1, \cdots, L-1, \\
&\qquad \Big|\sum_{i=l_1+1}^{l_2} n_i s_i\Big| \le (l_2 - l_1)\lambda_1 + 2\lambda_2, 1 \le l_1 < l_2 < L; \quad \Big|\sum_{i=1}^{L} n_i s_i\Big| \le L\lambda_1\Big\}, \\
\mathcal{S} &= \Big\{\boldsymbol{s} : \Big|\sum_{i=1}^{l} n_{k_i} s_{k_i}\Big| \le l\lambda_1 + l(L-l)\lambda_2, \quad \Big|\sum_{i=L-l+1}^{L} n_{k_i} s_{k_i}\Big| \le l\lambda_1 + l(L-l)\lambda_2, \\
&\qquad l = 1, \cdots, L, s_{k_1} \ge \cdots \ge s_{k_L}\Big\}.
\end{aligned}$$

The partition rule is useful for efficient computation, as it may reduce computation cost substantially. It can be used in several ways. First, the rule partitions nodes into disjoint connected subsets through the sample covariances s_{jkl}'s. This breaks the original large problem into smaller subproblems, owing to this necessary and sufficient rule. Second, **Algorithm 1** can be applied to each subproblem independently, permitting parallel computation.

Algorithm 2 integrates the partition rule in Theorem 6 with **Algorithm 1** to make the proposed method applicable to large-scale problems.

Algorithm 2 (A partition version of Algorithm 1):

Step 1. (Screening) Compute the sample-covariance matrix $\boldsymbol{S}_l$; $l = 1, \cdots, L$. Construct a $p \times p$ symmetric matrix $\boldsymbol{T} = (t_{jk})_{1 \le j,k \le p}$, with $t_{jk} = 0$ if $(s_{jk1}, \cdots, s_{jkL}) \in \mathcal{S}$ and $t_{jk} = 1$ otherwise. Treating $\boldsymbol{T}$ as an adjacency matrix of an undirected graph, we compute its maximum connected components to form a partition of nodes $\{\mathcal{J}_1, \cdots, \mathcal{J}_q\}$ using breadth-first search or depth-first search algorithm, c.f., [42]

Step 2. (Subproblems) For $i = 1, \cdots, q$, solve (2.2) for each subproblem consisting of nodes in $\mathcal{J}_i$, by applying **Algorithm 1** to obtain the solution $\hat{\boldsymbol{\Omega}}^{(i)} = \big(\hat{\boldsymbol{\Omega}}_1^{(i)}, \cdots, \hat{\boldsymbol{\Omega}}_L^{(i)}\big); i = 1, \cdots, q$.

Step 3. (Combining results) The final solution $\hat{\boldsymbol{\Omega}}_l = \mathrm{Diag}\big(\hat{\boldsymbol{\Omega}}_l^{(1)}, \cdots, \hat{\boldsymbol{\Omega}}_l^{(q)}\big); l = 1, \cdots, L$.

2.4 Theoretical analysis

This chapter investigates theoretical aspects of the proposed method. First we develop a general theory on maximum penalized likelihood estimation involving two types of L_0-constraints for pursuit of sparseness and clustering. Then we specialize the theory for estimation of multiple precision matrices in Section 4.3. Now consider a constrained L_0-version of (2.2):

$$\max_{\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})} L(\boldsymbol{\theta}), \text{ subject to } \sum_{j=1}^{d} \mathbb{I}(|\beta_j| \neq 0) \le C_1, \sum_{(jj')\in\mathcal{E}} \mathbb{I}\big(|\beta_j - \beta_{j'}| \neq 0\big) \le C_2. \quad (2.13)$$

as well as its computational surrogate

$$\max_{\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})} L(\boldsymbol{\theta}), \text{ subject to } \sum_{j=1}^{d} J_\tau(|\beta_j|) \le C_1, \sum_{(jj')\in\mathcal{E}} J_\tau\big(\big|\beta_j - \beta_{j'}\big|\big) \le C_2, \quad (2.14)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ with $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\eta}$ representing the off-diagonals and diagonals of $\boldsymbol{\Omega}$, and three non-negative tuning parameters (C_1, C_2, τ). Note that **Algorithm 1** yields a local minimizer of (2.14), relaxing it by solving a sequence of convex problems.

In what follows, we will prove that global minimizers of (2.13) and (2.14) reconstruct the ideal *oracle estimator* as if the true sparseness and clustering structures of the precision matrices were known in advance. As a result of the reconstruction, key properties of the oracle estimator are simultaneously achieved by the proposed method.

2.4.1 The oracle estimator and consistent graph

To define the oracle estimator, let $\mathcal{G}(\boldsymbol{\beta})$ denote a partition of $\mathcal{I} \equiv \{1, \cdots, d\}$ by the parameter $\boldsymbol{\beta}$, i.e. $\mathcal{G}(\boldsymbol{\beta}) = (\mathcal{I}_0(\boldsymbol{\beta}), \cdots, \mathcal{I}_{K(\boldsymbol{\beta})}(\boldsymbol{\beta}))$, with $\mathcal{I}_0(\boldsymbol{\beta}) = \mathcal{I} \setminus A(\boldsymbol{\beta})$ and $\mathcal{I}_k(\boldsymbol{\beta})$ satisfying $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j'}$; $j, j' \in \mathcal{I}_k(\boldsymbol{\beta})$; $k = 1, \cdots, K(\boldsymbol{\beta})$, where $K(\boldsymbol{\beta})$ is the number of nonzero clusters and $A(\boldsymbol{\beta}) \equiv \{i : \beta_i \neq 0\}$ is the support of $\boldsymbol{\beta}$. Let $\mathcal{G}^0 = \mathcal{G}(\boldsymbol{\beta}^0)$ be the true partition induced by $\boldsymbol{\beta}^0$, with $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \boldsymbol{\eta}^0)$ the true parameter value and $\boldsymbol{\beta}^0 \in \mathbb{R}^d$.

Definition 4 (Oracle estimator) *Given $\mathcal{G}^0$, the oracle estimator is defined as: $\hat{\boldsymbol{\theta}}^o = (\hat{\boldsymbol{\beta}}^o, \hat{\boldsymbol{\eta}}^o) = \operatorname{argmax}_{\boldsymbol{\beta}: \mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}^0} L(\boldsymbol{\theta})$, the corresponding maximum likelihood estimator.*

In (2.13) and (2.14), the edge set $\mathcal{E}$ of $\mathcal{U}$ is important for clustering. In order for simultaneous pursuit of sparseness and clustering structures to be possible, we may need $\mathcal{U}$ to be consistent with the clustering structure of the true precision matrices. In other words, a consistent graph is a minimal requirement for reconstruction of the oracle estimator, where there must exist a path connecting any nodes within the same true cluster.

Definition 5 (*Consistent* graph $\mathcal{U}$) *An undirected graph $\mathcal{U} = \big(\mathcal{I}, \mathcal{E}\big)$ is consistent with the true cluster $\mathcal{G}^0 = \{\mathcal{I}_0^0, \cdots, \mathcal{I}_{K_0}^0\}$, if the subgraph restricting nodes on $\mathcal{I}_j^0$ is connected; $j = 1, \cdots, K_0$.*

2.4.2 Non-asymptotic probability error bounds

Now we derive a non-asymptotic probability error bound for simultaneous sparseness and clustering pursuit, based on which we prove that (2.13) and (2.14) reconstruct the oracle estimator. This implies consistent identification of the sparseness and clustering structures of multiple graphical models, under one simple assumption, called the degree-of-separation condition.

Before proceeding, we introduce some notations. Given a graph $\mathcal{U} = (\mathcal{I}, \mathcal{E})$, let $\mathcal{S} = \big\{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) : |A(\boldsymbol{\beta})| \leq d_0, C(\boldsymbol{\beta}, \mathcal{E}) \leq c_0, \mathcal{G}(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}^0)\big\}$ be a constrained set with $C(\boldsymbol{\beta}, \mathcal{E}) = \sum_{(jj') \in \mathcal{E}} \mathbb{I}\big(|\beta_j - \beta_{j'}| \neq 0\big)$, where $d_0 = |A^0|$ with $A^0 = A(\boldsymbol{\beta}^0)$ as defined above. Given a partition $\mathcal{G}$, let $\mathcal{S}_{\mathcal{G}} = \big\{\boldsymbol{\theta} \in \mathcal{S} : \mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}\big\}$. Given an index set $A \subseteq \mathcal{I}$, let $\mathcal{S}_A = \big\{\boldsymbol{\theta} \in \mathcal{S} : \boldsymbol{A}(\boldsymbol{\beta}) = A\big\}$. Let $\mathcal{S}_i = \cup_{A: |A^0 \setminus A| = i} \mathcal{S}_A$, $S_i^\star = \max_{A: |A^0 \setminus A| = i} \big|\{\mathcal{G}(\boldsymbol{\beta}) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S}_A\}\big|$; $i = 0, \cdots, d_0$, and $S^* = \exp\big(\max_{0 \leq i \leq d_0} \frac{\log S_i^*}{\max(i,1)}\big)$. Roughly, $S^\star$ quantifies

complexity of the space of candidate precision matrices scaled by the number of nonzero entries.

The degree-of-separation condition will be used to ensure consistent reconstruction of the oracle estimator: For some constant $c_1 > 0$,

$$C_{\min}(\boldsymbol{\theta}^0) \geq c_1 \frac{\log d + \log S^*}{n}, \tag{2.15}$$

where $C_{\min}(\boldsymbol{\theta}^0) \equiv \inf_{\{\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})\in\mathcal{S}\}} \frac{-\log(1-h^2(\boldsymbol{\theta},\boldsymbol{\theta}^0))}{\max(|A^0\setminus A(\boldsymbol{\beta})|,1)}$ with $|\cdot|$ and $\setminus$ denoting the size of a set and that of set difference, respectively, $h(\boldsymbol{\theta},\boldsymbol{\theta}^0) = \big(\frac{1}{2}\int(g^{1/2}(\boldsymbol{\theta},y)-g^{1/2}(\boldsymbol{\theta}^0,y))^2 d\mu(y)\big)^{1/2}$ is the Hellinger-distance for densities with respect to a dominating measure μ.

We now define the bracketing Hellinger metric entropy of space $\mathcal{F}$, denoted by the function $H(\cdot,\mathcal{F})$, which is the logarithm of the cardinality of the u-bracketing (of $\mathcal{F}$) of the smallest size. That is, for a bracket covering $S(\varepsilon,m) = \{f_1^l, f_1^u, \cdots, f_m^l, f_m^u\} \subset \mathcal{L}_2$ satisfying $\max_{1\leq j\leq m}\|f_j^u - f_j^l\|_2 \leq \varepsilon$ and for any $f \in \mathcal{F}$, there exists a j such that $f_j^l \leq f \leq f_j^u$, *a.e.* P, then $H(u,\mathcal{F})$ is $\log(min\{m:\ S(u,m)\})$, where $\|f\|_2 = \int f^2(z)d\mu$. For more discussions about metric entropy of this type, see [43].

Assumption A: (Complexity of the parameter space) For some constant $c_0 > 0$ and any $0 < t < \varepsilon \leq 1$, $H(t,\mathcal{B}_{\mathcal{G}}) \leq c_0(\log p)^2, 1)|A|\log(2\varepsilon/t)$, where $\mathcal{B}_{\mathcal{G}} = \mathcal{F}_{\mathcal{G}} \cap \{h(\boldsymbol{\theta},\boldsymbol{\theta}^0) \leq 2\varepsilon\}$ is a local parameter space, and $\mathcal{F}_{\mathcal{G}} = \{g^{1/2}(\boldsymbol{\theta},y) : \boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\eta}) : \mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}\}$ be a collection of square-root densities indexed by any subset $\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}) : \boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\eta}) \in \mathcal{S}\}$.

Next we present our non-asymptotic probability error bounds for reconstruction of the oracle estimator $\hat{\boldsymbol{\theta}}^0$ by global minimizers of (2.13) and (2.14) in terms of $C_{min}(\boldsymbol{\theta}^0)$, n, d and d_0, where d_0 and d can depend on n. Consistency is established for reconstruction of $\hat{\boldsymbol{\theta}}^0$ as well as structure recovery. Note that $\hat{\boldsymbol{\theta}}^0$ is asymptotically optimal, hence the optimality translates into the global minimizers of (2.13) and (2.14).

Theorem 7 *(Global minimizer of* (2.13)*) Under* **Assumption A**, *if $\mathcal{U}$ is consistent with $\mathcal{G}^0$, then for a global minimizer of (2.13) $\hat{\boldsymbol{\theta}}^{l_0}$ with estimated grouping $\hat{\mathcal{G}}^{l_0} = \mathcal{G}(\hat{\boldsymbol{\beta}}^{l_0})$ at $(C_1,C_2) = (d_0,c_0)$ with $c_0 = C(\boldsymbol{\beta}^0,\mathcal{E})$,*

$$\mathbb{P}\Big(\hat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\Big) = \mathbb{P}\Big(\hat{\boldsymbol{\theta}}^{l_0} \neq \hat{\boldsymbol{\theta}}^o\Big) \ \leq\ \exp\Big(-c_2 n C_{\min}(\boldsymbol{\theta}^0) + 2\log d + \log S^\star\Big). \tag{2.16}$$

Under (2.15), $\mathbb{P}\Big(\hat{\mathcal{G}}^{l_0} = \mathcal{G}^0\Big) = \mathbb{P}\Big(\hat{\boldsymbol{\theta}}^{l_0} = \hat{\boldsymbol{\theta}}^o\Big) \to 1$ *as* $n, d \to \infty$.

For the constrained truncated L_1-likelihood, one additional condition–**Assumption B** is necessary. We requires the Hellinger-distance to be smooth so that the approximation of the truncated L_1-function to the L_0-function becomes adequate by tuning τ.

Assumption B: For some constants d_1-$d_3 > 0$,

$$-\log(1-h^2(\boldsymbol{\theta},\boldsymbol{\theta}^0)) \geq -d_1\log(1-h^2(\boldsymbol{\theta}^\tau,\boldsymbol{\theta}^0)) - d_3\tau^{d_2}d, \tag{2.17}$$

where $\boldsymbol{\theta}^\tau = (\boldsymbol{\beta}^\tau, \eta)$ with $\boldsymbol{\beta}^\tau = (\beta_1^\tau, \cdots, \beta_p^\tau)$, and $\beta_j^\tau = \frac{\sum_{j'\in\mathcal{I}_k}\beta_j'}{|\mathcal{I}_k|}$ for $j \in \mathcal{I}_k(\boldsymbol{\beta})$; $k = 0, 1, \cdots, K(\boldsymbol{\beta})$.

Theorem 8 *(Global minimizer of* (2.14)*) Assume that* **Assumption A** *with* $\mathcal{F}_\mathcal{G}$ *replaced by* $\mathcal{F}_\mathcal{G}^\tau = \{g^{1/2}(\boldsymbol{\theta}, y) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) : \mathcal{G}^\tau(\boldsymbol{\beta}) = \mathcal{G}\}$ *and* **Assumption B** *are met. If* $\mathcal{U}$ *is consistent with* $\mathcal{G}^0$*, then for a global minimizer of* (2.14) $\hat{\boldsymbol{\theta}}^g$ *with estimated grouping* $\hat{\mathcal{G}}^g = \mathcal{G}(\hat{\boldsymbol{\beta}}^g)$ *at* $(C_1, C_2) = (d_0, c_0)$ *with* $c_0 = C(\boldsymbol{\beta}^0, \mathcal{E})$ *and* $\tau \leq \big(\frac{(d_1-c_3)C_{\min}(\boldsymbol{\theta}^0)}{d_3 d}\big)^{1/d_2}$,

$$\mathbb{P}\Big(\hat{\mathcal{G}}^g \neq \mathcal{G}^0\Big) = \mathbb{P}\Big(\hat{\boldsymbol{\theta}}^g \neq \hat{\boldsymbol{\theta}}^o\Big) \quad \leq \quad \exp\Big(-c_3 n C_{\min}(\boldsymbol{\theta}^0) + 2\log d + \log S^\star\Big). \tag{2.18}$$

Under (2.15), $\mathbb{P}\Big(\hat{\mathcal{G}}^g = \mathcal{G}^0\Big) = \mathbb{P}\Big(\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o\Big) \to 1$ *as* $n, d \to \infty$.

2.4.3 An illustrative example

We now apply the general theory in Theorems 2 and 3 to the estimation of multiple precision matrices, in which the true precision matrices in each cluster are the same, with $g_0 \equiv \sum_{l=1}^{L-1}\sum_{j>k}\mathbb{I}\big(\omega_{jkl}^0 \neq \omega_{jk(l+1)}^0\big)$ the number of break points among these clusters. In this case, a serial graph $\mathcal{U}$ is considered for clustering.

Denote by p and L_0 the dimension of the precision matrix and the number of distinctive clusters, respectively. Let $H_l = \Big(\frac{\partial^2(-\log\det(\boldsymbol{\Omega}_l))}{\partial^2\boldsymbol{\Omega}}\Big)\Big|_{\boldsymbol{\Omega}_l=\boldsymbol{\Omega}_l^0}$ be the $p^2 \times p^2$ Hessian matrix of $-\log\det(\boldsymbol{\Omega}_l)$, whose $(jk, j'k')$ element is $tr(\boldsymbol{\Sigma}_l\Delta_{jk}\boldsymbol{\Sigma}_l\Delta_{j'k'})$, c.f., [44]. Define

$$\eta_{min} = \min\Big(\min_{(j,k,l):\omega_{jkl}^0\neq 0}\big|\omega_{jkl}^0\big|, \frac{1}{\sqrt{2}}\min_{(j,k,l):\omega_{jkl}^0\neq\omega_{jk(l+1)}^0}\big|\omega_{jkl}^0 - \omega_{jk(l+1)}^0\big|\Big)$$

to be the resolution level for simultaneous sparseness and clustering pursuit.

An application of Theorems 2 and 3 with $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ being off-diagonals and diagonals of $\boldsymbol{\Omega}$ leads to the following result.

Corollary 3 *(Multiple precision matrices with a serial graph) When $\mathcal{U}$ is a serial graph, all the results in Theorems 2 and 3 for simultaneous pursuit of sparseness and clustering hold under two simple conditions:*

$$C_{min}(\boldsymbol{\theta}^0) \geq c_4 \min_{1\leq l\leq L} c_{min}\big(H_l\big)\eta_{min}^2, \ \ and \ \ \log S^\star \leq 2g_0 \max(\log(d_0/g_0), 1), \quad (2.19)$$

for some constant $c_4 > 0$. Sufficiently, if

$$\min_{1\leq l\leq L} c_{min}\big(H_l\big)\eta_{min}^2 \geq c_0 \frac{\log\big(Lp(p-1)/2\big) - g_0 \max\big(\log(d_0/g_0), 1\big)}{n}, \quad (2.20)$$

holds for some constant $c_0 > 0$, then $\mathbb{P}\Big(\hat{\boldsymbol{\Omega}}^{\ell_0} \neq \hat{\boldsymbol{\Omega}}^o\Big)$ and $\mathbb{P}\Big(\hat{\boldsymbol{\Omega}}^{g} \neq \hat{\boldsymbol{\Omega}}^o\Big) \to 0$ as $n, d \to +\infty$.

Corollary 3 suggests that the amount of reconstruction improvement would be of the order of $1/L$ if the L precision matrices are identical. In general, the amount of improvement of joint estimation over separate estimation is $L/\log(L)$ when g_0 is small, i.e. $g_0 \max\big(\log(d_0/g_0), 1\big) \lesssim \log\big(Lp(p-1)/2\big)$, by contrasting the sufficient condition in (2.20) with that for a separate estimation approach in [17], where $\lesssim$ denotes inequality ignoring constant terms. Here g_0 describes similarity among L precision matrices with a small value corresponding to a high-degree of similarity shared among precision matrices.

2.5 Simulation

This chapter studies operational characteristics of the proposed method via simulation in sparse and nonsparse situations with different types of graphs in both low- and high-dimensional settings. In each simulated example, we compare our method against its convex counterpart for seeking the sparseness structure for each graphical model and identifying the grouping structure among multiple graphical models, and contrast the method against its counterpart seeking the sparseness structure alone. In addition, we also compare against a kernel smoothing method for time-varying networks [32, 5] in Examples 1-3, whenever appropriate. The smoothing method defines a weighted average over sample covariance matrices at time points as $\tilde{\boldsymbol{S}}_l(h) = \frac{\sum_{l'=1}^L w_{ll'}(h)\boldsymbol{S}_l'}{\sum_{l'=1}^L w_{ll'}(h)}$, with $w_{ll'}(h) = K(h^{-1}|l-l'|)$; $l = 1, \cdots, L$, where $K(x) = (1-|x|)\mathbb{I}(|x| < 1)$ is a triangular

kernel, h is a bandwidth, and $l = 1, \cdots, L$ denotes clusters. Then within each cluster l, the precision matrix estimate $\hat{\boldsymbol{\Omega}}_l(h, \lambda)$ is obtained by solving

$$\hat{\boldsymbol{\Omega}}_l(h,\lambda) \;=\; \underset{\boldsymbol{\Omega}_l \succ 0}{\operatorname{argmin}} \Big(-\log\det\big(\boldsymbol{\Omega}_l\big) + tr(\boldsymbol{\Omega}\tilde{\boldsymbol{S}}_l(h)) + \lambda \sum_{j<j'} |\omega_{jj'l}| \Big), \qquad (2.21)$$

using the glasso algorithm [27], and the final estimate is obtained through tuning over (h, λ)-grids. Two performance metrics are used to measure the accuracy of parameter estimation as well as that of correct identification of the sparseness and grouping structures.

In Examples 1-3, temporal clustering pursuit is performed over $\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_L$ through a serial graph $\mathcal{E} = \big\{\{(j,k,l),(j',k',l')\} : j = j', k = k', |l - l'| = 1\big\}$. That is, only adjacent matrices may be possibly clustered. In Example 4, general clustering pursuit is conducted through a complete graph $\mathcal{E} = \big\{\{(j,k,l),(j',k',l')\} : j = j', k = k', l < l'\big\}$.

For the accuracy of parameter estimation, the average entropy loss (EL) and average quadratic loss (QL) are considered, defined as

$$EL = \frac{1}{L}\sum_{l=1}^{L}\Big(tr\big(\boldsymbol{\Omega}_l^{-1}\hat{\boldsymbol{\Omega}}_l\big) - \log\det\big(\boldsymbol{\Omega}_l^{-1}\hat{\boldsymbol{\Omega}}_l\big)\Big), \quad QL = \frac{1}{L}\sum_{l=1}^{L} tr\Big(\big(\boldsymbol{\Omega}_l^{-1}\hat{\boldsymbol{\Omega}}_l - \boldsymbol{I}\big)^2\Big).$$

For the accuracy of identification, average false positive (FPV) and false negative (FNV) rates for sparseness pursuit, as well as those (FPG) and (FNG) for grouping are used:

$$\begin{aligned} FPV &= \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} = 0, \hat{\omega}_{jj'l} \neq 0)}{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} = 0)} \Big(1 - \mathbb{I}\big(\boldsymbol{\Omega}_{l,\text{off}} \neq \boldsymbol{0}\big)\Big) \\ FNV &= \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} \neq 0, \hat{\omega}_{jj'l} = 0)}{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} \neq 0)} \mathbb{I}\big(\boldsymbol{\Omega}_{l,\text{off}} \neq \boldsymbol{0}\big), \end{aligned}$$

$$\begin{aligned} FPG &= \frac{1}{|\mathcal{E}|}\sum_{l\sim l'} \frac{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} = \omega_{jj'l'}, \hat{\omega}_{jj'l} \neq \hat{\omega}_{jj'l'})}{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} = \omega_{jj'l'})} \Big(1 - \mathbb{I}\big(\boldsymbol{\Omega}_{l,\text{off}} \neq \boldsymbol{\Omega}_{l',\text{off}}\big)\Big), \\ FNG &= \frac{1}{|\mathcal{E}|}\sum_{l\sim l'} \frac{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} \neq \omega_{jj'l'}, \hat{\omega}_{jj'l} = \hat{\omega}_{jj'l'})}{\sum_{1\le j<j'\le p} \mathbb{I}(\omega_{jj'l} \neq \omega_{jj'l'})} \mathbb{I}\big(\boldsymbol{\Omega}_{l,\text{off}} \neq \boldsymbol{\Omega}_{l',\text{off}}\big), \end{aligned}$$

where $\boldsymbol{\Omega}_{l,\text{off}}$ denotes the off-diagonal elements of $\boldsymbol{\Omega}_l$. Note that FPV and FNG as well as FNV and FNG are not comparable due to normalization with and without the zero-group, respectively.

For tuning, we minimize a prediction criterion with respect to the tuning parameter(s) on an independent test set with the same sample size as the training set. The prediction criterion is $CV(\boldsymbol{\lambda}) = \frac{1}{L}\sum_{l=1}^{L}\Big(-\log\det\big(\hat{\boldsymbol{\Omega}}_l(\boldsymbol{\lambda})\big)+tr\big(\boldsymbol{S}_l^{\text{tune}}\hat{\boldsymbol{\Omega}}_l(\boldsymbol{\lambda})\big)\Big)$, where $\boldsymbol{S}_l^{\text{tune}}$ is the sample covariance matrix for the tuning data; $l = 1, \cdots, L$. Then the estimated tuning parameter is obtained: $\boldsymbol{\lambda}^{\star} = \text{argmin}_{\boldsymbol{\lambda}}\, CV(\boldsymbol{\lambda})$, which is used in the estimated precision matrices. Here minimization of $CV(\boldsymbol{\lambda})$ is performed through a simple grid search over the domain of the tuning parameter(s).

All simulations are performed based on 100 simulation replications. Three different types of networks are considered. Specifically, Example 1 concerns a chain network with small p and L but large n, where each $\boldsymbol{\Omega_l}$ is relatively sparse and a temporal change occurs at two different l values. Example 2 deals with a nearest neighbor networks for each Ω_l and the same temporal structure as in Example 2. Examples 3 and 4 study exponentially decaying networks in nonsparse precision matrices in high and low-dimensional situations with large and small L, respectively. In Examples 1-3 and Example 4, **Algorithms** 1 and 2 are respectively applied.

Example 1: Chain networks: This example estimates tridiagonal precision matrices as in [45]. Specifically, $\boldsymbol{\Omega}_l^{-1} = \boldsymbol{\Sigma}_l$ is AR(1)-structured with its ij-element being $\sigma_{ijl} = \exp(-|s_{il} - s_{jl}|/2)$, and $s_{1l} < s_{2l} < \cdots < s_{pl}$ are randomly chosen: $s_{il} - s_{(i-1)l} \sim$ Unif(0.5, 1); $i = 2, \cdots, p$, $l = 1, \cdots, L$. The following situations are considered: (I) $(n, p, L) = (120, 30, 4)$, $(n, p, L) = (120, 200, 4)$, with $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$, $\boldsymbol{\Omega}_3 = \boldsymbol{\Omega}_4$; (II) $(n, p, L) = (120, 20, 30)$, $(n, p, L) = (120, 10, 90)$, with $\boldsymbol{\Omega}_1 = \cdots = \boldsymbol{\Omega}_{L/3}$, $\boldsymbol{\Omega}^{(1+L/3)} = \cdots = \boldsymbol{\Omega}_{2L/3}$, $\boldsymbol{\Omega}_{1+2L/3} = \cdots = \boldsymbol{\Omega}_L$. Then, we study the proposed method's performance as a function of the number of graphs and the number of nodes.

Example 2: Nearest neighbor networks. This example concerns networks described in [28]. In particular, we generate p points randomly on a unit square, and compute the k nearest neighbors of each point based on the Euclidean distance. In the case of $k = 3$, three points are connected to each point. For each "edge" in the graph, the corresponding off-diagonal in a precision matrix is sampled independently according to the uniform distribution over $[-1, -0.5] \cup [0.5, 1]$, and the ith diagonal is set to be the sum of the absolute values of the ith row off-diagonals. Given the previous cluster, the matrices in the current cluster are obtained by randomly adding or deleting a small fraction of nonzero elements in the matrices from previous cluster. Finally, each row of a

precision matrix is divided by the square root of the product of corresponding diagonals ($\omega_{ij} \leftarrow \frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$) so that diagonals of the final precision matrices are one. The following scenarios are considered: (I) $(n, p, L) = (300, 30, 4)$ and $(n, p, L) = (300, 200, 4)$, where $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, $\mathbf{\Omega}_3 = \mathbf{\Omega}_4$, and (II) $(n, p, L) = (300, 20, 30)$ and $(n, p, L) = (300, 10, 90)$, where $\mathbf{\Omega}_1 = \cdots = \mathbf{\Omega}_{L/3}, \mathbf{\Omega}_{1+L/3} = \cdots = \mathbf{\Omega}_{2L/3}, \mathbf{\Omega}_{1+2L/3} = \cdots = \mathbf{\Omega}_l$. In (I), the first cluster of matrices $(\mathbf{\Omega}_1, \mathbf{\Omega}_2)$ are generated using the above mechanism, with the second cluster of matrices $(\mathbf{\Omega}_3, \mathbf{\Omega}_4)$ obtained by deleting one edge for each node in the network. In (II), the generating mechanism remains except that the third cluster of matrices $(\mathbf{\Omega}_{1+2L/3}, \cdots, \mathbf{\Omega}_L)$ are generated by adding an edge for each node in its previous adjacent network.

Example 3: Exponentially decaying networks. This example examines a nonsparse situation in which elements of precision matrices are nonzero, and decay exponentially with respect to their Euclidean distances to the corresponding diagonals. In particular, the (i, j)th entry of the lth precision matrix ω_{ijl} is $\exp\big(a_l|i - j|\big)$ with a_l sampled uniformly over $[1, 2]$. In this case, it is sensible to report the results for parameter estimation as opposed to identifying nonzeros As in Examples 1 and 2, several scenarios are considered: (I) $(p, L) = (30, 4)$, $(p, L) = (200, 4)$, and the sample size $n = 120$ or 300 with $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, $\mathbf{\Omega}_3 = \mathbf{\Omega}_4$, and (II) $(p, L) = (20, 30)$, $(p, L) = (10, 90)$, and the sample size $n = 120$ or 300 with $\mathbf{\Omega}_1 = \cdots = \mathbf{\Omega}_{L/3}, \mathbf{\Omega}_{1+L/3} = \cdots = \mathbf{\Omega}_{2L/3}, \mathbf{\Omega}_{1+2L/3} = \cdots = \mathbf{\Omega}_L$.

Example 4: Large precision matrices. This example utilizes the partition rule to treat large-scale simulations. First, we examine two cases $(n, p, L) = (120, 1000, 4)$ and $(n, p, L) = (500, 2000, 4)$ with $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$ and $\mathbf{\Omega}_3 = \mathbf{\Omega}_4$, where four precision matrices are considered with size 1000×1000 and 2000×2000 for pairwise clustering, where $\mathcal{U}$ is the complete graph. Here each precision matrix is set to be a block-diagonal matrix: $\mathbf{\Omega}_l = \text{Diag}(\mathbf{\Omega}_{l1}, \cdots, \mathbf{\Omega}_{lq})$; $l = 1, \cdots, L$, where $\mathbf{\Omega}_{1j} = \mathbf{\Omega}_{2j}, \mathbf{\Omega}_{3j} = \mathbf{\Omega}_{4j}$ are 20×20 matrices generated in the same fashion as that in **Examples 1**. Finally, the complete graph is used as opposed to the fused graph. Overall, the complexity is much higher than the previous examples.

As suggested by Tables 2.1-2.4, the proposed method performs well against its competitors in parameter estimation and correct identification of the sparseness and grouping structures across all the situations. With regard to accuracy of identification of

Table 2.1: Average entropy loss, denoted by EL, (SD in parentheses) and average quadratic loss, denoted by QL, (SD in parentheses), average false positive for sparseness pursuit, denote by FPV, (SD in parentheses), average false negative for sparseness pursuit, denoted as FNV, (SD in parentheses), average false positive for grouping, denoted by FPG, (SD in parentheses), and average false negative for grouping, denoted by FNG, (SD in parentheses), based on 100 simulations, for estimating precision matrices in Example 1 with $n = 120$. Here "Smooth", "Lasso", "TLP", "Our-con" and "Ours" denote estimation of individual matrices with kernel smoothing method proposed in [5], the L_1 sparseness penalty, that with non-convex TLP penalty, the convex counterpart of our method with the L_1-penalty for sparseness and clustering, and our non-convex estimates by solving (2.2) with penalty (2.6). The best performer is bold-faced.

(p, L)	Method	EL	QL	FPV	FNV	FPG	FNG
$(30, 4)$	Smooth	0.570(.005)	2.617(.266)	.312(.016)	.000(.000)	.392(.020)	.000(.000)
	Lasso	1.547(.074)	5.416(.393)	.200(.009)	.000(.000)	.377(.015)	.000(.000)
	TLP	0.746(.084)	4.688(.617)	.043(.006)	.001(.002)	.108(.008)	.000(.000)
	Our-con	1.288(.064)	3.700(.270)	.129(.016)	.000(.000)	.045(.020)	.251(.032)
	Ours	**0.525(.055)**	**3.494(.418)**	.040(.009)	.000(.000)	.009(.007)	.267(.043)
$(200, 4)$	Smooth	7.118(.173)	**22.45(2.61)**	.087(.017)	.000(.000)	.106(.018)	.000(.000)
	Lasso	36.48(.426)	69.21(1.97)	.013(.001)	.000(.000)	.027(.001)	.000(.000)
	TLP	5.305(.351)	33.67(2.22)	.004(.000)	.005(.003)	.014(.000)	.000(.000)
	Our-con	36.28(.422)	66.53(1.87)	.010(.001)	.000(.000)	.012(.001)	.122(.021)
	Ours	**3.500(.164)**	23.71(1.33)	.003(.000)	.000(.000)	.001(.000)	.280(.007)
$(20, 30)$	Smooth	1.122(.023)	1.983(.056)	.131(.006)	.000(.000)	.223(.006)	.000(.000)
	Lasso	1.685(.028)	3.770(.113)	.152(.005)	.000(.000)	.314(.008)	.000(.000)
	TLP	0.507(.023)	3.081(.180)	.077(.004)	.000(.000)	.198(.005)	.000(.000)
	Ours-con	1.593(.028)	3.256(.097)	.136(.007)	.000(.000)	.055(.003)	.024(.004)
	Ours	**0.236(.015)**	**1.812(.130)**	.068(.016)	.000(.000)	.020(.002)	.032(.004)
$(10, 90)$	Smooth	0.339(.007)	0.603(.017)	.271(.014)	.000(.000)	.420(.012)	.000(.000)
	Lasso	0.575(.010)	1.439(.038)	.273(.007)	.000(.000)	.541(.009)	.000(.000)
	TLP	0.250(.009)	1.404(.061)	.196(.006)	.000(.000)	.434(.008)	.000(.000)
	Our-con	0.519(.009)	1.190(.032)	.284(.018)	.000(.000)	.071(.005)	.008(.002)
	Ours	**0.100(.005)**	**0.748(.043)**	.017(.020)	.000(.000)	.028(.004)	.012(.002)

Table 2.2: Average entropy loss, denoted by EL, (SD in parentheses) and average quadratic loss, denoted by QL, (SD in parentheses), average false positive for sparseness pursuit, denote by FPV, (SD in parentheses), average false negative for sparseness pursuit, denoted as FNV, (SD in parentheses), average false positive for grouping, denoted by FPG, (SD in parentheses), and average false negative for grouping, denoted by FNG, (SD in parentheses), based on 100 simulations, for estimating precision matrices in Example 2 with $n = 300$. Here "Smooth", "Lasso", "TLP", "Our-con" and "Ours" denote estimation of individual matrices with kernel smoothing method proposed in [5], the L_1 sparseness penalty, that with non-convex TLP penalty, the convex counterpart of our method with the L_1-penalty for sparseness and clustering, and our non-convex estimates by solving (2.2) with penalty (2.6). The best performer is bold-faced.

(p, L)	Method	EL	QL	FPV	FNV	FPG	FNG
$(30, 4)$	Smooth	0.418(.025)	1.081(.072)	.387(.054)	.009(.008)	.543(.060)	.002(.002)
	Lasso	0.732(.041)	1.840(.122)	.229(.011)	.061(.013)	.466(.016)	.006(.005)
	TLP	0.772(.055)	2.290(.192)	.038(.005)	.184(.020)	.162(.010)	.037(.014)
	Our-con	0.591(.039)	1.373(.104)	.081(.019)	.033(.013)	.056(.035)	.146(.021)
	Ours	**0.359(.036)**	**1.012(.111)**	.044(.009)	.031(.015)	.004(.002)	.233(.012)
$(200, 4)$	Smooth	3.198(.069)	**7.823(.187)**	.129(.010)	.030(.006)	.161(.010)	.013(.002)
	Lasso	6.902(.140)	17.91(.435)	.049(.001)	.234(.010)	.110(.002)	.039(.004)
	TLP	8.350(.215)	23.71(.710)	.003(.001)	.493(.015)	.017(.001)	.116(.010)
	Our-con	6.151(.148)	15.58(.439)	.014(.005)	.198(.013)	.030(.010)	.142(.046)
	Ours	**2.977(.135)**	8.108(.399)	.001(.000)	.191(.010)	.001(.000)	.284(.012)
(20,30)	Smooth	0.409(.011)	0.839(.006)	.063(.007)	.024(.004)	.290(.007)	.005(.001)
	Lasso	0.470(.011)	1.157(.034)	.290(.006)	.034(.004)	.611(.008)	.001(.001)
	TLP	0.491(.017)	1.421(.057)	.081(.004)	.118(.007)	.341(.006)	.004(.001)
	Our-con	0.303(.010)	0.682(.025)	.071(.013)	.008(.002)	.044(.014)	.023(.002)
	Ours	**0.111(.006)**	**0.317(.019)**	.012(.005)	.006(.003)	.011(.002)	.032(.003)
$(10, 90)$	Smooth	0.123(.003)	0.248(.007)	.132(.013)	.008(.002)	.518(.008)	.001(.000)
	Lasso	0.170(.003)	0.419(.010)	.405(.010)	.007(.002)	.798(.008)	.000(.000)
	TLP	0.155(.005)	0.439(.014)	.175(.007)	.020(.003)	.609(.007)	.000(.000)
	Our-con	0.099(.008)	0.230(.007)	.135(.024)	.000(.000)	.043(.015)	.001(.001)
	Ours	**0.040(.002)**	**0.117(.005)**	.015(.014)	.000(.000)	.009(.001)	.001(.001)

Table 2.3: Average entropy loss, denoted by EL, (SD in parentheses) and average quadratic loss, denoted by QL, (SD in parentheses), based on 100 simulations, for estimating multiple precision matrices in Example 3. Here "Smooth", "Lasso", "TLP", "Our-con" and "Ours" denote estimation of individual matrices with kernel smoothing method proposed in [5], the L_1 sparseness penalty, that with non-convex TLP penalty, the convex counterpart of our method with the L_1-penalty for sparseness and clustering, and our non-convex estimates by solving (2.2) with penalty (2.6). The best performer is bold-faced.

Set-up	Method	$n = 120$		$n = 300$	
(p,L)		EL	QL	EL	QL
(30 , 4)	Smooth	**0.468(.042)**	**0.941(.097)**	**0.231(.056)**	**0.476(.034)**
	Lasso	1.158(.062)	2.534(.175)	0.736(.036)	1.434(.086)
	TLP	1.546(.100)	3.625(.262)	0.575(.045)	1.301(.121)
	Our-con	0.897(.066)	1.823(.166)	0.699(.038)	1.317(.085)
	Ours	0.501(.063)	1.143(.160)	0.247(.017)	0.524(.042)
(200, 4)	Smooth	6.882(.220)	13.26(.535)	2.578(.066)	4.843(.130)
	Lasso	10.37(.173)	21.92(.498)	5.449(.094)	10.84(.211)
	TLP	12.34(.202)	25.87(.560)	5.523(.153)	12.08(.365)
	Our-con	6.091(.199)	12.34(.484)	4.625(.098)	8.625(.209)
	Ours	**5.079(.265)**	**11.84(.658)**	**1.682(.038)**	**3.551(.096)**
(20 , 30)	Smooth	0.490(.021)	0.878(.042)	0.278(.008)	0.492(.015)
	Lasso	0.786(.020)	1.670(.052)	0.564(.012)	1.066(.027)
	TLP	0.987(.036)	2.454(.107)	0.355(.014)	0.819(.035)
	Our-con	0.653(.023)	1.281(.055)	0.528(.012)	0.959(.027)
	Ours	**0.317(.013)**	**0.730(.036)**	**0.183(.005)**	**0.391(.014)**
(10 , 90)	Smooth	0.230(.008)	0.409(.017)	0.115(.003)	0.203(.005)
	Lasso	0.318(.008)	0.694(.022)	0.205(.004)	0.398(.008)
	TLP	0.402(.012)	1.010(.043)	0.148(.004)	0.346(.011)
	Our-con	0.240(.008)	0.487(.019)	0.180(.004)	0.335(.007)
	Ours	**0.158(.005)**	**0.369(.014)**	**0.082(.002)**	**0.176(.005)**

Table 2.4: Average entropy loss, denoted by EL, (SD in parentheses) and average quadratic loss, denoted by QL, (SD in parentheses), average false positive for sparseness pursuit, denote by FPV, (SD in parentheses), average false negative for sparseness pursuit, denoted as FNV, (SD in parentheses), average false positive for grouping, denoted by FPG, (SD in parentheses), and average false negative for grouping, denoted by FNG, (SD in parentheses), based on 100 simulations, for estimating precision matrices in Example 4. Here "Lasso", "TLP", "Our-con" and "Ours" denote estimation of individual matrices with the L_1 sparseness penalty, that with non-convex TLP penalty, the convex counterpart of our method with the L_1-penalty for sparseness and clustering, and our non-convex estimates by solving (2.2) with penalty (2.6). The best performer is bold-faced.

(p, L)	Method	EL	QL	FPV	FNV	FPG	FNG
$(1000, 4)$	Lasso	378.5(2.09)	829.4(13.3)	.0005(.0000)	.0270(.0030)	.0020(.0000)	.0002(.0003)
	TLP	36.05(.1.78)	201.9(8.81)	.0004(.0000)	.0270(.0030)	.0020(.0000)	.0002(.0003)
	Our-con	377.3(2.07)	805.2(12.9)	.0004(.0000)	.0110(.0020)	.0017(.0000)	.0497(.0040)
	Ours	**26.8(1.35)**	**160.9(7.266)**	.0003(.0000)	.0130(.0020)	.0017(.0000)	.0267(.0027)
$(2000, 4)$	Lasso	225.6(.413)	358.1(1.13)	.0009(.0000)	.0000(.0000)	.0018(.0000)	.0000(.0000)
	TLP	9.160(.083)	54.17(.654)	.0007(.0000)	.0000(.0000)	.0015(.0000)	.0000(.0000)
	Our-con	225.6(.413)	358.1(1.12)	.0009(.0000)	.0000(.0000)	.0018(.0000)	.0000(.0000)
	Ours	**8.617(.081)**	**51.79(.657)**	.0006(.0000)	.0000(.0000)	.0005(.0000)	.1750(.0020)

the sparseness and clustering structures, the proposed method has the smallest false positives in terms of FPV and FPG, yielding sharper parameter estimation than the competitors. This says that shrinkage towards common elements is advantageous for parameter estimation in a low-or high-dimensional situation. Note that the largest improvement occurs for the most difficult situation in Example 4.

Compared with pursuit of sparseness alone–TLP, the amount of of improvement of our method is from 143% to 244% and 118% % to 236% in terms of the EL and QL when $n = 120$, and from 80.5% to 228% and 96.5% to 240% in terms of the EL and QL when $n = 300$, as indicated in Table 2.3. This comparison suggests that exploring the sparseness structure alone is inadequate for multiple graphical models. Pursuit of two types of structures appears advantageous in terms of performance, especially for large matrices.

Compared with its convex counterpart "our-con", our method leads to between a 19.9% and a 106% improvement, and between a 4.2% improvement and a 75.5% improvement in terms of the EL and QL when $n = 120$, and between a 18.3% improvement and a 120% improvement, and a 90.3% improvement and a 151% improvement in terms

of the EL and QL when $n = 300$; see Table 2.3. This is expected because more accurate identification of structures tends to yield better parameter estimation.

In contrast to the smoothing method [32, 5] for time-varying network analysis, across all cases except one low-dimensional case of $L = 4$ and $p = 30$ in Table 2.3, our method yields a 54.5% improvement and a 20.3% improvement in terms of the EL and QL when $n = 120$, and a 51.9% improvement and a 25.8% improvement in terms of the EL and QL when $n = 300$ when L is not too small, c.f. Tables 2.1-2.4.

To understand how the proposed method performs relative to (n, p, L), we examine Table 2.3 and Figure 2.1 in further detail. Overall, the proposed method performs better as n, L increases and worse as p increases. Interestingly, as suggested by Figure 2.1, the method performs better as L increases, which confirms with our theoretical analysis.

In summary, the proposed method achieves the desired objective of pursuing simultaneous both sparseness and clustering structures to battle the curse of dimensionality in a high-dimensional situation.

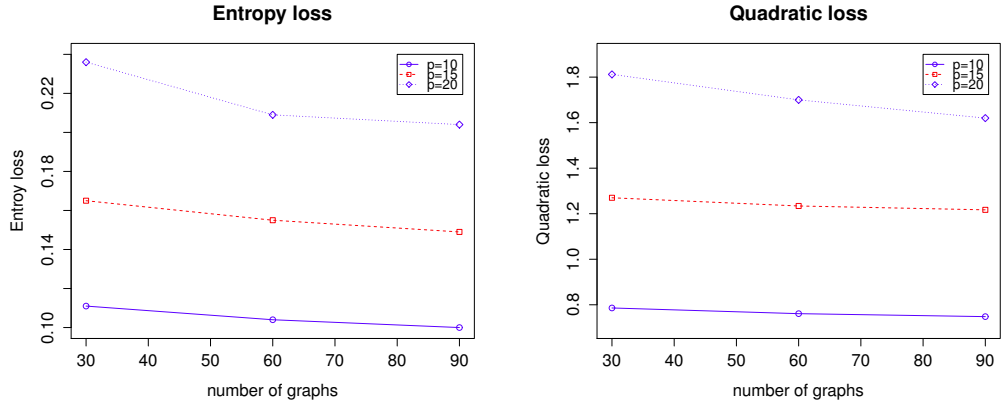


Figure 2.1: Average entropy and quadratic losses of the proposed method over different p and L values over 100 simulation replications in Example 1.

2.6 Real data analysis

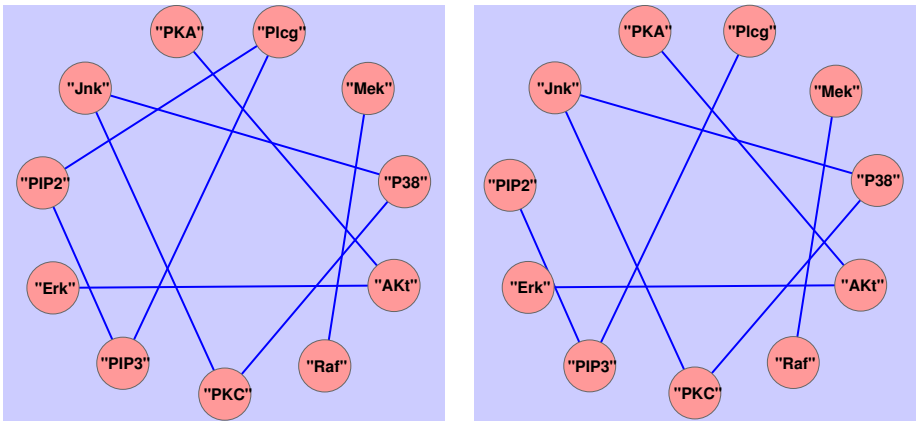
This chapter applies the proposed method to the multivariate single cell flow cytometry data in [6] to infer a signaling network or pathway; a consensus version of the network with eleven proteins is described in Figure 2.3. In this study, a multiparameter flow cytometry recorded the quantitative amounts of the eleven proteins in a single cell as an observation. To infer the network, experimental perturbations on various aspects of the network were imposed before the amounts of the eleven proteins were measured under each condition. The idea was that, if a chemical was applied to stimulate or inhibit the activity of a protein, then both the abundance of the protein and those of its downstream proteins in the network would be expected to increase or decrease, while those of non-related proteins would barely change. There were ten types of experimental perturbations on different targets: 1) activating a target (CD3) in the upstream of the network so that the whole network was expected to be perturbed; 2) activating a target (CD28) in the upsteam of the network; 3) activating a target (ICAM2) in the upsteam of the network; 4) activating PKC; 5) activating PKA; 6) inhibiting PKC; 7) inhibiting Akt; 8) inhibiting PIP2; 9) inhibiting Mek; 10) inhibiting a target (PI3K) in the upstream of the network. In [6], data were collected under nine experimental conditions and then used to infer a directed network; each of the nine experimental conditions was either a single type of perturbation or a combination of two or three types of perturbations. Interestingly, data were also collected under another five conditions, each of which was a combination of two of the previous nine conditions. Hence, the data offered an opportunity to infer the two networks under the two sets of the conditions: since the two sets of conditions largely overlapped, we would expect the two networks to be largely similar to each other; on the other hand, due to the difference between the two sets of the conditions, some deviations between the two networks were also anticipate. There were $n_1 = 7466$ and $n_2 = 4206$ observations under the two sets of the conditions respectively.

We apply the proposed method to the normalized data under the two sets of the conditions respectively. Due to the expected similarities between the two networks, we consider grouping to encourage common structure defined by connecting edges between the two networks. The tuning parameters are estimated by a three-fold cross-validation.

The reconstructed two undirected networks are now displayed in Figure 2.2, with 9 and 8 estimated (undirected) links for the two groups of conditions, being a subset of the 20 (directed) links in the gold standard signaling network as displayed in Figure 2.3, which is a consensus network that has been verified biologically, c.f. [6]. The reconstructed undirected graphs miss some edges as compared to the gold standard network, for instance, the links from protein "PKC" to "Raf" and "Mek". The three edges missed by [6], "PIP3" to "Akt", "Plcg" to "PKC", and "PIP2" to "PKC", are also missed by our method, possibly reflecting lack of information in the data due to no direct interventions imposed on "PIP3" and "Plcg".

Overall, the proposed method appears to work well in that the network inferred from the first set of conditions recovers one more dependence relationships than that from the second set of conditions, which is expected given that the second set of interventional conditions is less specific than the first one.

Here we analyze the data by contrasting the network constructed under the nine conditions with $n_1 = 7466$ against that under the five conditions with $n_2 = 4206$. Of particular interest is the detection of network structural changes between the two sets of conditions.



(a) Under the nine conditions. (b) Under the five conditions.

Figure 2.2: Reconstructed networks for simultaneous pursuit of clustering and sparsity.

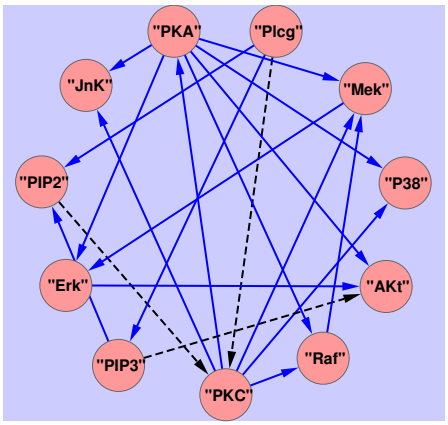


Figure 2.3: Signaling network reproduced from Figure 3(A) of [6], where the black dashed line represents links that have been missed by methods in [6].

2.7 Conclusion and Discussion

This thesis proposes a novel method to pursue two disparate types of structures—sparseness and clustering for multiple Gaussian graphical models. The proposed method is equipped with an efficient algorithm for large graphs, which is integrated with a partition rule to break down a large problem into many separate small problems to solve. For data analysis, we have considered signaling network inference in a low-dimensional situation. Worthy of note is that the proposed method can be equally applied to high-dimensional data, such as reconstructing and comparing gene regulatory networks across four subtypes of glioblastoma multiforme based on gene expression data [36].

To make the proposed method useful in practice, inferential tools need to be further developed. A Monte Carlo method may be considered given the level of complexity of the underlying problems. Moreover, the general approach developed here can be expanded to other types of graphical models, for instance, dynamic network models or time-varying graphical models [32]. This enables us to build time dependency into a model through, for example, a Markov property. Further investigation is necessary.

References

[1] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498, 2010.

[2] Seyoung Kim and Eric P Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8):e1000587, 2009.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[4] Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[5] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2):295–319, 2010.

[6] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523, 2005.

[7] Clemens R Scherzer, Aron C Eklund, Lee J Morse, Zhixiang Liao, Joseph J Locascio, Daniel Fefer, Michael A Schwarzschild, Michael G Schlossmacher, Michael A Hauser, Jeffery M Vance, et al. Molecular markers of early parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, 104(3):955–960, 2007.

[8] 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*, 67:91–108, 2005.

[9] X. Shen and H.C. Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739, 2010.

[10] Alessandro Rinaldo et al. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009.

[11] Wei Pan, Benhuai Xie, and Xiaotong Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484, 2010.

[12] Xiaotong Shen and Jianming Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97(457):210–221, 2002.

[13] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19, 2007.

[14] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

[15] Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832, 2013.

[16] Jean-Baptiste Veyrieras, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard. High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214, 2008.

[17] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107:223–232, 2012.

[18] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[19] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[20] Pham Dinh Tao et al. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

[21] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[22] Hua Zhong, Xia Yang, Lee M Kaplan, Cliona Molony, and Eric E Schadt. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics*, 86(4):581–591, 2010.

[23] Xianghong Zhou, Ming-Chih J Kao, and Wing Hung Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20):12783–12788, 2002.

[24] Wei Zhang, Jun Zhu, Eric E Schadt, and Jun S Liu. A bayesian partition method for detecting pleiotropic and epistatic eqtl modules. *PLoS computational biology*, 6(1):e1000642, 2010.

[25] Leonardo Bottolo, Enrico Petretto, Stefan Blankenberg, François Cambien, Stuart A Cook, Laurence Tiret, and Sylvia Richardson. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189(4):1449–1459, 2011.

[26] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[27] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

[28] H. Li and J. Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302, 2006.

[29] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[30] G.V. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *Arxiv preprint arXiv:0807.3734*, 2008.

[31] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[32] M. Kolar and E.P. Xing. On time varying undirected graphs. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

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

[34] B. Li, H. Chun, and H. Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107:152–167, 2012.

[35] Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, and Satoru Miyano. Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. *Bioinformatics*, 26(8):1064–1072, 2010.

[36] R. Verhaak, K.A Hoadley, E. Purdom, V. Wang, Y. Qi, M. D Wilkerson, C. R. Miller, L. Ding, T. Golub, J.P. Mesirov, et al. An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr and nf1. *Cancer cell*, 17(1):98, 2010.

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

[38] Sen Yang, Zhisong Pan, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *arXiv preprint arXiv:1209.2139*, 2012.

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

[40] Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

[41] Y. Zhu, X. Shen, and W. Pan. Simultaneous grouping pursuit and feature selection over an undirected graph. *Journal of the American Statistical Association*, 108:713–725, 2013.

[42] T.H. Cormen. *Introduction to algorithms*. The MIT press, 2001.

[43] A.N. Kolmogorov and V.M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[44] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

[45] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521–541, 2009.

[46] Gary Chartrand. *Introduction to graph theory*. Tata McGraw-Hill Education, 2006.

[47] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[48] W.H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, pages 339–362, 1995.

Appendix A

Proofs

A.1 Technical details for Chapter 1

Proof of Lemma 1: Before proceeding, we introduce some notations. Let $\widetilde{\boldsymbol{X}}$ be a matrix with column vectors $(\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_p)$, where $\tilde{\boldsymbol{x}}_k = \boldsymbol{x}_k$ if $k \in \big(\cup_{j=1}^{K_0} \mathcal{I}_{j1}^0\big) \cup \mathcal{I}_0^0$; $\tilde{\boldsymbol{x}}_k = -\boldsymbol{x}_k$ otherwise. In other words, $\widetilde{\boldsymbol{X}}$ is generated by flipping signs of columns of $\boldsymbol{X}$ when their indices are in $\cup_{j=1}^{K_0}\mathcal{I}_{j2}^0$. For any partition $\mathcal{G} = (\mathcal{I}_0, \mathcal{I}_1, \cdots, \mathcal{I}_K)$ with $\mathcal{I}_i = \mathcal{I}_{i1} \cup \mathcal{I}_{i2}, i = 1, \cdots, K$, let $S_{\mathcal{G}}(k) = 1$ if $k \in \big(\cup_{i=1}^{K} \mathcal{I}_{i1}\big) \cup \mathcal{I}_0$ and $S_{\mathcal{G}}(k) = -1$ otherwise. For $\mathcal{G} \in \mathcal{S}$, let $A = \mathcal{I} \setminus \mathcal{I}_0$, and $A_0 = \mathcal{I} \setminus \mathcal{I}_0^0$. Denote by $s_k = S_{\mathcal{G}^0}(k) S_{\mathcal{G}}(k)$; $k = 1, \cdots, p$.

To lower bound C_{min}, note that $\tilde{c}_{min} = \min_{|B| \le 2|\mathcal{I} \setminus \mathcal{I}_0^0|, \mathcal{I} \setminus \mathcal{I}_0^0 \subseteq B} \lambda_{min}\Big(n^{-1} \widetilde{\boldsymbol{X}}_B^T \widetilde{\boldsymbol{X}}_B\Big) = c_{min}$, because $\widetilde{\boldsymbol{X}}_B^T \widetilde{\boldsymbol{X}}_B = \boldsymbol{X}_B^T \boldsymbol{X}_B$ for any B by definition. For $\mathcal{G} \in \mathcal{S}$, write $\boldsymbol{X}_{A_0} \boldsymbol{\beta}_{A_0}^0 - \boldsymbol{X}_{\mathcal{G}} \boldsymbol{\alpha}$ as

$$\sum_{i=1}^{K_0} \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_i^0 \cap \mathcal{I}_j} (S_{\mathcal{G}}^0(k) \beta_k^0 - s_k \alpha_j) \tilde{\boldsymbol{x}}_k + \sum_{i=1}^{K_0} \sum_{k \in \mathcal{I}_i^0 \setminus A} S_{\mathcal{G}}^0(k) \beta_k^0 \tilde{\boldsymbol{x}}_k + \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_j \setminus (\mathcal{I} \setminus \mathcal{I}_0^0)} s_k \alpha_j \tilde{\boldsymbol{x}}_k.$$

Then $\|(I - \boldsymbol{P}_{\mathcal{G}}) \boldsymbol{X}_{A_0} \boldsymbol{\beta}_{A_0}^0\|^2 = \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \|\boldsymbol{X}_{A_0} \boldsymbol{\beta}_{A_0}^0 - \boldsymbol{X}_{\mathcal{G}} \boldsymbol{\alpha}\|^2$ is lower bounded by

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \Big(\sum_{i=1}^{K_0} \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_i^0 \cap \mathcal{I}_j} (S_{\mathcal{G}^0}(k) \beta_k^0 - s_k \alpha_j)^2 + \sum_{i=1}^{K_0} \sum_{k \in \mathcal{I}_i^0 \setminus A} (\beta_k^0)^2 + \sum_{j=1}^{K} |\mathcal{I}_j \setminus A_0| \alpha_j^2 \Big) c_{min} n \equiv I.$$

If $\mathcal{I}_i^0 \setminus A \neq \emptyset$ for some i; $1 \leq i \leq K_0$, then $I \geq nc_{min} \sum_{k \in \mathcal{I}_i^0 \setminus A} (\beta_k^0)^2 \geq nc_{min}\eta^2$. Otherwise, $\mathcal{I}_i^0 \setminus A = \emptyset$; $i = 1, \cdots, K_0$, implying that $A_0 \subseteq A$. Note further that $|A| \leq |A_0|$ for $\mathcal{G} \in \mathcal{S}$ by assumption. Then $A_0 = A$. Hence

$$I = \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \Big(\sum_{i=1}^{K_0} \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_i^0 \cap \mathcal{I}_j} (S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2 \Big) c_{\min} n.$$

Next two cases are examined.

For each j; $1 \leq j \leq K$, (a) if there exist two indices i', i'' with $1 \leq i' \neq i'' \leq K_0$ such that $\mathcal{I}_{i'}^0 \cap \mathcal{I}_j \neq \emptyset$ and $\mathcal{I}_{i''}^0 \cap \mathcal{I}_j \neq \emptyset$, then

$$\begin{aligned} I &\geq nc_{min} \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \Big(\sum_{k \in \mathcal{I}_{i'}^0 \cap \mathcal{I}_j} (S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2 + \sum_{k \in \mathcal{I}_{i''}^0 \cap \mathcal{I}_j} (S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2 \Big) \\ &\geq nc_{min} \min_{(j,j'):|\beta_j^0| \neq |\beta_{j'}^0|} \frac{1}{2}(|\beta_j^0| - |\beta_{j'}^0|)^2 \geq nc_{min}\eta^2; \end{aligned}$$

otherwise, (b) there exists at most one index i^* with $1 \leq i^* \leq K_0$ such that $\mathcal{I}_j \subseteq \mathcal{I}_{i^*}^0$, or $\mathcal{G}^0$ is coarser than $\mathcal{G}$. This implies that $C_2(\mathcal{G}, \mathcal{E}) \geq C_2(\mathcal{G}^0, \mathcal{E}) = c_0$, which in turn yields that $C_2(\mathcal{G}, \mathcal{E}) > c_0$ when $\mathcal{G} \neq \mathcal{G}^0$ by graph consistency. This contradicts to the tuning assumption that $C_2(\mathcal{G}, \mathcal{E}) \leq c_0$. The bound of I in (a) thus establishes (1.13).

For (1.14), two cases are considered for any $\mathcal{G} \in \mathcal{S}$: (c) if there exists an index subset of length l^* $\{i_1, \cdots, i_{l^*}\} \subseteq \{1, \cdots, K_0\}$ and that of length $(l^* - 1)$ $\{j_1, \cdots, j_{l^*-1}\} \subseteq \{1, \cdots, K\}$ such that $\mathcal{I}_{i_1}^0 \cup \cdots \cup \mathcal{I}_{i_{l^*}}^0 \subseteq \mathcal{I}_{j_1} \cup \cdots \cup \mathcal{I}_{j_{l^*-1}}$ for some l^* with $1 \leq l^* \leq K$; otherwise, (d) for any l with $1 \leq l \leq K$, $\{i_1, \cdots, i_l\}$, $(\mathcal{I}_{i_1}^0 \cup \cdots \cup \mathcal{I}_{i_l}^0) \nsubseteq (\mathcal{I}_{j_1} \cup \cdots \cup \mathcal{I}_{j_k}) \neq \emptyset$ for $k < l$.

For (c), let $\mathcal{J} = (A \cup A_0) \setminus (\mathcal{I}_{i_1}^0 \cup \cdots \cup \mathcal{I}_{i_{l^*}}^0)$, $L(\boldsymbol{X}_\mathcal{J}) = \boldsymbol{X}_\mathcal{J}\boldsymbol{\beta}_\mathcal{J}^0 - \sum_{k \in \mathcal{J}} \big(\sum_{j=1}^K \alpha_j \mathbb{I}(k \in \mathcal{I}_j) \big) \boldsymbol{x}_k$, $\boldsymbol{\alpha} = (\alpha_{j_1}, \cdots, \alpha_{j_{l^*-1}}) \in \mathbb{R}^{l^*-1}$ and $a_{ts}^{(m)} = \sum_{k \in A_{i_t m}} \pm \mathbb{I}(k \in \mathcal{I}_{j_s})$; $t = 1, \cdots, l^*$, $s = 1, \cdots, l^* - 1$, $m = 1, \cdots, n_t$. For any $\mathcal{G} \in \mathcal{S}$, $\|(I - \boldsymbol{P}_\mathcal{G})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}^0\|^2$ is lower bounded

by

$$\min_{\boldsymbol{\alpha}} \Big\| \sum_{t=1}^{l^*} \beta_{i_t}^0 \sum_{k\in\mathcal{I}_{i_t}^0} \tilde{\boldsymbol{x}}_k - \sum_{k\in\mathcal{I}_{i_1}^0\cup\cdots\cup\mathcal{I}_{i_{l^*}}^0} \tilde{\boldsymbol{x}}_k \sum_{s=1}^{l^*-1} (\pm\alpha_{j_s}) \mathbb{I}(k\in\mathcal{I}_{j_s}) + L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2$$

$$\geq \min_{\boldsymbol{\alpha}} \Big\| \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} |A_{i_t m}| \beta_{i_t}^0 \boldsymbol{z}_{i_t m} - \sum_{t=1}^{l} \sum_{m=1}^{n_t} \Big(\boldsymbol{z}_{i_t m} \big(\sum_{s=1}^{l^*-1} \alpha_{j_s} \sum_{k\in A_{i_t m}} \pm\mathbb{I}(k\in\mathcal{I}_{j_s}) \big) \Big)$$

$$+ L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2$$

$$\geq \min_{\boldsymbol{\alpha}, a_{ts}^{(m)}} \Big\| \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} (|A_{i_t m}| \beta_{i_t}^0 - \sum_{s=1}^{l^*-1} \alpha_{j_s} a_{ts}^{(m)}) \boldsymbol{z}_{i_t m} + L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2$$

$$\geq n c_{min}^G \min_{\boldsymbol{\alpha}, a_{ts}^{(m)}} \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} (|A_{i_t m}| \beta_{i_t}^0 - \sum_{s=1}^{l^*-1} \alpha_{j_s} a_{ts}^{(m)})^2 \geq n c_{min}^G \min_{\boldsymbol{\alpha}, \boldsymbol{A}} \|\boldsymbol{\gamma} - \boldsymbol{A}\boldsymbol{\alpha}\|^2,$$

implying (1.14).

For (d), we will show that it does not occur under sufficient preciseness. Suppose that (d) does. By Hall's Theorem [46], there exists a matching of $\{\mathcal{I}_1^0 \cup \cdots \cup \mathcal{I}_{K_0}^0\}$ into $\{\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_K\}$. Without loss of generality, we may assume $\mathcal{I}_1 \cap \mathcal{I}_1^0 \neq \emptyset, \cdots, \mathcal{I}_{K_0} \cap \mathcal{I}_{K_0}^0 \neq \emptyset$. For $D \subseteq \mathcal{I} = \{1, \cdots, p\}$, let $d_{\mathcal{E}}(D) = \sum_{i,i'\in D; i<i'} \mathbb{I}\big((i,i')\in\mathcal{E}\big)$, and $\mathcal{I}_{ij} = \mathcal{I}_i^0 \cap \mathcal{I}_j$. Then

$$2\big(C_2(\mathcal{G},\mathcal{E}) - C_2(\mathcal{G}^0,\mathcal{E})\big) = 2\big(d_{\mathcal{E}}(\mathcal{I}) - \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_j)\big) - 2\big(d_{\mathcal{E}}(\mathcal{I}) - \sum_{i=0}^{K_0} d_{\mathcal{E}}(\mathcal{I}_i^0)\big)$$

$$= \Big(\sum_{i=0}^{K_0} \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) \Big) - \Big(\sum_{j=0}^{K} \sum_{i=0}^{K_0} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij}) \Big). \tag{A.1}$$

To simplify (A.1), consider two cases: (e) if $\mathcal{I}_i^0 \not\subseteq \mathcal{I}_i$ thus $\mathcal{I}_i^0 \setminus \mathcal{I}_{ii} \neq \emptyset$ for any i; $0 \leq i \leq K_0$; otherwise (f) the set $\mathcal{I}_* \equiv \{i : \mathcal{I}_i^0 \subseteq \mathcal{I}_i\}$ is nonempty.

For (e), note that $\mathcal{I}_{ii} \neq \emptyset$, hence that $\mathcal{I}_i^0 \setminus \mathcal{I}_{ij} \neq \emptyset$ for any $i \neq j$; $0 \leq i \leq K_0, 0 \leq j \leq K$. By sufficiently preciseness, $d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) > 2 d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij}) > d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij})$; $i = 0, \cdots, K_0, j = 0, \cdots, K$, implying that $C_2(\mathcal{G},\mathcal{E}) > C_2(\mathcal{G}^0,\mathcal{E}) = c_0$ in (A.1), which contradicts to the tuning assumption that $C_2(\mathcal{G},\mathcal{E}) \leq c_0$.

For (f), let $\mathcal{I}_*^1 = \{0, 1, \cdots, K_0\} \setminus \mathcal{I}_*$ and $\mathcal{I}_*^2 = \{0, 1, \cdots, K\} \setminus \mathcal{I}_*$. Now, $\mathcal{I}_i^0 \subseteq \mathcal{I}_i, i \in \mathcal{I}_*$. Since $|\cup_{i=1}^{K_0} \mathcal{I}_i^0| \geq |\cup_{j=1}^{K} \mathcal{I}_j|$, $1 \leq |\mathcal{I}_*| < K_0$. Hence $\mathcal{I}_{ij} = \emptyset, i \in \mathcal{I}_*, j \neq i$ and

$\mathcal{I}_i^0 \setminus \mathcal{I}_{ii} = \emptyset, i \in \mathcal{I}_*$. Now (A.1) becomes

$$\sum_{i\in\mathcal{I}_*^1}\sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) - \sum_{j\in\mathcal{I}_*} d_{\mathcal{E}}(\mathcal{I}_j \setminus \mathcal{I}_j^0, \mathcal{I}_j^0) - \sum_{i\in\mathcal{I}_*^1}\sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij}). \quad \text{(A.2)}$$

By sufficiently preciseness, $\sum_{i\in\mathcal{I}_*^1}\sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) > 2\sum_{i\in\mathcal{I}_*^1}\sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij})$. This together with

$$\sum_{j\in\mathcal{I}_*} d_{\mathcal{E}}(\mathcal{I}_j \setminus \mathcal{I}_j^0, \mathcal{I}_j^0) \le \sum_{j\in\mathcal{I}_*}\sum_{i\in\mathcal{I}_*^1} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j^0) \le \sum_{i\in\mathcal{I}_*^1}\sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij})$$

yields that $C_2(\mathcal{G}, \mathcal{E}) > C_2(\mathcal{G}^0, \mathcal{E}) = c_0$ in (A.2), which is impossible as before. Consequently (f) does not occur under sufficiently preciseness. This completes the proof.

Proof of Theorem 1: The proof is similar to the convergence proof in [15]. Hence it will be omitted.

Proof of Theorem 3: Before proceeding, we introduce some notations. Define $\widehat{\mathcal{G}} = \big(\hat{\mathcal{I}}_0, \hat{\mathcal{I}}_1, \cdots, \hat{\mathcal{I}}_K\big)$ with $\hat{\mathcal{I}}_i = \hat{\mathcal{I}}_{i1} \cup \hat{\mathcal{I}}_{i2}$; $i = 1, \cdots, K$ as follows. First, $|\hat{\beta}_j^g|$'s are ordered by their values. Second, check any two consecutive ordered values of $|\hat{\beta}_j^g|$, and set j_1 and j_2 to be in one group if $||\hat{\beta}_{j_1}^g| - |\hat{\beta}_{j_2}^g|| \le \tau$. Third, let $\hat{\mathcal{I}}_0$ be the group whose range contains zero, and $\hat{\mathcal{I}}_0 = \emptyset$ otherwise. Finally, for each $1 \le i \le K$, partition $\hat{\mathcal{I}}_i$ into $\hat{\mathcal{I}}_{i1}$ and $\hat{\mathcal{I}}_{i2}$ by grouping components $\hat{\beta}_j$'s of the same sign together. Consequently, (i) $\max_{j\in\hat{\mathcal{I}}_0} |\hat{\boldsymbol{\beta}}_j^g| \le \tau$; (ii) $\big||\hat{\beta}_{j_1}^g| - |\hat{\beta}_{j_2}^g|\big| \le \tau$ for any $1 \le j_1, j_2 \le K$; (iii) $\hat{\beta}_{j_1}^g \hat{\beta}_{j_2}^g < 0$ for any $j_1 \in \mathcal{I}_{i1}^1, j_2 \in \hat{\mathcal{I}}_{i2}$; $i = 1, \cdots, K$.

Next we show that $\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}$ when $\widehat{\mathcal{G}} = \mathcal{G}^0$. Now $p_1 = \mathcal{I} \setminus \hat{\mathcal{I}}_0^0 = p_0$. By (1.5), $\frac{1}{\tau}\sum_{j\in\hat{\mathcal{I}}_0} |\hat{\beta}_j^g| + p_1 \le p_0$, with $p_0 = p_1$, yields that $\hat{\beta}_j^g = 0$; $j \in \mathcal{I}_0^1$. In addition, the second constraint of (1.5) implies $\sum_{i=1}^{K}\sum_{j,j'\in\mathcal{I}_i,(j,j')\in\mathcal{E}} \frac{\big||\hat{\beta}_j^g| - |\hat{\beta}_{j'}^g|\big|}{\tau} \le 0$, yielding that $\hat{\beta}_j^g = -\hat{\beta}_{j'}^g$; $j \in \hat{\mathcal{I}}_{i1}$, $j' \in \hat{\mathcal{I}}_{i2}, (j, j') \in \mathcal{E}$ and $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$; $j_1, j_2 \in \hat{\mathcal{I}}_{i1}$ or $j_1, j_2 \in \hat{\mathcal{I}}_{i2}, (j_1, j_2) \in \mathcal{E}$. By graph consistency of $\mathcal{E}$, $\mathcal{E}|_{\hat{\mathcal{I}}_i}$ is connected, implying that $\hat{\beta}_j^g = -\hat{\beta}_{j'}^g$; $j \in \hat{\mathcal{I}}_{i1}$, $j' \in \hat{\mathcal{I}}_{i2}$ and $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$; $j_1, j_2 \in \hat{\mathcal{I}}_{i1} \cup \hat{\mathcal{I}}_{i2}$. This further implies that $\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}$, hence that $\{\widehat{\mathcal{G}} = \mathcal{G}^0\} \subseteq \{\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}\}$. Thus

$$\mathbb{P}(\hat{\boldsymbol{\beta}}^g \ne \hat{\boldsymbol{\beta}}^{ol}, \widehat{\mathcal{G}} \ne \mathcal{G}^0) \le \mathbb{P}\Big(S(\hat{\boldsymbol{\beta}}^g) - S(\hat{\boldsymbol{\beta}}^{ol}) \le 0, \widehat{\mathcal{G}} \ne \mathcal{G}^0\Big) \equiv I, \quad \text{(A.3)}$$

To bound I, we first obtain lower bounds of $S(\hat{\boldsymbol{\beta}}^g) - S(\hat{\boldsymbol{\beta}}^{ol})$. Let $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \cdots, \bar{\beta}_p)$, with $\bar{\beta}_j = sign(\hat{\beta}_j^g)\frac{\sum_{j'\in\hat{\mathcal{I}}_i} |\hat{\beta}_{j'}^g|}{|\hat{\mathcal{I}}_i|}$; $j \in \hat{\mathcal{I}}_i$, $i = 1, \cdots, K$ and $\bar{\beta}_j = 0$; $j \in \hat{\mathcal{I}}_0$. Then $|\bar{\beta}_j - \hat{\beta}_j^g| \le$

$(|\hat{\mathcal{I}}_i|-1)\tau$ for $j \in \hat{\mathcal{I}}_i$; $i = 0, \cdots, K$. Note that

$$\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}\|^2 \geq \|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{Y}\|^2 = \|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + (\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}^g})\boldsymbol{\epsilon}\|^2,$$
$$\|\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\|^2 \leq \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)\|\bar{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^g\|^2 \leq \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)\tau^2\sum_{i=0}^{K}(|\hat{\mathcal{I}}_i|-1)^2|\hat{\mathcal{I}}_i|$$
$$\leq \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2.$$

Using the inequality $\|U+V\|^2 \geq \frac{a-1}{a}\|U\|^2-(a-1)\|V\|^2$ for any real vectors $U, V \in \mathbb{R}^p$ and $a > 0$, we have

$$\begin{aligned}
S(\hat{\boldsymbol{\beta}}^g) &= \frac{1}{2}\big\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}+\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\big\|^2 \\
&\geq \frac{a-1}{2a}\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}\|^2 - \frac{a-1}{2}\|\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\| \\
&\geq \frac{a-1}{2a}\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + (\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{\epsilon}\|^2 - \frac{(a-1)\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2}{2} \\
&\geq \frac{a-1}{2a}\Big(\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 + \|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{\epsilon}\|^2 + 2\boldsymbol{\epsilon}^T(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} - \frac{\lambda}{a-1}\Big),
\end{aligned}$$

where $\lambda = a(a-1)\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2$. This yields that

$$\begin{aligned}
&2a\Big(S(\hat{\boldsymbol{\beta}}^g) - S(\hat{\boldsymbol{\beta}}^{ol})\Big) = 2a\Big(S(\hat{\boldsymbol{\beta}}^g) - \frac{1}{2}\|(I-\boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\|^2\Big) \\
&\geq 2(a-1)\boldsymbol{\epsilon}^T(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + (a-1)\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 - \\
&\boldsymbol{\epsilon}^T\Big(\boldsymbol{I}+(a-1)\boldsymbol{P}_{\widehat{\mathcal{G}}}\Big)\boldsymbol{\epsilon} - \lambda \equiv -L(\widehat{\mathcal{G}}) + b(\widehat{\mathcal{G}}),
\end{aligned}$$

where $L(\mathcal{G}) \equiv \big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)^T\Big(\boldsymbol{I}+(a-1)\boldsymbol{P}_{\mathcal{G}}\Big)\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)$, $b(\mathcal{G}) = a(a-1)\|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2-\lambda$. Note that $L(\mathcal{G}) = L_1(\mathcal{G})+L_2(\mathcal{G})$, where $L_1(\mathcal{G}) = \big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)^T\Big(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}}\Big)\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)$, which follows $\chi^2_{k,\Lambda}$ of freedom $n-K$ and non-central parameter $\Lambda = (a-1)^2\sigma^{-2}\|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \geq (a-1)^2nC_{min}/\sigma^2$ and $L_2(\mathcal{G}) = a\boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$ is independent of $L_1(\mathcal{G})$.

Recall that $\mathcal{S} = \big\{\mathcal{G} \neq \mathcal{G}^0 : C_1(\mathcal{G}) \leq p_0; C_2(\mathcal{G},\mathcal{E}) \leq c_0\big\}$. Let $\widehat{A} = \mathcal{I}\setminus\widehat{\mathcal{I}}_0$. By Markov's

inequality with any $t < \frac{1}{2a}$, it follows from (A.3) that

$$\begin{aligned} I &\leq \sum_{A:|A_0\setminus A|=i} \sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{P}\Big(L(\mathcal{G}) \geq b(\mathcal{G}), \widehat{\mathcal{G}} = \mathcal{G}, \widehat{A} = A\Big) \\ &\leq \sum_{A:|A_0\setminus A|=i} \sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_1(\mathcal{G})\Big)\mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_2(\mathcal{G})\Big)\exp\Big(-\frac{t}{\sigma^2}b(\mathcal{G})\Big) \\ &= \sum_{i=1}^{p_0} \sum_{A:|A_0\setminus A|=i} S_i^* \frac{\exp\Big(\frac{t(a-1)^2 niC_{\min}}{(1-2t)\sigma^2}\Big)\exp\Big(-\frac{t}{\sigma^2}(-\lambda + a(a-1)niC_{\min})\Big)}{(1-2at)^{\frac{K_i^*}{2}}(1-2t)^{\frac{n-K_i^*}{2}}} \\ &\leq \sum_{i=1}^{p_0}\binom{p_0}{i}\sum_{j=0}^{i}\binom{p-p_0}{j}\frac{S_i^*}{(1-2t)^{\frac{n}{2}}}\exp\Big(-n\frac{t(a-1)iC_{min}}{\sigma^2}\frac{1-2at}{1-2t}\Big)\Big(\frac{1-2t}{1-2at}\Big)^{\frac{K_i^*}{2}} \end{aligned}$$

where $S_i^* \equiv \max_{A\in\mathcal{A},|A_0\setminus A|=i}|\mathcal{S}_A|$ and $K_i^* \equiv \max_{\mathcal{G}\in\mathcal{S}_A,|A_0\setminus A|=i} K(\mathcal{G})$, as defined. This, together with the fact that $\binom{p_0}{p_0-i} \leq p_0^i$, $\sum_{j=1}^{i}\binom{p-p_0}{j} \leq (p-p_0)^i$ and $(p-p_0)p_0 \leq \frac{p^2}{4}$, yields

$$I \leq \sum_{i=1}^{p_0}\frac{p^2}{4}S_i^*\exp\Big(-n\frac{t(a-1)iC_{min}}{\sigma^2}\frac{1-2at}{1-2t}\Big)\Big(\frac{1-2t}{1-2at}\Big)^{K_i^*/2}\frac{1}{(1-2t)^{n/2}} \quad \text{(A.4)}$$

provided that $\frac{t}{\sigma^2}\lambda \leq 1$. Let $K^* = \max_{1\leq i\leq p_0} K_i^*/i, \log(S^*) = \max_{1\leq i\leq p_0}\log(S_i^*)/i$. For simplification, choose $t = \frac{1}{4(a-1)}$, $c = \frac{2a-3}{a-2} > 2$, and a to satisfy $2\frac{n}{\log S^*} > a > 4 + \frac{n}{4\log S^*}$. Then (A.4) becomes:

$$\begin{aligned} I &\leq \sum_{i=1}^{p_0}\frac{p^2}{4}S_i^*\exp\Big(-n\frac{1}{4c\sigma^2}iC_{min}\Big)c^{K_i^*/2}\frac{1}{(1-2t)^{n/2}} \\ &\leq \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 20\sigma^2\frac{\log|\mathcal{S}|}{n}\big)\Big), \end{aligned}$$

provided that $\tau \leq \frac{2\sigma}{p}\sqrt{\frac{\log p}{2np\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)}}$. This leads to (1.8).

For the risk property, let $D = 25\sigma^2$ and $G = \{\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 \geq D\}$. Then

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 = \frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g - \boldsymbol{X}\boldsymbol{\beta}^0\|^2(\mathbb{I}(G) + \mathbb{I}(G^c)) \equiv T_1 + T_2.$$

For T_1, note that $\frac{1}{4n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 - \frac{1}{2n}\|\boldsymbol{\epsilon}\|^2 \leq \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^g\|^2 \leq \frac{1}{2n}\|\boldsymbol{\epsilon}\|^2$. By Markov's inequality with $t = \frac{1}{3}$, $T_1 = \int_D^\infty \mathbb{P}\big(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 \geq x\big)dx$ is upper bounded by

$$\begin{aligned} &\int_D^\infty \mathbb{P}\Big(\frac{1}{n}\|\boldsymbol{\epsilon}\|^2 \geq \frac{x}{4}\Big)dx \leq \int_D^\infty \mathbb{E}\exp\big(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\big)\exp\big(-nt\frac{x}{4\sigma^2}\big)dx \\ &\leq \int_D^\infty \exp\big(-\frac{n}{12\sigma^2}(x - 24\sigma^2)\big)dx = \frac{12\sigma^2}{n}\exp\big(-\frac{n}{12}\big), \end{aligned}$$

implying that $T_1 = o(\frac{p_0}{n}\sigma^2)$. For T_2, then,

$$\begin{aligned} T_2 &\leq D\mathbb{P}(\hat{\boldsymbol{\beta}}^g \neq \hat{\boldsymbol{\beta}}^{ol}) + \frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 \\ &= 25\sigma^2\mathbb{P}(\hat{\boldsymbol{\beta}}^g \neq \hat{\boldsymbol{\beta}}^{ol}) + \frac{K_0}{n}\sigma^2 = (o(1)+1)\frac{K_0}{n}\sigma^2. \end{aligned}$$

The desired result then follows. This completes the proof.

Proof of Theorem 2: The proof is similar to that of Theorem 3 with some minor modifications. In the present case, let $\widehat{\mathcal{G}}^{l_0}$ be a grouping associated with $\hat{\boldsymbol{\beta}}^{l_0}$. Then $\hat{\boldsymbol{\beta}}^{l_0} = \hat{\boldsymbol{\beta}}^{ol}$ if $\widehat{\mathcal{G}}^{l_0} = \mathcal{G}^0$. This means $\{\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\} = \{\widehat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\}$. Then

$$\mathbb{P}(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}) \leq \sum_{i=0}^{p_0} \mathbb{P}\Big(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, \widehat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\Big) \equiv I$$

Note that $S(\hat{\boldsymbol{\beta}}^{l_0}) \equiv \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{l_0}\big\|^2 \geq \frac{1}{2}\big\|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})(\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon})\big\|^2$. Then

$$\begin{aligned} &2\Big(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol})\Big) \geq \big\|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})(\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon})\big\|^2 - \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\|^2 \\ &= 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{\epsilon}\|^2 - \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\| \\ &\geq 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 - \boldsymbol{\epsilon}^T\boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}}\boldsymbol{\epsilon} \\ &\equiv -L(\widehat{\mathcal{G}}^{l_0}) + b(\widehat{\mathcal{G}}^{l_0}), \end{aligned} \tag{A.5}$$

where $L(\mathcal{G}) \equiv L_1(\mathcal{G}) + L_2(\mathcal{G}) = 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$, $b(\mathcal{G}) = \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2$, and $L_1(\mathcal{G}) \equiv -2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}$ and $L_2(\mathcal{G}) \equiv \boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$, and $L_1(\mathcal{G})$ are $L_2(\mathcal{G})$ are independent. Recall that $\mathcal{S} = \big\{\mathcal{G} : \mathcal{G} \neq \mathcal{G}^0; C_1(\mathcal{G}) \leq p_0; C_2(\mathcal{G}, \mathcal{E}) \leq c_0\big\}$. Let $\widehat{A} = \mathcal{I} \setminus \widehat{\mathcal{I}}_0$, Then, for any $0 < t < 1/2$ by Markov's inequality,

$$\begin{aligned} I &\leq \sum_{A\in\mathcal{A}}\sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{P}\Big(L(\mathcal{G}) \geq b(\mathcal{G}), \widehat{\mathcal{G}} = \mathcal{G}, \widehat{A} = A\Big) \\ &\leq \sum_{A\in\mathcal{A}}\sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_1(\mathcal{G})\Big)\mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_2(\mathcal{G})\Big)\exp\Big(-\frac{t}{\sigma^2}b(\mathcal{G})\Big) \\ &= \sum_{i=1}^{p_0}\sum_{A\in\mathcal{A},|A_0\setminus A|=i} S_i^*\exp\Big(-\frac{t-t^2}{2\sigma^2}niC_{min}\Big)\frac{1}{(1-2t)^{K_i^*}} \\ &\leq \sum_{i=1}^{p_0}\binom{p_0}{p_0-i}\sum_{j=0}^{i}\binom{p-p_0}{j}S_i^*\exp\Big(-\frac{t-t^2}{2\sigma^2}niC_{min}\Big)\frac{1}{(1-2t)^{K_i^*}} \end{aligned}$$

where $S_i^* \equiv \max_{A\in\mathcal{A},|A_0\setminus A|=i}|\mathcal{S}_A|, K(\mathcal{G}), K_i^* \equiv \max_{\mathcal{G}\in\mathcal{S}_A,|A_0\setminus A|=i} K(\mathcal{G})$, as defined. This, together with the fact that $\binom{p_0}{p_0-i} \leq p_0^i$, $\sum_{j=1}^{i}\binom{p-p_0}{j} \leq (p-p_0)^i$ and $(p-p_0)p_0 \leq \frac{p^2}{4}$,

yields

$$I \;\leq\; \sum_{i=1}^{p_0} \frac{p^2}{4} S_i^* \;\exp\Big(-\frac{t-t^2}{2\sigma^2} n i C_{min}\Big) \frac{1}{(1-2t)^{K_i^*}}$$

Let $K^* = \max_{1\leq i\leq p_0} \frac{K_i^*}{i}$ and $\log S^* = \max_{1\leq i\leq p_0} \frac{\log S_i^*}{i}$. To simplify the bound we choose $t = \frac{e-1}{2e} > \frac{3}{10}$, where $\frac{t-t^2}{2} > \frac{1}{10}$

$$I \;\leq\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 10\sigma^2\frac{\log S^*}{n}\big)\Big)$$

This leads to (1.7).

The proof for the risk property is the same and is omitted. This completes the proof.

Proof of Corollary 1: Easily, $K^* \leq K_i^* \leq K_0$. Note that for any $A \subset \mathcal{I}$ with $|A| \neq p_0$, $|S_A| \leq \sum_{i=0}^{K_0-1} \binom{|A|}{i} \leq \sum_{i=0}^{K_0-1} \binom{p_0}{i}$. Thus, $S_i^* = \max_{A\in\mathcal{A},|A_0\setminus A|=i} |\mathcal{S}_A| \leq \sum_{i=1}^{K_0-1} \binom{p_0}{i} \leq p_0^{K_0}$ and $S^* = \exp\big(\max_{1\leq i\leq p_0} \frac{\log S_i^*}{i}\big) \leq \max_{1\leq i\leq p_0} S_i^* \leq p_0^{K_0}$. Using the bounds derived in Theorem 2 and 3, we obtain the desired results.

A.2 Technical details for Chapter 2

Proof of Theorem 1: The equivalence follows directly from Theorem 4.1 in [47]. □

Next we present two lemmas to be used in the proof of Theorem 2.

Lemma 3 *For any* $\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_m \in \mathbb{R}^L$, (A.6) *and* (A.7) *are equivalent:*

$$\exists\ |b_1|, \cdots, |b_m| \leq 1 \ \textit{s.t.}\ \boldsymbol{x}_0 + b_1\boldsymbol{x}_1 + \cdots + b_m\boldsymbol{x}_m = \boldsymbol{0}, \tag{A.6}$$

$$\textit{for}\ \forall\ \boldsymbol{c} \in \mathbb{R}^L, |\boldsymbol{c}^T\boldsymbol{x}_0| \leq |\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}^T\boldsymbol{x}_m|. \tag{A.7}$$

Proof: If (A.6) holds, then for any $\boldsymbol{c} \in \mathbb{R}^L$, $|\boldsymbol{c}^T\boldsymbol{x}_0| = |b_1\boldsymbol{c}^T\boldsymbol{x}_1 + \cdots + b_m\boldsymbol{c}^T\boldsymbol{x}_m| \leq |b_1||\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |b_m||\boldsymbol{c}^T\boldsymbol{x}_m|$, which is no greater than $|\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}^T\boldsymbol{x}_m|$, implying (A.7). For the converse, assume that for any $\boldsymbol{c} \in \mathbb{R}^L$, $|\boldsymbol{c}^T\boldsymbol{x}_0| \leq |\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}^T\boldsymbol{x}_m|$. Consider the following convex minimization:

$$\min_{\{b_1,\cdots,b_m\}} \sum_{i=1}^{m} B(b_i) \quad \text{subject to } \boldsymbol{x}_0 + b_1\boldsymbol{x}_1 + \cdots + b_m\boldsymbol{x}_m = \boldsymbol{0}, \tag{A.8}$$

where $B(x)$ is an indicator function with $B(x) = 0$ when $|x| \leq 1$ and $B(x) = +\infty$ otherwise. First, we need to show that the constraint set in (A.8) is nonempty. Suppose

that it is empty. Let $\boldsymbol{c}_0 = (\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{x}_1,\cdots,x_m)})\boldsymbol{x}_0$, where $\boldsymbol{P}_{(\boldsymbol{x}_1,\cdots,x_m)}$ is the projection matrix onto the linear space spanned by $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$. Since $\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$ are linearly independent, we have that $\|\boldsymbol{c}_0\|_2 > 0$. Therefore $|\boldsymbol{c}_0^T\boldsymbol{x}_0| = \|\boldsymbol{c}_0\|_2^2 > 0 = |\boldsymbol{c}_0^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}_0^T\boldsymbol{x}_m|$, contracting to that $|\boldsymbol{c}^T\boldsymbol{x}_0| \le |\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}^T\boldsymbol{x}_m|$. Hence the constraint set of (A.8) is nonempty and we denote its optimal value by $p^\star$. Next we convert (A.8) to its dual by introducing dual variable $\boldsymbol{\nu} \in \mathbb{R}^L$ for the equality constraints in (A.8) through Lagrange multipliers:

$$\max_{\{\boldsymbol{\nu}\in\mathbb{R}^L\}} \boldsymbol{\nu}^T\boldsymbol{x}_0 - |\boldsymbol{\nu}^T\boldsymbol{x}_1| - \cdots - |\boldsymbol{\nu}^T\boldsymbol{x}_m|. \tag{A.9}$$

By the assumption that $|\boldsymbol{c}^T\boldsymbol{x}_0| \le |\boldsymbol{c}^T\boldsymbol{x}_1| + \cdots + |\boldsymbol{c}^T\boldsymbol{x}_m|$ for any $\boldsymbol{c}$, the maximal of (A.9) $d^\star$ must satisfy $d^\star \le 0$. Hence $d^\star = 0$ because it is attained by $\boldsymbol{\nu} = \boldsymbol{0}$. Moreover, Slater's condition holds because constraint set of (A.8) is nonempty. By the strong duality principle, the duality gap is zero, and hence that $p^\star = d^\star = 0$. Consequently, a minimizer of (A.8) $(b_1,\cdots,b_m)$ exists with $|b_1| \le 1,\cdots,|b_m| \le 1$, satisfying the constraints $\boldsymbol{x}_0 + b_1\boldsymbol{x}_1 + \cdots + b_m\boldsymbol{x}_m = \boldsymbol{0}$. This implies (A.7). This completes the proof. □

Lemma 4 *For $\boldsymbol{s} = (s_1,\cdots,s_L)$ and a connected graph $\mathcal{G} = (V,\mathcal{E})$, there exist $|g_l| \le 1$, $|g_{ll'}| \le 1$, $g_{ll'} = -g_{l'l}$; $1 \le l,l' \le L$ such that*

$$\begin{cases} n_1 s_1 + \lambda_1 g_1 + \lambda_2 \sum_{l'\sim 1} g_{1l'} &= 0 \\ \qquad\vdots & \vdots \\ n_L s_L + \lambda_1 g_L + \lambda_2 \sum_{l'\sim L} g_{Ll'} &= 0, \end{cases} \tag{A.10}$$

is equivalent to $|\sum_{l\in\mathcal{I}} n_l s_l| \le \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I},\mathcal{I}^c)$ for any $\mathcal{I} \subseteq V$ with $d(\mathcal{I},\mathcal{I}^c) = \sum_{l\in\mathcal{I},l'\in\mathcal{I}^c} \mathbb{I}\big(l \sim l'\big)$.

Proof: First, for some $|g_l| \le 1$, $|g_{ll'}| \le 1, g_{ll'} = -g_{l'l}$; $1 \le l,l' \le L$, if (A.10) holds then,

$$|\sum_{l\in\mathcal{I}} n_l s_l| \;=\; \lambda_1\Big|\sum_{l=1} g_l\Big| + \lambda_2\Big|\sum_{l\in\mathcal{I}}\sum_{l'\sim l} g_{ll'}\Big| = \lambda_1\Big|\sum_{l=1} g_l\Big| + \lambda_2\Big|\sum_{l\in\mathcal{I}}\sum_{l'\in\mathcal{I}^c} \mathbb{I}(l\sim l')g_{ll'}\Big|,$$

which is no greater than $\lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I},\mathcal{I}^c)$ for any $\mathcal{I} \subseteq V$. Conversely, by Lemma 3, it suffices to show that for any $\boldsymbol{c} \in \mathbb{R}^L$,

$$|\sum_{l=1}^{L} c_l n_l s_l| \le \lambda_1 \sum_{l=1}^{L} |c_l| + \lambda_2 \sum_{l\sim l'} |c_l - c_{l'}| \tag{A.11}$$

provided that $|\sum_{l\in\mathcal{I}} n_l s_l| \leq \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I},\mathcal{I}^c)$ for any $\mathcal{I}\subseteq V$. To this end, for any permutation $(k_1,\cdots,k_L)\in\sigma(1,\cdots,L)$ and $l=1,\cdots,L$, define convex region $\mathcal{C}_{lk_1\cdots k_L} = \{\boldsymbol{c}=(c_1,\cdots,c_L): c_{k_1}\geq\cdots\geq c_{k_l}\geq 0\geq\cdots\geq c_{k_L}\}$, where $\sigma(1,\cdots,L)$ denotes the set of all possible permutation of $(1,\cdots,L)$. It's easy to see that $\cup_{l=1}^L \cup_{(k_1,\cdots,k_L)\in\sigma(1,\cdots,L)} \mathcal{C}_{lk_1\cdots k_L} = \mathbb{R}^L$ Then, consider function $g(\boldsymbol{c}) = |\sum_{l=1}^L c_l n_l s_l| - \lambda_1\sum_{l=1}^L |c_l| - \lambda_2\sum_{l\sim l'}|c_l - c_{l'}|$. Note that, $g(\boldsymbol{c})$ over each region $\mathcal{C}_{lk_1\cdots k_L}$ is a convex function. By the maximal principle, its maximum (over each region) can be attained at the extreme points of $\mathcal{C}_{lk_1\cdots k_L}$. It is easy to show that the extreme points must be of the form $\boldsymbol{c}=(t\mathbf{1}_{\mathcal{I}},\mathbf{0}_{\mathcal{I}^c})$ for some $\mathcal{I}\subseteq V$ and $t\neq 0$, that is the non-zero components must to equal to each other. Hence, $g(\boldsymbol{c})$ evaluated at the extreme points of $\mathcal{C}_{lk_1\cdots k_L}$ reduces to $|\sum_{l\in\mathcal{I}} n_l s_l| - \lambda_1|\mathcal{I}| - \lambda_2 d(\mathcal{I},\mathcal{I}^c)$ for some $\mathcal{I}\subseteq V$, which, by assumption, is always nonpositive. This completes the proof. □

Proof of Theorem 5: We shall use the KKT condition of (2.12), or local optimality, which is in the form of

$$n_l\hat{\boldsymbol{\Omega}}_l^{-1} + n_l\boldsymbol{S}_l + \lambda_1\partial\|\hat{\boldsymbol{\Omega}}_l\|_{1,\text{off}} + \lambda_2\sum_{l':l\sim l'}\partial\|\hat{\boldsymbol{\Omega}}_l - \hat{\boldsymbol{\Omega}}_{l'}\|_{1,\text{off}} = \mathbf{0}, \quad l=1,\cdots,L, \quad \text{(A.12)}$$

where $\hat{\boldsymbol{\Omega}}_l^{-1}$ is the inversion of matrices $\hat{\boldsymbol{\Omega}}_l$ and $\partial\|\cdot\|_1$ denotes the subgradient of the ℓ_1 function. If $\hat{\boldsymbol{\omega}}_{jkl}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$, then $\left(\hat{\boldsymbol{\Omega}}_l^{-1}\right)_{jk}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$. By Lemma 4, we must have $(s_{jk1},\cdots,s_{jkL})\in\mathcal{S}$ for any $j\in\mathcal{J},k\in\mathcal{J}^c$. Conversely, if $(s_{jk1},\cdots,s_{jkL})\in\mathcal{S}$ for any $j\in\mathcal{J},k\in\mathcal{J}^c$, again by Lemma 4, the KKT condition in (A.12) holds at $\hat{\boldsymbol{\omega}}_{jkl}=0, l=1,\cdots,L$ for jkth components for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$. Hence, $\hat{\boldsymbol{\omega}}_{jkl}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$. This completes the proof. □

Proof of Theorem 6: Let $(\hat{\boldsymbol{\Omega}}_1^{(m)},\cdots,\hat{\boldsymbol{\Omega}}_L^{(m)})$ be the DC solution at iteration m. If the diagonal matrix is initialized as in **Algorithm 1**, then an application of **Theorem 5** on $(\hat{\boldsymbol{\Omega}}_1^{(1)},\cdots,\hat{\boldsymbol{\Omega}}_L^{(1)})$ yields that $(s_{jk1},\cdots,s_{jkL})\in\mathcal{S}$ for any $j\in\mathcal{J},k\in\mathcal{J}^c$, implying that $\hat{\omega}_{jkl}^{(1)}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$. Next, we prove by induction that if $(s_{jk1},\cdots,s_{jkL})\in\mathcal{S}$ for any $j\in\mathcal{J},k\in\mathcal{J}^c$, then $\hat{\omega}_{jkl}^{(m)}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$ holds for any $m\geq 1$. Suppose that $\hat{\omega}_{jkl}^{(m-1)}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$ holds for some $m\geq 2$, then at DC iteration m, $|\hat{\omega}_{jkl}^{(m-1)}|=0\leq\tau, |\hat{\omega}_{jkl}^{(m-1)}-\hat{\omega}_{jkl'}^{(m-1)}|=0\leq\tau$. This, together with **Theorem 5**, again implies that $\hat{\omega}_{jkl}^{(m)}=0$ for any $j\in\mathcal{J},k\in\mathcal{J}^c,1\leq l\leq L$. Using the finite convergence of the DC algorithm, c.f., **Theorem 1**, we have

$(s_{jk1}, \cdots, s_{jkL}) \in \mathcal{S}$ for any $j \in \mathcal{J}, k \in \mathcal{J}^c$, implying that $\hat{\omega}^{dc}_{jkl} = 0$ for any $j \in \mathcal{J}, k \in \mathcal{J}^c, 1 \le l \le L$. Conversely, if for some $\mathcal{J}$ $\hat{\omega}^{dc}_{jkl} = 0$ for any $j \in \mathcal{J}, k \in \mathcal{J}^c, 1 \le l \le L$, consider the next DC iteration, we have $\hat{\omega}^{m*+1}_{jkl} = 0$ for any $j \in \mathcal{J}, k \in \mathcal{J}^c, 1 \le l \le L$. Using the same argument as above with the converse part of **Theorem 5**, we obtain that $(s_{jk1}, \cdots, s_{jkL}) \in \mathcal{S}$ for any $j \in \mathcal{J}, k \in \mathcal{J}^c$. This completes the proof. □

Proof of Corollary 2: For the fused graph, let $\mathcal{I} = \{1, \cdots, l\}$, $\{L-l+1, \cdots, L\}$, $\{l_1+1, \cdots, l_2\}$ and $\{1, \cdots, L\}$, then if $|\sum_{l\in\mathcal{I}} n_l s_l| \le \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I}, \mathcal{I}^c)$ then $\left|\sum_{i=1}^{l} n_i s_i\right| \le l\lambda_1 + \lambda_2$, $\left|\sum_{i=L-l+1}^{L} n_i s_i\right| \le l\lambda_1 + \lambda_2$, $\left|\sum_{i=l_1+1}^{l_2} n_i s_i\right| \le (l_2 - l_1)\lambda_1 + 2\lambda_2$, $\left|\sum_{i=1}^{L} n_i s_i\right| \le L\lambda_1$. Conversely, if $\mathcal{I} = \{1, \cdots, L\}$, then $\left|\sum_{i\in\mathcal{I}} n_i s_i\right| \le \lambda_1|\mathcal{I}| + \lambda_2 d(\mathcal{I}, \mathcal{I}^c) = L\lambda_1$. Next, assume that $\mathcal{I} \neq \{1, \cdots, L\}$, and write $\mathcal{I} = \cup_{k=1}^{q}\{i_k, i_k + 1, \cdots, i_k + l_k\}$ with $i_1 \le i_1 + l_1 < i_2 < i_2 + l_2 < \cdots < i_q < i_q + l_q$. Then

$$\begin{aligned}\left|\sum_{i\in\mathcal{I}} n_i s_i\right| &\le \sum_{k=1}^{q}\left|\sum_{i=i_k}^{i_k+l_k} n_i s_i\right| \le \lambda_1\sum_{k=1}^{q} l_k + 2(q-2)\lambda_2 + \big(\mathbb{I}(i_1 \neq 1) + \mathbb{I}(i_q + l_q \neq L) + 2\big)\lambda_2 \\ &= |\mathcal{I}|\lambda_1 + 2(q-1)\lambda_2 + \big(\mathbb{I}(i_1 \neq 1) + \mathbb{I}(i_q + l_q \neq L)\big)\lambda_2 = |\mathcal{I}|\lambda_1 + d(\mathcal{I}, \mathcal{I}^c)\lambda_2.\end{aligned}$$

In the case of the complete graph, set $\mathcal{I} = \{k_1, \cdots, k_l\}$, $\{k_{L-l+1}, \cdots, k_L\}$, given $s_{k_1} \le \cdots \le s_{k_L}$, then we have $\left|\sum_{i=1}^{l} n_{k_i} s_{k_i}\right| \le l\lambda_1 + l(L-l)\lambda_2$, $\left|\sum_{i=L-l+1}^{L} n_{k_i} s_{k_i}\right| \le l\lambda_1 + l(L-l)\lambda_2$. Conversely, for any $\mathcal{I}$, $\left|\sum_{i\in\mathcal{I}} n_i s_i\right| \le \max\left(\left|\sum_{i=1}^{|\mathcal{I}|} n_{k_i} s_{k_i}\right|, \left|\sum_{i=L-|\mathcal{I}|+1}^{L} n_{k_i} s_{k_i}\right|\right) \le l\lambda_1 + l(L-l)\lambda_2$. This completes the proof. □.

Proof of Theorem 7: The proof uses a large deviation probability inequality of [48] to treat one-sided log-likelihood ratios with constraints. This enables us to obtain sharp results without a moment condition on both tails of the log-likelihood ratios.

Recall that $\mathcal{S} = \big\{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) : |A(\boldsymbol{\beta})| \le d_0, C(\boldsymbol{\beta}, \mathcal{E}) \le c_0, \mathcal{G}(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}^0)\big\}$ and $\mathcal{S}_A = \big\{\boldsymbol{\theta} \in \mathcal{S} : A(\boldsymbol{\beta}) = A\big\}$. Let a class of candidate subsets be $\mathcal{A} \equiv \{A \neq A^0 : |A| \le d_0\}$ for sparseness pursuit. Note that any $A \subset \{1, \cdots, d\}$ can be partitioned into $(A \setminus A^0) \cup (A \cap A^0)$. Then we partition $\mathcal{S}$ accordingly with $\mathcal{S} = \cup_{i=0}^{d_0} \cup_{A\in\mathcal{B}_i} \mathcal{S}_A$, where $\mathcal{B}_i = \mathcal{A} \cap \{A : |A^0 \setminus A| = i\}$, with $|\mathcal{B}_i| = \binom{d_0}{d_0-i}\sum_{j=0}^{i}\binom{d-d_0}{j}, i = 0, \cdots, d_0$. Moreover, $\mathcal{S}_A = \cup_{\mathcal{G}\in\{\mathcal{G}(\boldsymbol{\beta}):\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})\in\mathcal{S}_A\}}\mathcal{S}_{\mathcal{G}}$, where $\mathcal{S}_{\mathcal{G}} = \big\{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S} : \mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}\big\}$. So $\mathcal{S} = \cup_{i=0}^{d_0} \cup_{A\in\mathcal{B}_i} \cup_{\mathcal{G}\in\{\mathcal{G}(\boldsymbol{\beta}):\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})\in\mathcal{S}_A\}}\mathcal{S}_{\mathcal{G}}$.

To bound the error probability, note that if $\hat{\mathcal{G}}^{l_0} = \mathcal{G}^0$ then $\hat{\boldsymbol{\theta}}^{\ell_0} = \hat{\boldsymbol{\theta}}^o$ then $\hat{\boldsymbol{\theta}}^{\ell_0} = \hat{\boldsymbol{\theta}}^o$, by Definition 4. Conversely, if $\hat{\boldsymbol{\theta}}^{\ell_0} = \hat{\boldsymbol{\theta}}^o$ or $\hat{\boldsymbol{\beta}}^{\ell_0} = \hat{\boldsymbol{\beta}}^o$, then $\hat{\mathcal{G}}^{l_0} = \mathcal{G}^0$. Thus $\{\hat{\mathcal{G}}^{l_0} = \mathcal{G}^0\}\{\hat{\boldsymbol{\theta}}^{\ell_0} = \hat{\boldsymbol{\theta}}^o\}$. So $\{\hat{\boldsymbol{\theta}}^{\ell_0} \neq \hat{\boldsymbol{\theta}}^o\} \subseteq \{L(\hat{\boldsymbol{\theta}}^{\ell_0}) - L(\hat{\boldsymbol{\theta}}^o) \ge 0\} \subseteq \{l(\hat{\boldsymbol{\theta}}^{\ell_0}) - l(\boldsymbol{\theta}^0) \ge 0\}$. This together with

$\{\hat{\boldsymbol{\theta}}^{\ell_0} \neq \hat{\boldsymbol{\theta}}^o\} \subseteq \{\hat{\boldsymbol{\theta}}^{\ell_0} \in \mathcal{S}\}$ implies that $\{\hat{\boldsymbol{\theta}}^{\ell_0} \neq \hat{\boldsymbol{\theta}}^o\} \subseteq \{l(\hat{\boldsymbol{\theta}}^{\ell_0}) - l(\boldsymbol{\theta}^0) \geq 0\} \cap \{\hat{\boldsymbol{\theta}}^{\ell_0} \in \mathcal{S}\}$. Consequently, $I \equiv \mathbb{P}\big(\hat{\boldsymbol{\theta}}^{\ell_0} \neq \hat{\boldsymbol{\theta}}^o\big) \leq \mathbb{P}\Big(L(\hat{\boldsymbol{\theta}}^{\ell_0}) - L(\boldsymbol{\theta}^0) \geq 0; \hat{\boldsymbol{\theta}}^{\ell_0} \in \mathcal{S}\Big)$ is upper bounded by

$$\begin{aligned} & \sum_{i=0}^{d_0} \sum_{A \in \mathcal{B}_i} \sum_{\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}): \boldsymbol{\theta} \in \mathcal{S}_A\}} \mathbb{P}^*\Big(\sup_{\boldsymbol{\theta} \in \mathcal{S}_\mathcal{G}} \big(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^0)\big) \geq 0\Big) \\ \leq & \sum_{i=0}^{d_0} \sum_{A \in \mathcal{B}_i} \sum_{\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}): \boldsymbol{\theta} \in \mathcal{S}_A\}} \mathbb{P}^*\Big(\sup_{\boldsymbol{\theta} \in \mathcal{M}} \big(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^0)\big) \geq 0\Big), \end{aligned}$$

where $\mathcal{M} = \big\{-\log(1-h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)) \geq \max(i,1) C_{\min}(\boldsymbol{\theta}^0),\ \boldsymbol{\theta} \in \mathcal{S}_\mathcal{G}\big\}$, $\mathbb{P}^*$ is the outer measure and the last two inequalities use the fact that $\mathcal{S} = \cup_{i=0}^{d_0} \cup_{A \in \mathcal{B}_i} \cup_{\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}): \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S}_A\}} \mathcal{S}_\mathcal{G}$ and $\mathcal{S}_\mathcal{G} \subseteq \{\boldsymbol{\theta} : \max(|A^0 \setminus A|, 1) C_{\min}(\boldsymbol{\theta}^0) \leq -\log(1 - h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0))\}$ for $\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S}_A\}$.

For I, we apply Theorem 1 of [48] to bound each term. Towards this end, we verify their entropy condition (3.1) for the local entropy over $\mathcal{S}_\mathcal{G}$ for $\mathcal{G} \in \{\mathcal{G}(\boldsymbol{\beta}) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S}_A\}$, $A \in \mathcal{B}_i$ and $i = 0, \cdots, d_0$. Under **Assumption A** $\varepsilon = \varepsilon_{n,p_0,p} = (2c_0)^{1/2} c_4^{-1} \log(2^{1/2}/c_3) \log p (\frac{p_0}{n})^{1/2}$ satisfies there with respect to $\varepsilon > 0$, that is,

$$\sup_{\{0 \leq |A| \leq p_0\}} \int_{2^{-8}\varepsilon^2}^{2^{1/2}\varepsilon} H^{1/2}(t/c_3, \mathcal{B}_A) dt \leq p_0^{1/2} 2^{1/2} \varepsilon \log(2/2^{1/2} c_3) \leq c_4 n^{1/2} \varepsilon^2. \qquad \text{(A.13)}$$

for some constant $c_3 > 0$ and c_4, say $c_3 = 10$ and $c_4 = \frac{(2/3)^{5/2}}{512}$. By **Assumption A**, $C_{\min}(\boldsymbol{\theta}^0) \geq \varepsilon_{n,p_0,p}^2$ implies (A.13), provided that $d_0 \geq (2c_0)^{1/2} c_4^{-1} \log(2^{1/2}/c_3)$.

Now, let $S_i^\star = \max_{A \in \mathcal{B}_i} \#\{\mathcal{G} : \mathcal{I}_0^c = A\}$ and $\log(S^\star) = \max_{1 \leq i \leq p_0} \log(S_i^\star)/i$. Using inequalities for binomial coefficients: $\sum_{j=0}^i \binom{d-d_0}{j} \leq (d-d_0)^i$ and $\binom{d_0}{i} \leq d_0^i$, $|\mathcal{B}_i| = \binom{d_0}{d_0-i} \sum_{j=0}^i \binom{d-d_0}{j} \leq \big(d(d-d_0)\big)^i \leq (d^2/4)^i$, we have, by Theorem 1 of [48], that for a constant $c_2 > 0$, say $c_2 = \frac{4}{27} \frac{1}{1926}$,

$$\begin{aligned} I &\leq \sum_{i=0}^{d_0} |\mathcal{B}_i| S_i^\star \exp\big(-c_2 n i C_{\min}(\boldsymbol{\theta}^0)\big) \leq \sum_{i=0}^{d_0} \big(\frac{d^2}{4}\big)^i S_i^\star \exp\big(-c_2 n i C_{\min}(\boldsymbol{\theta}^0)\big) \\ &\leq \exp\big(-c_2 n C_{\min}(\boldsymbol{\theta}^0) + 2\log d + \log(S^\star)\big). \end{aligned}$$

This completes the proof. □

Proof of Theorem 8: The proof is similar to that of Theorem 7 with some minor modifications. Given τ and $\boldsymbol{\beta} \in \mathbb{R}^d$, a partition $\mathcal{G}^\tau(\boldsymbol{\beta}) = (\mathcal{I}_0(\boldsymbol{\beta}), \cdots, \mathcal{I}_{K(\boldsymbol{\beta})}(\boldsymbol{\beta}))$ associated with $\boldsymbol{\beta}$ is defined to satisfy the following (i) $\max_{j \in \mathcal{I}_0(\boldsymbol{\beta})} |\boldsymbol{\beta}_j| \leq \tau$; (ii) $\big|\beta_{j_1} - \beta_{j_2}\big| \leq \tau$ for any j_1, j_2 in different groups. Let $A^\tau(\boldsymbol{\beta}) = \mathcal{I} \setminus \mathcal{I}_0(\boldsymbol{\beta})$.

The rest of the proof is basically the same as that in Theorem 2 with a modification that $\mathcal{G}(\boldsymbol{\beta})$ and $A(\boldsymbol{\beta})$ are replaced by $\mathcal{G}^\tau(\boldsymbol{\beta})$ and $A^\tau(\boldsymbol{\beta})$ respectively. Here, $\mathcal{S} = \big\{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) : \sum_{j=1}^p J_\tau(|\beta_j|) \le d_0, \sum_{(jj')\in\mathcal{E}} J_\tau\big(\big|\beta_j - \beta_{j'}\big|\big) \le c_0, \mathcal{G}^\tau(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}^0)\big\}$, $C^\tau(\boldsymbol{\beta}, \mathcal{E}) = \sum_{(jj')\in\mathcal{E}} I(|\beta_j - \beta_{j'}| \neq 0)$, $\mathcal{S}_A = \big\{\boldsymbol{\theta} \in \mathcal{S} : A^\tau(\boldsymbol{\beta}) = A\big\}$ and $\mathcal{S}_\mathcal{G} = \big\{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathcal{S} : \mathcal{G}^\tau(\boldsymbol{\beta}) = \mathcal{G}\big\}$.

Next, we show that $\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o$ if and only if $\hat{\mathcal{G}}^g = \mathcal{G}^0$, where $\hat{\mathcal{G}}^g \equiv \mathcal{G}^\tau(\hat{\boldsymbol{\beta}}^g)$. Now $d_1 \equiv |\mathcal{I} \setminus \hat{\mathcal{I}}_0^0| = d_0$. By (2.14), $\frac{1}{\tau}\sum_{j\in\hat{\mathcal{I}}_0} |\hat{\beta}_j^g| + d_1 \le d_0$, with $d_0 = d_1$, yields that $\hat{\beta}_j^g = 0$; $j \in \mathcal{I}_0^1$. In addition, the second constraint of (2.14) implies $\sum_{i=1}^K \sum_{jj'\in\mathcal{I}_i,(jj')\in\mathcal{E}} \frac{|\hat{\beta}_j^g - \hat{\beta}_{j'}^g|}{\tau} \le 0$, yielding that $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$ for any $j_1, j_2 \in \hat{\mathcal{I}}_i, (j_1, j_2) \in \mathcal{E}$, $i = 1, \cdots, K$. By graph consistency of $\mathcal{U}$, $\mathcal{U}$ is connected over $\hat{\mathcal{I}}_i$, implying that $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$ for any $j_1, j_2 \in \hat{\mathcal{I}}_i$, $i = 1, \cdots, K$. This further implies that $\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^o$ and $\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o$, meaning that that $\{\widehat{\mathcal{G}}^g = \mathcal{G}^0\} \subseteq \{\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o\}$. On the other hand, it is obvious that if $\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o$ then $\{\widehat{\mathcal{G}}^g = \mathcal{G}^0\}$. Hence, $\{\widehat{\mathcal{G}}^g = \mathcal{G}^0\} = \{\hat{\boldsymbol{\theta}}^g = \hat{\boldsymbol{\theta}}^o\}$ from which we conclude that $\{\hat{\boldsymbol{\theta}}^g \neq \hat{\boldsymbol{\theta}}^o\} \subseteq \{\hat{\boldsymbol{\theta}}^g \in \mathcal{S}\}$. This together with $\{\hat{\boldsymbol{\theta}}^g \neq \hat{\boldsymbol{\theta}}^o\} \subseteq \{L(\hat{\boldsymbol{\theta}}^g) - L(\boldsymbol{\theta}^o) \ge 0\} \subseteq \{L(\hat{\boldsymbol{\theta}}^g) - L(\boldsymbol{\theta}^0) \ge 0\}$ implies that $\mathbb{P}\big(\hat{\boldsymbol{\theta}}^g \neq \hat{\boldsymbol{\theta}}^o\big) \le \mathbb{P}\Big(L(\hat{\boldsymbol{\theta}}^g) - L(\boldsymbol{\theta}^0) \ge 0; \hat{\boldsymbol{\theta}}^g \in \mathcal{S}\Big)$ is bounded by

$$
\begin{aligned}
&\sum_{i=0}^{d_0} \sum_{A\in\mathcal{B}_i} \sum_{\mathcal{G}\in\{\mathcal{G}^\tau(\boldsymbol{\beta}):\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})\in\mathcal{S}_A\}} \mathbb{P}^*\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_\mathcal{G}} \big(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^0)\big) \ge 0\Big) \\
\le\ &\sum_{i=0}^{d_0} \sum_{A\in\mathcal{B}_i} \sum_{\mathcal{G}\in\{\mathcal{G}^\tau(\boldsymbol{\beta}):\boldsymbol{\theta}\in\mathcal{S}_A\}} \mathbb{P}^*\Big(\sup_{\left\{-\log(1-h^2(\boldsymbol{\theta},\boldsymbol{\theta}^0)) \ge d_1 \max(i,1) C_{\min}(\boldsymbol{\theta}^0) - d_3\tau^{d_2} d, \boldsymbol{\theta}\in\mathcal{S}_\mathcal{G}\right\}} \\
&\big(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^0)\big) \ge 0\Big),
\end{aligned}
$$

where the last step uses the fact that

$$
\begin{aligned}
\{\boldsymbol{\theta} \in \mathcal{S}_\mathcal{G}\} &\subseteq \{-\log(1 - h^2(\boldsymbol{\theta}^\tau, \boldsymbol{\theta}^0)) \ge \max(i,1) C_{\min}(\boldsymbol{\theta}^0)\} \\
&\subseteq \{-\log(1 - h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)) \ge d_1 \max(i,1) C_{\min}(\boldsymbol{\theta}^0) - d_3\tau^{d_2} d\},
\end{aligned}
$$

under **Assumption B**. Then, for some constant c_3, $\mathbb{P}\big(\hat{\boldsymbol{\theta}}^g \neq \hat{\boldsymbol{\theta}}^o\big)$ is upper bounded by

$$
\begin{aligned}
&\sum_{i=0}^{d_0} |\mathcal{B}_i| S_i^\star \exp\big(-c_3 n i C_{\min}(\boldsymbol{\theta}^0)\big) \le \sum_{i=0}^{d_0} \big(\frac{d^2}{4}\big)^i S_i^\star \exp\big(-c_3 n i C_{\min}(\boldsymbol{\theta}^0)\big) \\
\le\ &\exp\big(-c_3 n C_{\min}(\boldsymbol{\theta}^0) + 2\log d + \log(S^\star)\big),
\end{aligned}
$$

provided that $\tau \le \big(\frac{(d_1 - c_3) C_{\min}(\boldsymbol{\theta}^0)}{d_3 d}\big)^{1/d_2}$. This completes the proof. □

Proof of Corollary 3: First we derive an upper bound of S^*. Let p_l^0 the number of nonzero elements of the precision matrix in the sth cluster for $l = 1, \cdots, L$. Let $p_0 = d_0/L = \frac{p_1^0+\cdots+p_L^0}{L}$ be the average number of nonzero elements. For any $\boldsymbol{\theta} \in \mathcal{S}_A$ with $|A^0 \setminus A| = i$, let $\mathcal{Q} = \{(j,k) : j > k, \exists l, x_{jul} \neq 0\}$ and $|\mathcal{Q}| = q_0$. Let $a_{jk} = \#\{l : x_{jkl} \neq 0\}$ for $(j,k) \in \mathcal{Q}$. Note that $q_0 \leq p^0$ and $\sum_{(j,k)\in\mathcal{Q}} a_{jk} \leq d_0$ since $|A| \leq d_0$. By the definition of S_i^*, we have

$$\begin{aligned} S_i^* &\leq \sum_{\sum_{(j,k)\in\mathcal{Q}} r_{jk} \leq g_0} \prod_{(j,k)\in\mathcal{Q}} \binom{a_{jk}-1}{r_{jk}} = \sum_{g=0}^{g_0} \binom{\sum_{(j,k)\in\mathcal{Q}} a_{jk} - q_0}{g} \\ &\leq \sum_{g=0}^{g_0} \binom{d_0 - p_0}{g} \leq (g_0+1)\big(e\frac{d_0-p_0}{g_0}\big)^{g_0}. \end{aligned} \tag{A.14}$$

This together with $\log(1+g_0) \leq g_0$ implies $\log S^* \leq 2g_0 \max(\log(d_0/g_0), 1)$. To lower bound $C_{min}(\boldsymbol{\theta}^0)$, we proceed similarly with the proof of **Proposition 2** in [17]. Specifically, note that $h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = 1 - \prod_{l=1}^L \big(1 - h^2(\boldsymbol{\Omega}_l, \boldsymbol{\Omega}_l^0)\big)$. Thus,

$$-\log(1 - h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)) = \sum_{l=1}^L \Big(-\log\big(1 - h^2(\boldsymbol{\Omega}_l, \boldsymbol{\Omega}_l^0)\big)\Big). \tag{A.15}$$

An application of Proposition 2 of [17] yields that each term in (A.15) is lower bounded by $c^\star \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^0\|_2^2$. Therefore, $-\log(1 - h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)) \geq c^\star \min_{1\leq l\leq L} c_{min}\big(H_l\big) \sum_{l=1}^L \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^0\|_2^2$. Now if $A_0 \setminus A \neq \emptyset$, we have $\sum_{l=1}^L \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^0\|_2^2 \geq |A_0 \setminus A| \min_{(j,k,l):\omega_{jkl}\neq 0} \omega_{jkl}^2$. If $A_0 \setminus A = \emptyset$, then by definition of $\mathcal{S}$, there must exist (j,k,l) such that $\omega_{jkl} = \omega_{jk(l+1)}$ and $\omega_{jkl}^0 \neq \omega_{jk(l+1)}^0$. Here

$$\begin{aligned} \sum_{l=1}^L \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^0\|_2^2 &\geq (\omega_{jkl} - \omega_{jkl}^0)^2 + (\omega_{jk(l+1)} - \omega_{jk(l+1)}^0)^2 \geq \frac{1}{2}(\omega_{jkl}^0 - \omega_{jk(l+1)}^0)^2 \\ &\geq \frac{1}{2} \min_{\{(j,k,l):\ \omega_{jkl}^0 \neq \omega_{jk(l+1)}^0\}} \big(\omega_{jkl}^0 - \omega_{jk(l+1)}^0\big)^2. \end{aligned}$$

A combination of both the cases yield that

$$-\log(1 - h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0))/\max\big(|A_0 \setminus A|, 1\big) \geq c^\star \min_{1\leq l\leq L} c_{min}\big(H_l\big)\eta_{min}^2.$$

which, after taking infimum over $\mathcal{S}$, leads to $C_{min}(\boldsymbol{\theta}^0) \geq c^\star c_{min}\big(H_l\big)\eta_{min}^2$. This, together with, the upper bound on $\log S^\star$ in Theorems 7 and 8, gives a sufficient condition for simultaneous pursuit of sparseness and clustering: $\min_{1\leq l\leq L} c_{min}\big(H_l\big)\eta_{min}^2 \geq$

$c_0 \frac{\log(p^2 L) - g_0 \max(\log(d_0/g_0), 1)}{n}$, for some $c_0 > 0$. Moreover, under this condition, $\mathbb{P}\Big(\hat{\mathbf{\Omega}}^{\ell_0} \neq \hat{\mathbf{\Omega}}^{o}\Big)$ and $\mathbb{P}\Big(\hat{\mathbf{\Omega}}^{g} \neq \hat{\mathbf{\Omega}}^{o}\Big) \to 0$ as $n, d \to +\infty$. This completes the proof. $\square$