Journal of Machine Learning Research 27 (2026) 1-63 Submitted 5/24; Revised 4/26; Published 6/26

Simultaneous Identification of Sparse Structures and Communities in Heterogeneous Graphical Models

Dapeng Shi DAPENGSHI@CUHK.EDU.HK
Department of Statistics and Data Science
The Chinese University of Hong Kong
Shatin, Hong Kong, 999077, China

Tiandong Wang * TD_WANG@FUDAN.EDU.CN
Shanghai Center for Mathematical Sciences
Fudan University, Shanghai, 200433, China
Shanghai Academy of Artificial Intelligence for Science
Shanghai, 200003, China

Zhiliang Ying ZY9@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: Xiaotong Shen

Abstract

Exploring and detecting community structures hold significant importance in genetics, social sciences, biology, neuroscience and finance, among others. Graphical models are a useful and key tool for community detection through the exploration of sets of variables with group-like properties. In this paper, within the framework of Gaussian graphical models, we propose a decomposition of the underlying graphical structure into low-rank diagonal blocks and a sparse component. This new approach captures the non-overlapping community structure, while uncovering the underlying connectivity, both within and between communities. We show the significance of this decomposition through two modeling perspectives and propose a three-stage estimation procedure with an efficient algorithm to estimate the sparse structure and communities. We provide conditions to ensure local identifiability and extend the traditional irrepresentability condition to an adaptive form by constructing an effective norm that guarantees the consistency of model selection for the adaptive ℓ_1 penalized estimator in the second stage. We also provide the clustering error bound for the K-means procedure in the third stage. We conduct extensive numerical experiments to assess the performance of the proposed method and compare it with existing methods. Finally, we apply our method to a dataset of stock returns, demonstrating its capability to accurately identify non-overlapping community structures.

Keywords: Gaussian graphical models, low-rank diagonal blocks, community detection, adaptive penalty, K-means clustering

∗. T. Wang is the corresponding author.

1. Introduction

Probabilistic graphical models are widely used in various domains such as gene regulatory networks (Li et al., 2012; Fan et al., 2009), social networks (Guo et al., 2015; Tan et al., 2014), brain connectivity networks (Li and Solea, 2018; Pircalabelu and Claeskens, 2020; Eisenach et al., 2020), and business networks (Zhao et al., 2024). In graphical models, nodes represent random variables, while edges denote conditional independence/dependence among different random variables. Some commonly used models include Gaussian graphical models (Friedman et al., 2007), Ising models (Chen et al., 2016), and more general exponential family graphical models (Wainwright et al., 2008).

The focus of this paper is on Gaussian graphical models. The problem of structure estimation in Gaussian graphical models has been extensively investigated in the literature under the assumption of sparsity (Friedman et al., 2007; Yuan and Lin, 2007; Tan et al., 2014; Ravikumar et al., 2011; Meinshausen and Bühlmann, 2006; Fan et al., 2009). However, recent attention has shifted towards addressing the estimation challenges posed by heterogeneous graph structures. Two primary types of heterogeneity characterize these structures. The first arises from the hierarchical structure of the population, leading to partially different graph structures in distinct subpopulations (Guo et al., 2011; Danaher et al., 2014; Hao et al., 2018). Additionally, graph structures may depend on individual characteristics, leading to subject-specific graphical models. To address such heterogeneity, decomposition-based approaches to graph estimation have been proposed, wherein each substructure represents the subgraph structure generated by different covariates (Zhang and Li, 2023; Wang et al., 2022; Niu et al., 2023).

Another form of heterogeneity arises from the community attributes of nodes. Nodes within the same community tend to be densely connected, whereas nodes in different communities exhibit sparse connections. Consequently, nodes are typically more homogeneous within communities than across communities. Community detection, a well-explored task in network data analysis (Holland et al., 1983; Amini et al., 2013; Amini and Levina, 2018), focuses on identifying these groups of nodes. However, a fundamental distinction between probabilistic graphical models and complex network models lies in the fact that, in the latter, the graph structure is predetermined at the outset of the analysis. In contrast, our objective here is to estimate the graph structure based on the data.

For node heterogeneity, decomposition-based methods can be used to identify communities and characterize the remaining sparse structures that contain information both between and within communities. When hub nodes with large degrees emerge within a community, Tan et al. (2014) and Tarzanagh and Michailidis (2018) proposed to use the ℓ_2 group Lasso penalty (Yuan and Lin, 2006) and the ℓ_1 penalty (Tibshirani, 1996) to identify hubs and the remaining sparse structure, respectively. It is noteworthy that the sparse structure may also provide important insights into the underlying model structure. For example, identifying the remaining sparse structure proves to be beneficial for distinguishing subtypes when analyzing the differential gene expressions (Danaher et al., 2014; Hao et al., 2018).

To the best of our knowledge, there has been limited exploration to simultaneously detect communities with comparable degrees and identify sparse structures through the decomposition of graph structures. The majority of the existing literature concentrates on estimating the overall graph structure without leveraging the additional information available through

decomposition. For example, Tan et al. (2015) proposed a two-stage method that initially clusters nodes and then assumes the overall graph to be block diagonal. Wilms and Bien (2022) introduced a tree-based node-aggregation method that encodes side information to construct relationships within communities, even in scenarios where overlapping communities are allowed. These approaches, however, only facilitate the estimation of the subgraph structure corresponding to each community but overlook the relationship between communities. In addition, Pircalabelu and Claeskens (2020) conducted community detection by constructing a new pseudo-covariance matrix through the perturbation of community information to the observed covariance matrix. Eisenach et al. (2020) employed the factor model to incorporate community information and developed inference tools for the graph structure of the factors. Moreover, Ma et al. (2021) and Yuan and Zhang (2014) explored partial graphical models in a similar spirit. Although these methods consider sparse connections between communities, they are ineffective in extracting information within communities.

In this paper, we introduce a three-stage estimation procedure aiming at simultaneously identifying both sparse structures and communities by properly decomposing the graph structure. Our key contributions are summarized as follows:

1. We present a novel decomposition of the overall graph structure. Specifically, we decompose the graph structure of the regression residual into two components. One component consists of low-rank diagonal blocks, capturing the non-overlapping community structure. The other component is sparse, revealing the underlying connectivity within and between communities.

2. We also provide a three-stage estimation procedure with efficient algorithms to recover the sparse structure and communities. Especially after removing globally influential factors, we propose an adaptive ℓ_1 penalized estimation method to obtain a low-rank matrix with certain diagonal-block sparsity, aiming to capture the initial structure of the communities. Furthermore, we apply the K-means clustering method based on the resulting estimator to obtain specific membership labels. Besides, we develop a fast and efficient algorithm to implement the proposed graph decomposition procedure.

3. To ensure the identifiability of the sparse and community parts, we extend the classical low-rank manifold and its corresponding tangent space to accommodate low-rank diagonal blocks. Building upon this extension, we introduce a practical norm that generalizes the traditional irrepresentability condition and establish the model selection consistency for the adaptive ℓ_1 penalized estimation in the second stage. We further establish a clustering error bound for the K-means procedure in the third stage.

4. Finally, we apply the proposed method to the synthetic and real datasets. The results show that the proposed method can recover both community structure and sparse conditional dependencies with good interpretability. Meanwhile, in terms of uncovering sparse structures, the proposed method performs favorably against some well-established methods.

The remainder of the paper is organized as follows. Section 2 gives a general framework for the decomposition of graph structure, which consists of two modeling strategies. One is inspired by the stochastic block model widely used in network analysis, while the other

arises when there exist grouped latent variables. Section 3 gives a detailed statement of the proposed adaptive ℓ_1 penalized estimation method, including the optimization algorithm, weight construction, and the choice of tuning parameters. We then develop theoretical results on identifiability and asymptotic properties in Subsections 4.1 and 4.2, respectively. In Sections 5 and 6, we illustrate the proposed framework on several synthetic and real datasets. Concluding remarks are collected in Section 7.

1.1 Notation

We first introduce some notation to be used throughout the paper. Let $\mathbb{R}^p$ be the p-dimensional vector space, and $\mathbb{R}^{p\times q}$, $\mathbb{S}^{p\times p}$ be the sets of $p\times q$ matrices and symmetric $p\times p$ matrices, respectively. For a positive integer m, let $[m]\triangleq\{1,\ldots,m\}$. For two positive sequences a_n and b_n, $a_n\lesssim b_n$ if $\varlimsup_{n\to\infty} a_n/b_n<\infty$, and $a_n\asymp b_n$ if $0<\varliminf_{n\to\infty} a_n/b_n<\varlimsup_{n\to\infty} a_n/b_n<\infty$. Also, for two sequences of positive random variables x_n, y_n, $x_n\lesssim_P y_n$ if $\lim_{m\to\infty}\sup_n \mathbb{P}(x_n/y_n>m)=0$.

For a matrix $A=(a_{ij})_{p\times q}$, the submatrix corresponding to some rectangle set S is denoted by $A_S\triangleq(a_{ij})_{(i,j)\in S}$. The support of A is defined as $\mathrm{Supp}(A)=\{(i,j)|a_{ij}\neq 0\}$ and the rank of A is $\mathrm{rank}(A)$. We let $\mathbf{0}_{p\times q}$ be the matrix of all zeros in $\mathbb{R}^{p\times q}$. We also denote the following norms for a matrix A:

$$\|A\|_0\triangleq\#\{(i,j)|a_{ij}\neq 0\},\qquad \|A\|_{\mathrm{off},1}\triangleq\sum_{i\neq j}|a_{ij}|,$$
$$\|A\|_\infty\triangleq\max_{ij}|a_{ij}|,\qquad \|A\|_F\triangleq\sqrt{\sum_{ij}|a_{ij}|^2}.$$

We use $v(A)\in\mathbb{R}^{p^2}$ to denote the vectorization of A and $\mathrm{diag}(A)\in\mathbb{R}^{p\times p}$ to denote the diagonal matrix of $A\in\mathbb{R}^{p\times p}$. We use $A\succeq 0$ if A is positive semidefinite and $A\succ 0$ if A is positive definite. In addition, for $A\succeq 0$, we let $\|A\|_2$ and $\|A\|_*$ be its respective spectral norm and nuclear norm.

Finally, for a linear subspace $\mathcal{T}$, let $\mathcal{T}^\perp$ be the orthogonal complement of $\mathcal{T}$, and $\mathcal{T}_1\oplus\mathcal{T}_2$ be the orthogonal direct sum space of two linear spaces $\mathcal{T}_1,\mathcal{T}_2$.

2. Graphical Models with Community-based Decomposition

We now introduce a decomposition for the underlying graph structure to uncover community structure, with a focus on non-overlapping communities. We start by assuming that the observed p-dimensional variables $\boldsymbol{X}=(X_1,\cdots,X_p)^T$ and q-dimensional covariates $\boldsymbol{C}=(C_1,\cdots,C_q)^T$ follow the regression model:

$$\boldsymbol{X}|\boldsymbol{C}=B^T\boldsymbol{C}+\boldsymbol{R},\quad \boldsymbol{R}\sim N(0,\Sigma),$$

where $B\in\mathbb{R}^{q\times p}$ represents the regression coefficient matrix. Each component in $\boldsymbol{X}$ is associated with a node in an undirected graph $G(E,V)$, where $V=\{1,\cdots,p\}$ denotes the set of nodes, and $E=\{\{a,b\}\,|a\neq b\in V\}$ the set of undirected edges. Let

$\widetilde{\Theta} \triangleq (B\,\mathrm{Var}(\boldsymbol{C})B^T + \Sigma)^{-1} = (\tilde{\theta}_{ij})_{p\times p}$ be the precision matrix of $\boldsymbol{X}$, with $\tilde{\theta}_{ij} = 0$ equivalent to $\{i, j\} \notin E$ under the Gaussian assumption for both $\boldsymbol{C}$ and $\boldsymbol{R}$ (Lauritzen, 1996). Although the precision matrix $\widetilde{\Theta}$ generally does not exhibit community structures, the precision matrix of the residual $\boldsymbol{R}$ may show some community structures after properly accounting for the covariates with global effects. In particular, we decompose the precision matrix $\Theta \triangleq \Sigma^{-1}$ of $\boldsymbol{R}$ as:

$$\Theta = L + S, \quad L \succeq 0 \quad (\text{or } -L \succeq 0), \tag{1}$$

where $L \in \mathbb{S}^{p\times p}$ consists of low-rank diagonal blocks up to row or column permutations to capture the non-overlapping communities, and $S \in \mathbb{S}^{p\times p}$ is sparse (see Assumptions 1 and 3 for a detailed description). Notice that when there is only one block, such decomposition is the same as the setup in Chandrasekaran et al. (2012). Such sparse plus low-rank decomposition can be locally identifiable under mild conditions, which has been carefully studied (Chandrasekaran et al., 2011; Candes et al., 2011). Additionally, it is worth noting that we do not assume that S has the sparse structure with the same diagonal blocks as L. This implies that within the off-diagonal blocks of L, which are entirely composed of zeros, S does not necessarily have to be zero.

Such decomposition is equivalent to assuming that with global covariates removed, the remaining dependence structure is driven by community-specific latent factors and sparse connections. By representing L as a diagonal-block low-rank matrix, we capture these localized co-movements, e.g. sector-specific drivers in finance (De Nard et al., 2021; Žignić et al., 2024), pathway-level regulators in genomics (Yu et al., 2024; Xue et al., 2023), and system-wide latent processes in neuroscience (Wang and Guo, 2023; Uddin et al., 2023). Also, the sparse matrix S captures the remaining idiosyncratic or direct dependence that may cross community boundaries, including cross-sector supply chain links, inter-pathway regulatory interactions, and sparse functional connectivity between brain regions. This decomposition separates broad group-specific effects from sparse pairwise interactions.

2.1 Latent Community Graphical Models

The first perspective is motivated by the stochastic block model (SBM) in network analysis (Amini and Levina, 2018; Li et al., 2021; Lei and Rinaldo, 2015). Suppose there exist m non-overlapping latent communities. Suppose each node i belongs to the k-th latent community, and denote its community loading by $\boldsymbol{a}_i := (0, \cdots, a_{ik}, \cdots, 0)^T \in \mathbb{R}^r$. In the classical SBM, one typically has $a_{ik} \in \{0, 1\}$. Here we relax this constraint by assuming $a_{ik} \in \mathbb{R}$, which quantifies the strength of the k-th community effect on node i. We also allow each community effect to be multidimensional. Specifically, for a node in community k, write $\boldsymbol{a}_i = (0, \cdots, a_{i1}, \cdots, a_{ir_k}, \cdots, 0)^T$. For nodes i and j belonging to different communities, assume that $\mathrm{Supp}(\boldsymbol{a}_i) \cap \mathrm{Supp}(\boldsymbol{a}_j) = \emptyset$, and let $\sum_{k=1}^m r_k = r$.

Hence, we model the precision matrix $\Theta = (\theta_{ij})_{p\times p}$ of $\boldsymbol{R}$ as:

$$\theta_{ij} = (\boldsymbol{a}_i)^T \boldsymbol{a}_j + s_{ij}, \tag{2}$$

where, after removing the effects of $\boldsymbol{C}$, an edge in E is driven by two components. The first is the community effect $(\boldsymbol{a}_i)^T \boldsymbol{a}_j$. The second is the remaining sparse effect s_{ij}. Let

$A = (\boldsymbol{a}_1^T, \cdots, \boldsymbol{a}_p^T)^T$ and $L = AA^T$. Then model (2) is exactly the model (1) with $L \succeq 0$. We refer to such models as latent community graphical models.

Note that graphical and network models differ in focus. In network analysis, for the general decomposition $\Theta = APA^T + S$ with a connectivity probability matrix P, inference typically targets edge probabilities. Across-community sparsity is often induced by small off-diagonal entries in P. In contrast, graphical models emphasize conditional independence, where even small off-diagonal entries in P may still imply dense conditional dependence. Therefore, generating community structures in graphical models requires sparsity in the precision matrix itself. In our setting, P is assumed diagonal.

2.2 Grouped Latent Variable Graphical Models

The second perspective is an extension of the classical latent variable graphical model (Chandrasekaran et al., 2012). Let $\boldsymbol{X} \in \mathbb{R}^p$ denote the observed variables. Assume there exist m groups of latent variables (unobserved variables) $\boldsymbol{Z}_1 \in \mathbb{R}^{r_1}, \cdots, \boldsymbol{Z}_m \in \mathbb{R}^{r_m}$, where $\boldsymbol{Z}_i = (Z_{i,1}, \cdots, Z_{i,r_i})^T$ and $r = \sum_{i=1}^m r_i$. Conditioning on $\boldsymbol{C}$, we assume

$$(\boldsymbol{X}^T, \boldsymbol{Z}_1^T, \cdots, \boldsymbol{Z}_m^T)^T = B_{(OH)}^T \boldsymbol{C} + \boldsymbol{R}_{(OH)}, \qquad \boldsymbol{R}_{(OH)} \sim N(0, \Theta_{(OH)}^{-1}),$$

where $\Theta_{(OH)}$ is the precision matrix of the joint errors and $B_{(OH)} \in \mathbb{R}^{q \times (p+r)}$ denotes the joint regression coefficients on $\boldsymbol{C}$.

Divide $\Theta_{(OH)}$ into submatrices $\Theta_O, \Theta_{O,H}, \Theta_H$, which specify the dependence among the observed variables, between the observed and latent variables, and among the latent variables, respectively. Let $B_O \in \mathbb{R}^{q \times p}$ be the first p columns of $B_{(OH)}$ associated with the regression coefficients of $\boldsymbol{X}$ on $\boldsymbol{C}$, and $\boldsymbol{R}_O$ be the first p components of the error term $\boldsymbol{R}_{(OH)}$. Then the regression becomes

$$\boldsymbol{X}|\boldsymbol{C} = B_O^T \boldsymbol{C} + \boldsymbol{R}_O, \quad \boldsymbol{R}_O \sim N(0, \Theta^{-1}), \quad \Theta = \Theta_O - \Theta_{O,H}(\Theta_H)^{-1}\Theta_{O,H}^T. \tag{3}$$

Now assume $\boldsymbol{X}$ admits the partition $\{\boldsymbol{X}_{G_1}, \cdots, \boldsymbol{X}_{G_m}\}$. For each $i \in [m]$, we impose the conditional independence assumption

$$(\boldsymbol{X}_{G_i}, \boldsymbol{Z}_i) \perp \boldsymbol{Z}_{-i} \mid (\boldsymbol{C}, \boldsymbol{X}_{-G_i}), \quad \forall i \in [m], \tag{4}$$

where $\boldsymbol{Z}_{-i} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_{i-1}, \boldsymbol{Z}_{i+1}, \cdots, \boldsymbol{Z}_m)$, and $\boldsymbol{X}_{-G_i} = (\boldsymbol{X}_{G_1}, \cdots, \boldsymbol{X}_{G_{i-1}}, \boldsymbol{X}_{G_{i+1}}, \cdots, \boldsymbol{X}_{G_m})$. Under (4), each latent group $\boldsymbol{Z}_i$ interacts only with the observed variables $\boldsymbol{X}_{G_i}$, and cross-group couplings between $(\boldsymbol{X}_{G_i}, \boldsymbol{Z}_i)$ and $\boldsymbol{Z}_{-i}$ vanish (up to permutations that group variables by $\{G_i\}$ and $\{Z_i\}$). Hence, $\Theta_{O,H}$ admits a diagonal-block form $\Theta_{O,H} = \mathrm{diag}(\Theta_{G_1,H_1}, \ldots, \Theta_{G_m,H_m})$. Therefore,

$$L = \Theta_{O,H}\Theta_H^{-1}\Theta_{O,H}^T = \mathrm{diag}(L_1, \ldots, L_m), \qquad L_i = \Theta_{G_i,H_i}\Theta_{H_i}^{-1}\Theta_{G_i,H_i}^T \succeq 0,$$

and each block L_i has rank at most $\dim(\boldsymbol{Z}_i)$.

This assumption generalizes Chandrasekaran et al. (2012) by allowing important latent variables to exist in multiple communities. In this formulation, responses in each community are connected only to the latent variables within that community. Sparse connections between communities are established solely through the observed variables. Under the sparsity assumption for Θ_O, the decomposition (3) is also a case of the model (1).

In the sequel, our analysis is solely based on the assumption that $L \succeq 0$, though it can be easily extended to $-L \succeq 0$.

3. Proposed Method

3.1 A Three-stage Estimation Procedure

We propose a three-stage estimation procedure to estimate the parameters (B, S, L) and community memberships. In the first stage, we estimate B using the least squares method, provided that $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$ is nonsingular,

$$\widehat{B} = (\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}})^{-1}\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{X}},$$

where $\widetilde{\boldsymbol{C}} = (\tilde{\boldsymbol{c}}_1, \cdots, \tilde{\boldsymbol{c}}_n)^T \in \mathbb{R}^{n\times q}, \tilde{\boldsymbol{c}}_i = (\tilde{c}_{i1}, \cdots, \tilde{c}_{iq})^T, i = 1, \cdots, n$ and $\widetilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_n)^T \in \mathbb{R}^{n\times p}, \tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \cdots, \tilde{x}_{ip})^T, i = 1, \cdots, n$. Then we obtain a residual estimate by the plug-in method, i.e., $\widehat{\boldsymbol{R}} = (\boldsymbol{I}_n - \widetilde{\boldsymbol{C}}(\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}})^{-1}\widetilde{\boldsymbol{C}}^T)\widetilde{\boldsymbol{X}}$ with $\boldsymbol{I}_n$ being the n dimensional identity matrix.

In the second stage, we focus on the consistent estimation of (S, L). Let $\widehat{\Sigma} = \frac{1}{n}\widehat{\boldsymbol{R}}^T\widehat{\boldsymbol{R}}$ be the empirical estimator of Σ, and consider the following constrained maximum likelihood estimation problem:

$$\begin{aligned}(\widehat{S}, \widehat{L}) = &\underset{(S,L)}{\operatorname{argmax}} \log\det(S+L) - \operatorname{tr}\left(\widehat{\Sigma}(S+L)\right)\\ s.t. \quad &\begin{cases} L+S \succ 0, \quad L \succeq 0,\\ \|L\|_0 \le l_0, \quad \|S\|_0 \le s_0,\\ \operatorname{rank}(L) \le r. \end{cases}\end{aligned} \tag{5}$$

Note that this optimization problem is non-convex due to the use of the $\|\cdot\|_0$ norm and the rank constraint, thus computationally challenging. To overcome this problem, Tan et al. (2023), Richard et al. (2012), and Oymak et al. (2015) suggest relaxing the constraint in (5) for L to a combined penalty function incorporating both the ℓ_1 norm and the nuclear norm $\|\cdot\|_*$. However, imposing only $\|L\|_0 \le l_0$ (or an ℓ_1, SCAD, or MCP penalty) does not yield a diagonal-block solution. To encourage such a structure, we propose the following adaptive ℓ_1-penalized estimation problem:

$$\begin{aligned}(\widehat{S}, \widehat{L}) = &\underset{(S,L)}{\operatorname{argmin}}\, \ell_n(S, L) + \mathbf{Pen}(S, L)\\ &s.t. \quad L+S \succ 0, \quad L \succeq 0,\end{aligned} \tag{6}$$

where $\ell_n(S, L) = -\log\det(S+L) + \operatorname{tr}\left(\widehat{\Sigma}(S+L)\right)$ and

$$\mathbf{Pen}(S, L) = \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\sum_{i,j=1}^{p} w_{ij}|L_{ij}|, \tag{7}$$

and w_{ij} are pre-specified data-driven weights. Different from the existing methods (Tan et al., 2023; Richard et al., 2012), we use an adaptive ℓ_1 penalty form introduced by Zou (2006) and Huang et al. (2008), upon which we conduct the consistency analysis for model selection. Intuitively, if the weight estimate w_{ij} closely approximates the inverse of the true value L^*_{ij}, then applying small penalties to the diagonal-block elements of L^* and large (potentially infinite) penalties to the off-diagonal-block elements will yield an estimate with a diagonal-block structure. We will discuss this further in Section 4.2.

In the third stage, to uncover the membership labels, we apply the K-means clustering algorithm (Hastie et al., 2009) to $\widehat{L}$. Specifically, let $\{\hat{l}_i\}_{i=1}^p$ represent the row vectors of $\widehat{L}$, and then we seek to minimize the following loss function:

$$\left\{\widehat{\mathcal{C}}_i\right\}_{i=1}^m = \underset{\mathcal{C}_1,\cdots,\mathcal{C}_m}{\operatorname{argmin}} \sum_{i=1}^m \sum_{k,j\in\mathcal{C}_i} \|\hat{l}_k - \hat{l}_j\|^2, \tag{8}$$

where $(\mathcal{C}_1,\cdots,\mathcal{C}_m)$ represents a partition of $[p]$, and $\|\cdot\|$ here denotes the Euclidean norm of vectors. In addition, when $m = r$ can be considered as a prior, we also apply the K-means clustering to a specifically transformed counterpart of $\widehat{L}$. For instance, we define the correlation matrix associated with $\{\hat{l}_i\}$ as $\mathrm{Cor}(\mathrm{abs}(\widehat{L})) \triangleq (\mathrm{cor}(\mathrm{abs}(\hat{l}_i), \mathrm{abs}(\hat{l}_j)))_{p\times p}$, where $\mathrm{cor}(\mathrm{abs}(\hat{l}_i), \mathrm{abs}(\hat{l}_j))$ is the Pearson correlation between the absolute values of $\hat{l}_i$ and $\hat{l}_j$.

Note that here we do not use the spectral clustering approach (Lei and Rinaldo, 2015; Jin, 2015) due to its dependence on the accuracy of rank estimation in the second stage. Also, in what follows, we assume the number of communities (clusters) m is known a priori in the third stage.

3.2 Algorithm

Since $\widehat{B}$ is computationally efficient due to its closed form and equation (8) has well-established algorithms for solving it (Hastie et al., 2009), our main focus here is to develop an efficient alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) to solve the convex optimization problem in (6). Since the global convergence of ADMM with three blocks (or more) of variables is ambiguous in general (Ma et al., 2013), we rewrite the problem (6) as a convex minimization problem with two blocks of variables and two separable functions as follows (Richard et al., 2012; Ma et al., 2013; Tan et al., 2023). This form can be viewed as a special case of the consensus problem discussed in Boyd et al. (2011), whose global convergence has been well studied. Specifically, we rewrite the problem in (6) as

$$\begin{aligned} (\widehat{\Theta}, \widehat{S}, \widehat{L}_1, \widehat{L}_2) = & \underset{(\Theta,S,L_1,L_2)}{\operatorname{argmin}} \ \ell_n(\Theta) + \mathbf{Pen}_1(S, L_1, L_2), \\ s.t. \quad & \begin{cases} \Theta = L_1 + S, \quad L_1 = L_2, \\ \Theta \succ 0, \quad L_1, L_2 \succeq 0, \end{cases} \end{aligned} \tag{9}$$

where $\ell_n(\Theta) = -\log\det(\Theta) + \mathrm{tr}\left(\widehat{\Sigma}\Theta\right)$ and $\mathbf{Pen}_1(S, L_1, L_2)$ is defined as follows

$$\mathbf{Pen}_1(S, L_1, L_2) = \gamma_n \|S\|_{\mathrm{off},1} + \delta_n \|L_1\|_* + \tau_n \sum_{i,j=1}^p w_{ij}|L_{2,ij}|.$$

Let $Y_1 = (\Theta, S, L_1, L_2) \in \mathbb{R}^{p\times 4p}$, $Y_2 = (\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1, \widetilde{L}_2) \in \mathbb{R}^{p\times 4p}$. Then the optimization problem in (9) is equivalent to

$$\begin{aligned} \min \quad & f(Y_1) + \psi(Y_2) \\ s.t. \quad & Y_1 - Y_2 = 0, \end{aligned} \tag{10}$$

Algorithm 1 ADMM Algorithm for Solving (10).

Input: The residual $\widehat{\mathbf{R}}$ computed in the first stage. Tuning parameters $(\mu, \gamma_n, \delta_n, \tau_n) > 0$. Suitable weights $w_{ij} > 0$. Stopping error $\varepsilon > 0$.

Initialization:

Let primal variables S^1, $L_1^1 = L_2^1$ all be identity matrix, and $\Theta^1 = S^1 + L_1^1$.

Set $Y_1^1 = Y_2^1 = (\Theta^1, S^1, L_1^1, L_2^1)$.

Let dual variables Γ_0 be zero matrix.

while the stopping criterion $\|Y_1^{k+1} - Y_1^k\|_F^2 / \|Y_1^k\|_F^2 \le \varepsilon$ is not met **do**

For the k-th iteration, denote $Y_2^k = (\widetilde{\Theta}^k, \widetilde{S}^k, \widetilde{L}_1^k, \widetilde{L}_2^k)$, $\Gamma^k = (\Gamma_\Theta^k, \Gamma_S^k, \Gamma_{L_1}^k, \Gamma_{L_2}^k)$.

Update Θ^{k+1}:

$$\Theta^{k+1} = U \operatorname{diag}(\xi) U^T, \quad \xi_i = \frac{\sigma_i + \sqrt{\sigma_i^2 + 4\mu}}{2},$$

where $U \operatorname{diag}(\sigma) U^T$is the eigendecomposition for $\widetilde{\Theta}^k - \mu\widehat{\Sigma} - \Gamma_\Theta^k$.

Update S^{k+1}:

$$\left(S^{k+1}\right)_{ij} = \begin{cases} \operatorname{prox}_{\mu\gamma_n, \|\cdot\|_1}\left(\widetilde{S}_{ij}^k - (\Gamma_S^k)_{ij}\right), i \neq j \\ \widetilde{S}_{ij}^k - \left(\Gamma_S^k\right)_{ij}, \quad i = j. \end{cases}$$

Update L_1^{k+1}:

$$L_1^{k+1} = \mathcal{P}_{S^+}\left(\left(\operatorname{prox}_{\mu\tau_n w_{ij}, \|\cdot\|_1}\left((\widetilde{L}_1^k)_{ij} - (\Gamma_{L_1}^k)_{ij}\right)\right)_{(i,j)\in[p]\times[p]}\right),$$

where S^+ denotes the positive semidefinite cone in $\mathbb{S}^{p\times p}$.

Update L_2^{k+1}:

$$L_2^{k+1} = \operatorname{prox}_{\mu\delta_n, \|\cdot\|_*}(\widetilde{L}_2^k - \Gamma_{L_2}^k).$$

Let $Y_1^{k+1} = (\Theta^{k+1}, S^{k+1}, L_1^{k+1}, L_2^{k+1})$ and partition the matrix $T^k \triangleq Y_1^{k+1} + \Gamma^k = (T_\Theta^k, T_S^k, T_{L_1}^k, T_{L_2}^k)$ into four blocks in the same form as Y_2^k.

Update $(\widetilde{\Theta}^{k+1}, \widetilde{S}^{k+1}, \widetilde{L}_1^{k+1}, \widetilde{L}_2^{k+1})$ (see Appendix A.2 for detailed calculation):

$$\begin{cases} \widetilde{\Theta}^{k+1} = \frac{3}{5}T_\Theta^k + \frac{2}{5}T_S^k + \frac{1}{5}T_{L_1}^k + \frac{1}{5}T_{L_2}^k \\ \widetilde{S}^{k+1} = \frac{2}{5}T_\Theta^k + \frac{3}{5}T_S^k - \frac{1}{5}T_{L_1}^k - \frac{1}{5}T_{L_2}^k \\ \widetilde{L}_1^{k+1} = \frac{1}{5}T_\Theta^k - \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \\ \widetilde{L}_2^{k+1} = \frac{1}{5}T_\Theta^k - \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \end{cases}.$$

Denote $Y_2^{k+1} = (\widetilde{\Theta}^{k+1}, \widetilde{S}^{k+1}, \widetilde{L}_1^{k+1}, \widetilde{L}_2^{k+1})$. Update Γ^{k+1}:

$$\Gamma^{k+1} = \Gamma^k + \left(Y_1^{k+1} - Y_2^{k+1}\right).$$

Update the iteration step: $k \leftarrow k + 1$.

end while

Collect the last iteration value denoted by $Y_1^{\text{final}} = (\Theta^{\text{final}}, S^{\text{final}}, L_1^{\text{final}}, L_2^{\text{final}})$.

Output: The final result $(\Theta^{\text{final}}, S^{\text{final}}, L_2^{\text{final}})$.

where

$$f(Y_1) \triangleq \ell_n(\Theta) + \mathbf{Pen}_1(S, L_1, L_2),$$

$$\psi(Y_2) \triangleq \mathcal{I}\{\widetilde{\Theta} - \widetilde{L}_1 - \widetilde{S} = 0\} + \mathcal{I}\{\widetilde{L}_1 - \widetilde{L}_2 = 0\},$$

and the indicator function is defined as

$$\mathcal{I}(X \in \mathcal{X}) = \begin{cases} 0, & \text{if } X \in \mathcal{X} \\ +\infty, & \text{otherwise} \end{cases}.$$

Define the augmented Lagrangian function associated with (10) as

$$\mathcal{L}_\mu(Y_1, Y_2, \Gamma) = f(Y_1) + \psi(Y_2) + \Gamma \cdot (Y_1 - Y_2) + \frac{1}{2\mu}\|Y_1 - Y_2\|_F^2,$$

where $\Gamma \in \mathbb{R}^{p\times 4p}$ is the Lagrange multiplier and for two $p \times q$ matrices A and B

$$A \cdot B \triangleq \sum_{i=1}^{p}\sum_{j=1}^{q} A_{ij}B_{ij} = \mathrm{Tr}\left(A^T B\right).$$

Here $\mu > 0$ is the penalty parameter. Then we divide (10) into solvable subproblems through the scaled ADMM algorithm:

$$\begin{cases} Y_1^{k+1} = \underset{Y_1}{\operatorname{argmin}}\, \mathcal{L}_\mu(Y_1, Y_2^k, \Gamma^k) = \underset{Y_1}{\operatorname{argmin}}\, f(Y_1) + \dfrac{1}{2\mu}\left\|Y_1 - Y_2^k + \Gamma^k\right\|_F^2 \\ Y_2^{k+1} = \underset{Y_2}{\operatorname{argmin}}\, \mathcal{L}_\mu(Y_1^{k+1}, Y_2, \Gamma^k) = \underset{Y_2}{\operatorname{argmin}}\, \psi(Y_2) + \dfrac{1}{2\mu}\left\|Y_1^{k+1} - Y_2 + \Gamma^k\right\|_F^2, \\ \Gamma^{k+1} = \Gamma^k + \left(Y_1^{k+1} - Y_2^{k+1}\right). \end{cases}$$

We now introduce some useful proximal mappings (see Appendix A for details) to design the algorithm. The proximal mapping for $p(X) = \|X\|_1$ is (equivalent to the soft-thresholding operator)

$$\mathrm{prox}_{\lambda,\|\cdot\|_1}(Z) \triangleq \mathrm{sign}(Z)(Z - \lambda \mathbf{1}_p \mathbf{1}_p^T)_+,$$

where $(\cdot)_+ = \max(\cdot, 0)$ and $\mathbf{1}_p = (1, \cdots, 1)^T \in \mathbb{R}^p$. The proximal mapping for $p(X) = \|X\|_*, X \succeq 0$ is

$$\mathrm{prox}_{\lambda,\|\cdot\|_*}(Z) \triangleq U\widetilde{\Sigma}U^T,$$

where $Z = U\Sigma U^T, \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ and $\widetilde{\Sigma} = \mathrm{diag}\left((\sigma_1 - \lambda)_+, \ldots, (\sigma_p - \lambda)_+\right)$. In addition, the proximal mapping for $p(X) = \mathcal{I}(X \in \mathcal{X})$ is the projection mapping $\mathcal{P}_\mathcal{X}(Z)$, denoting the projection of Z on set $\mathcal{X}$ with respect to the inner product defined above. The complete algorithm is summarized in Algorithm 1, whose detailed description is deferred to Appendix A.

3.3 Weight Assignment and Choices of Tuning Parameters

According to Huang et al. (2008), compared to the standard ℓ_1 penalty, incorporating a high penalty for zero elements and a low penalty for nonzero elements in the true values of L within the adaptive ℓ_1 penalty reduces the estimation bias and improves the accuracy of variable selection. We start by implementing the method given in Chandrasekaran et al. (2012) to determine an initial weight assignment:

$$\begin{aligned}(\bar{S}, \bar{L}) &= \underset{(\Theta,S,L)}{\operatorname{argmin}}\, \ell_n(S, L) + \lambda_n \|S\|_{\text{off},1} + \gamma_n \|L\|_* \\ s.t. \quad &\Theta = L + S, \quad \Theta \succ 0, \quad L \succeq 0.\end{aligned} \tag{11}$$

Denote $\bar{L} = (\bar{l}_{ij})_{p\times p}$, then the initial weight estimate $\widetilde{W}_n$ is set as $\tilde{w}_{ij} \triangleq \bar{l}_{ij}^a$, for $i, j = 1, \ldots, p$, and we assume $a > 0$ which is typically set as 1 or 2. Subsequently, the adaptive weight $W_n = (w_{ij})_{p\times p}$ assigned in (7) is:

$$w_{ij} = \frac{1}{\tilde{w}_{ij}}. \tag{12}$$

Moreover, we rely on the cross-validation (CV) method to choose tuning parameters, which in general is more accurate than using BIC but with higher computational costs. We randomly split the rows of residual $\widehat{\boldsymbol{R}}$ into T segments of equal size, and denote the sample covariance matrix using the data in the t-th segment ($t = 1, ..., T$) by $\widehat{\Sigma}^t$. Let $(\widehat{S}_\Xi^{-t}, \widehat{L}_\Xi^{-t})$ be the estimator obtained by solving (6) but using all data excluding the t-th segment with the tuning parameters $\Xi \equiv (\gamma_n, \delta_n, \tau_n)$. We then choose $\widehat{\Xi}$ that minimizes the average predictive negative log-likelihood:

$$\text{CV}(\Xi) = \sum_{t=1}^{T} \left[\text{tr}\left(\widehat{\Sigma}^t (\widehat{S}_\Xi^{-t} + \widehat{L}_\Xi^{-t}) \right) - \log\left(\det(\widehat{S}_\Xi^{-t} + \widehat{L}_\Xi^{-t}) \right) \right],$$

i.e. we set tuning parameters as $\widehat{\Xi} = \text{argmin}_{\Xi \in \mathbb{R}_+^3} \text{CV}(\Xi)$.

4. Theoretical Properties

In this section, we will discuss the identifiability issues in the model (1) and provide theoretical guarantees for the three-stage estimation procedure. Throughout the discussion, we will always denote the true parameters by (B^*, S^*, L^*).

4.1 Identifiability

The identifiability for B^* can be ensured by assuming that the fixed design matrix $\widetilde{\boldsymbol{C}}$ is of column full rank. We mainly present conditions for identifiability of (S^*, L^*) by resorting to the tools developed by Chandrasekaran et al. (2012). To proceed, we will describe two submanifolds of $\mathbb{S}^{p\times p}$ and their tangent spaces.

Denote $s_0 = \|\mathbf{O}(S^*)\|_0$ with $\mathbf{O}(S) \triangleq S - \text{diag}(S)$. Consider the set of sparse matrices with s_0 nonzero elements except the diagonal, let

$$\mathcal{S}(s_0) = \left\{ M \in \mathbb{S}^{p\times p} : \|\,\text{Supp}(\mathbf{O}(M))\|_0 \le s_0 \right\},$$

It is a subspace of $\mathbb{S}^{p\times p}$ and thus can also be viewed as a submanifold of $\mathbb{S}^{p\times p}$. The tangent space at any smooth point $S \in \mathcal{S}(s_0)$ around S^* is all given by

$$\Omega(S) = \{N : \mathrm{Supp}(\mathbf{O}(N)) \subseteq \mathrm{Supp}(\mathbf{O}(S^*))\}.$$

To describe the tangent space of the low-rank diagonal blocks L up to row and column permutations near L^*, we make the first assumption for L^*.

Assumption 1 *Denote the eigendecomposition of L^* by $L^* = U_1^*\Sigma_1^*U_1^{*T}$, where $\Sigma_1^* = \mathrm{diag}(\sigma_1^*, \cdots, \sigma_r^*)$, $U_1^* \in \mathbb{R}^{p\times r}, U_1^{*T}U_1^* = I_r$. Assume that*

(1) (Eigengap) $\sigma_1^* > \cdots > \sigma_r^* > 0$.

(2) (Diagonal-block) L^* has a diagonal-block structure up to column and row permutations. Without loss of generality, let $L^* = \mathrm{diag}(L_1^*, \cdots, L_m^*), L_i^* \in \mathbb{S}^{d_i\times d_i}, \sum_{i=1}^m d_i = p$ and $\forall i \in [m]$, $L_i^* \neq 0$ in the entry-wise sense. Besides, denote $\mathrm{rank}(L_i^*) = r_i$, $\sum_{i=1}^m r_i = r$.

We make some remarks on this assumption. Part (1) is the eigengap assumption commonly imposed in the literature of matrix perturbation (Yu et al., 2014; Fan et al., 2018), which is necessary to identify the eigenspace of L^*. Part (2) states that L^* has m diagonal blocks up to row and column permutations with each block being of low-rank. To fix the idea, we directly assume that L^* is diagonal and this condition emphasizes that the sparsity of the truth L^* only comes from off-diagonal blocks. To clarify this point, define $G_{ij}^* = \{(k,l)|k \in \{d_{i-1}+1, \cdots, d_{i-1}+d_i\}, l \in \{r_{j-1}+1, \cdots, r_{j-1}+r_j\}\}$ for $i, j \in [m]$ and define $d_0 = 0, r_0 = 0$. Then $\{G_{ij}^*\}_{(i,j)\in[m]\times[m]}$ forms a partition of $[p]\times[r]$ and $(U_1^*)_{G_{ij}^*} = \mathbf{0}_{d_i\times r_j}$ for $i \neq j \in [m]$. It is easy to see that when $r = m$, part (2) will be automatically satisfied, which means that each block has at most one dimension, i.e., each row of U_1^* has at most one nonzero element. Essentially, part (2) requires the sparsity pattern of the eigenspace to explain the zero off-diagonal blocks of L^*, while allowing the eigenspace within each diagonal block to be as dense as possible. Furthermore, in the context of this paper, different diagonal blocks represented by L_i signify distinct communities (clusters). To more accurately identify the members of each community, we define the true labels of different communities as $\mathcal{C}_i^* = \{d_{i-1}+1, \cdots, d_{i-1}+d_i\}, i = 1, \cdots, m$ and $\{\mathcal{C}_i^*\}_{i=1}^m$ forms a partition of $[p]$.

Now we consider the set of diagonal-block matrices with the rank of each block less than r_i. Let

$$\mathcal{LS}(m,r) = \left\{L = \mathrm{diag}(L_1, \cdots, L_m) : \mathrm{rank}(L_i) \leq r_i, L_i \in \mathbb{S}^{d_i\times d_i}\right\}.$$

When $m = 1$, this set is the classical algebraic variety of low-rank matrices $\mathcal{L}(r_i) \triangleq \mathcal{LS}(1, r_i)$ with tangent space $\mathcal{T}_1(L_i) = \{U_{1i}Y_{1i}^T + Y_{1i}U_{1i}^T : Y_{1i} \in \mathbb{R}^{d_i\times r_i}\}$, where U_{1i} is the r_i leading eigenvectors of L_i (Chandrasekaran et al., 2011) (when we say "leading eigenvectors' we are comparing the magnitudes of the eigenvalues by neglecting the signs). Under Assumption 1, all the smooth points $L \in \mathcal{LS}(m,r)$ form a submanifold of $\mathbb{S}^{p\times p}$ satisfying $\mathrm{rank}(L_i) = \mathrm{rank}(L_i^*) = r_i$, and $\mathrm{Supp}(L_i) = \mathrm{Supp}(L_i^*)$, which is a simple application of property of product manifold (details see Appendix B.1). The tangent space at any smooth point of this submanifold can be viewed as an intersection of two spaces. We summarize this property in the following proposition.

Proposition 2 *The tangent space at any smooth point $L = U_1 \Sigma U_1^T \in \mathcal{LS}(m, r)$ is given by*

$$\mathcal{T}(L) = \left\{ N = \mathrm{diag}(N_1, \cdots, N_m) : N_i \in \mathbb{S}^{d_i \times d_i}, N_i \in \mathcal{T}_1(L_i) \right\} = \mathcal{T}_1(L) \cap \mathcal{T}_2(L^*),$$

where $\mathcal{T}_1(L) = \left\{ U_1 Y^T + Y U_1^T : Y \in \mathbb{R}^{p \times r} \right\}$, $\mathcal{T}_2(L^*) = \{ N = \mathrm{diag}(N_1, \cdots, N_m) : \mathrm{Supp}(N_i) \subseteq \mathrm{Supp}(L_i^*) \}$. *Moreover,*

$$\mathcal{T}^\perp(L) = \mathcal{T}_1^\perp(L) + \mathcal{T}_2^\perp(L^*) = \mathcal{V}(L) \oplus \mathcal{T}_2^\perp(L^*),$$

with $\mathcal{V}(L) = \left\{ N = \mathrm{diag}(N_1, \cdots, N_m) : N_i \in \mathbb{S}^{d_i \times d_i}, N_i \in \mathcal{T}_1^\perp(L_i) \right\}$.

Denote the spaces evaluated at the true value by $\Omega(S^*) = \Omega^*$, $\mathcal{V}(L^*) = \mathcal{V}^*$, $\mathcal{T}_2(L^*) = \mathcal{T}_2^*$, $\mathcal{T}(L^*) = \mathcal{T}^*$ and $\mathcal{T}_2^\perp(L^*) = \mathcal{T}_2^{*\perp}$. In addition, for each i, let $S_i^* \triangleq S^*_{\mathcal{C}_i^* \times \mathcal{C}_i^*}$ be the submatrix in the diagonal of S^* exhibiting the same block size as L_i^*. For identifiability, we need the following transversality condition (Chandrasekaran et al., 2012; Chen et al., 2016; Zhao et al., 2024; Hsu et al., 2011).

Assumption 3 *The intersection of the tangent spaces evaluated at S^* and L^* is trivial, i.e., $\Omega^* \cap \mathcal{T}^* = \mathbf{0}_{p \times p}$.*

According to Proposition 2, this assumption is equivalent to $\Omega(S_i^*) \cap \mathcal{T}_1(L_i^*) = \mathbf{0}_{d_i \times d_i}$ for $i = 1, \ldots, m$, which means that the identifiability of (S^*, L^*) can be reduced to that of (S_i^*, L_i^*) within each diagonal block.

Theorem 4 *Under Assumptions 1 (2) and 3, (S^*, L^*) is locally identifiable with respect to $\mathcal{S}(s_0) \times \mathcal{LS}(m, r)$. Moreover, if Assumption 1 (1) also holds, then U_1^* is also locally identifiable up to sign flips.*

Remark 5 *For convenience, let $\mathcal{E}(r)$ be the set of $r \times r$ sign flip matrices, i.e., for any $M \in \mathbb{R}^{p \times r}$ and $P \in \mathcal{E}(r)$, and MP is a matrix whose k-th column equals to the k-th column of M multiplied by 1 or -1. Theorem 4 claims that there exists a neighborhood $\mathcal{N}(S^*, L^*)$ of (S^*, L^*) in $\mathcal{S}(s_0) \times \mathcal{LS}(m, r)$ such that if it holds that $S + L = S^* + L^*$ for $(S, L) \in \mathcal{N}(S^*, L^*)$, then $S = S^*, L_i = L_i^*, U_{1i} = U_{1i}^* P_i$ for some $P_i \in \mathcal{E}(r_i)$, where $L = \mathrm{diag}(L_1, \cdots, L_m)$ and U_{1i}, U_{1i}^* are the r_i leading eigenvectors of L_i, L_i^* respectively.*

Remark 6 *The model is identifiable, at least with respect to $\mathcal{S}(s_0) \times \mathcal{LS}(m, r)$ in the neighborhood of (S^*, L^*) defined by $\left\{ (S, L) : \mathrm{Supp}(S) = \mathrm{Supp}(S^*), L = \mathrm{diag}(L_1, \cdots, L_m), \mathrm{rank}(L_i) = r_i, L_i = U_{1i}^* Y^T + Y U_{1i}^{*T}, Y \in \mathbb{R}^{p \times r_i} \right\}$. Appendix C examines the effect of varying initial values and shows that the proposed algorithm consistently converges to a neighborhood of the ground truth, suggesting that the identifiable region around the true model is sufficiently large.*

4.2 Asymptotic Analysis

In this section, we provide some theoretical guarantees for the proposed three-stage estimation procedure for $(\widehat{B}, \widehat{S}, \widehat{L})$ and $\{\widehat{\mathcal{C}}_i\}_{i=1}^m$ in the asymptotic regime, where we fix p, q and let $n \to \infty$ only. For $\widehat{B}$ in the first stage, we make the following commonly used assumption, which ensures that $\widehat{B}$ is $\sqrt{n}$-consistent for B^* from the classical linear regression theory.

Assumption 7 *There exists a positive definite matrix C^* such that $\frac{1}{n}\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} \to C^*$ as $n \to \infty$.*

When it comes to $\delta\|L\|_* + \|L\|_1$ type penalty in the second stage, the existing studies mainly focus on the estimation error problem (Tan et al., 2023; Richard et al., 2012) or signal processing problem (Oymak et al., 2015). Next, we establish model-selection consistency and derive a convergence rate under the adaptive penalty (7).

To fix ideas, we temporarily ignore S and focus on L. Let $\tilde{\ell}(L) = \ell_n(S, L)$. According to the optimality condition (Boyd and Vandenberghe, 2004) to the problem (6), we want to solve the following equation for smooth points $L \in \mathcal{LS}(m, r)$ with $L \succeq 0$,

$$-\nabla\tilde{\ell}(L) \in \partial(\delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1),$$

where $\odot$ is the Hadamard product. It is equivalent to solving the equations (see Lemma 24 in Appendix B.2.1 for the expression of this subgradient)

$$\begin{aligned}\mathcal{P}_{\mathcal{T}(L)}(-\nabla\tilde{\ell}(L)) &= \delta_n UU^T + \tau_n\mathcal{P}_{\mathcal{T}(L)}(W_n \odot \operatorname{sign}(L^*)),\\ \mathcal{P}_{\mathcal{T}^\perp(L)}(-\nabla\tilde{\ell}(L)) &= \tau_n\mathcal{P}_{\mathcal{T}^\perp(L)}(W_n \odot \operatorname{sign}(L^*)) + \delta_n F_1 + \tau_n W_n \odot F_2,\end{aligned}$$

where $F_1 \in \mathcal{T}_1^\perp(L), F_2 \in \mathcal{T}_2^{*\perp}, \|F_1\|_\infty \le 1, \|F_2\|_2 \le 1$. Second, to solve these two equations, we resort to the primal-dual witness technique (Ravikumar et al., 2011; Chandrasekaran et al., 2012; Chen et al., 2016), i.e., we obtain a solution $\tilde{L}$ from the first equation. Then we substitute this solution into the second equation to construct suitable dual elements F_1 and F_2. Finally, we establish the uniqueness of this solution. This procedure encounters two primary challenges. The first one is related to the extra terms $\tau_n\mathcal{P}_{\mathcal{T}(L)}(W_n \odot \operatorname{sign}(L^*))$ and $\tau_n\mathcal{P}_{\mathcal{T}^\perp(L)}(W_n \odot \operatorname{sign}(L^*))$, which can affect the existence of the solution. The second difficulty lies in identifying dual elements F_1 and F_2 that satisfy the second equation through some suitable norm. Unlike the existing literature that mainly focuses on the decomposable penalty function (Negahban et al., 2012; Candes and Recht, 2013), the penalty involved in this problem does not satisfy the decomposable condition, making it challenging to directly use the dual norm of this penalty to identify the dual elements.

The first difficulty can be solved by appropriate weight selection, and the extra terms will vanish with rate γ_n in probability. In fact, we need the weight matrix W_n in (7) to satisfy the following assumption (Zou, 2006; Huang et al., 2008):

Assumption 8 *Let J_1 be the support set of L^*. The initial estimator $\widetilde{W}_n = (\tilde{w}_{ij})_{p\times p}$ defined in* (12) *is r_n-consistent*

$$r_n \max_{i,j}\left|(\widetilde{W}_n)_{ij} - (\Lambda_n)_{ij}\right| = O_P(1), \quad r_n \to \infty,$$

where Λ_n are unknown nonrandom constants satisfying

$$\max_{(i,j)\notin J_1}|(\Lambda_n)_{ij}| \le M_{n2} = O(\frac{1}{r_n^{1-\alpha}}), \quad \max_{(i,j)\in J_1}\frac{1}{(\Lambda_n)_{ij}} \le M_{n1} = o(r_n^{1-\alpha}),$$

for some $\alpha \in (0, 1)$.

Remark 9 *This assumption follows the adaptive-weighting framework of Huang et al. (2008). It assumes that the initial estimator $(\widetilde{W}_n)_{ij}$ actually estimates some proxy $(\Lambda_n)_{ij}$ of $(L_n^*)_{ij}$, so that the weight $(W_n)_{ij} = |(\widetilde{W}_n)_{ij}|^{-1}$ is not too large for $(L_n^*)_{ij} \neq 0$ and not too small for $(L_n^*)_{ij} = 0$. According to Theorem 4.1 in Chandrasekaran et al. (2012), the choice of weights in section 3.3 satisfies this assumption by taking $r_n = \sqrt{n}$ for fixed p and $\Lambda_n = L^*$. By reparameterization, we first denote $\delta_n = \rho_1\gamma_n, \tau_n = \frac{\rho_2\gamma_n}{r_n^{1-\alpha}}$. Then according to Lemma 21 in Appendix B.2.1, we find $\tau_n\mathcal{P}_{\mathcal{T}(L)}(W_n \odot \operatorname{sign}(L^*))$ and $\tau_n\mathcal{P}_{\mathcal{T}^\perp(L)}(W_n \odot \operatorname{sign}(L^*))$ are all $o_P(\gamma_n)$.*

To overcome the second difficulty, we construct the following useful norm, which is only defined on $\Omega^{*\perp} \times \mathcal{T}^\perp(L)$ with $\mathcal{T}^\perp(L) = \mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp}$,

$$h_{\rho_1,\rho_2,L,\widetilde{W}_n}(S, L') = \max\left\{\|S\|_\infty, \frac{\|L_1\|_2}{\rho_1}, \frac{r_n^{1-\alpha}\|\widetilde{W}_n \odot L_2\|_\infty}{\rho_2}\right\},$$

where $L' = L_1 \oplus L_2$ is the orthogonal direct sum decomposition of L' restricted to $\mathcal{V}(L)\oplus\mathcal{T}_2^{*\perp}$, i.e., $L_1 = \mathcal{P}_{\mathcal{V}(L)}(L'), L_2 = \mathcal{P}_{\mathcal{T}_2^{*\perp}}(L')$. This norm is well-defined for fixed L due to the uniqueness of the decomposition and $\mathcal{V}(L)$ has an invariant dimension for any smooth point $L \in \mathcal{LS}(m, r)$ according to Proposition 2. It can be verified that with probability tending to 1, combining $\tau_n\mathcal{P}_{\mathcal{T}^\perp(L)}(W_n \odot \operatorname{sign}(L^*)) = o_P(\gamma_n)$ with

$$h_{\rho_1,\rho_2,L,\Lambda_n}(0, \mathcal{P}_{\mathcal{T}^\perp(L)}(-\nabla\tilde{\ell}(L))) < \gamma_n$$

imply the existence of dual elements F_1, F_2 (see Lemma 24 in Appendix B.2.1).

After completing the above preparations, we present the adaptive irrepresentability condition that is important for the consistency of the adaptive ℓ_1 penalized estimator defined in equation (6). The core quantity involved here is the positive definite Fisher information matrix evaluated at the true parameter, i.e.,

$$\mathcal{I}^* \triangleq \mathcal{I}(\Theta^*) = \Theta^{*-1} \otimes \Theta^{*-1} = (S^* + L^*)^{-1} \otimes (S^* + L^*)^{-1},$$

where $\otimes$ represents the matrix Kronecker product and $\mathcal{I}^* \in \mathbb{R}^{p^2\times p^2}$. We also treat $\mathcal{I}^*$ as a tensor product operator, i.e., $\mathcal{I}^* : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$. To simplify notation, we use $\mathcal{I}^*(M)$ and $\mathcal{I}^*v(M)$ interchangeably, with $\mathcal{I}^*v(M) = v(\mathcal{I}^*(M))$. Define two linear operators $\mathbf{G} : \Omega^* \times \mathcal{T}^* \to \Omega^* \times \mathcal{T}^*$ and $\mathbf{G}^\perp : \Omega^* \times \mathcal{T}^* \to \Omega^{*\perp} \times \mathcal{T}^{*\perp}$,

$$\mathbf{G}(S, L) = \left(\mathcal{P}_{\Omega^*}\left\{\mathcal{I}^*(S + L)\right\}, \mathcal{P}_{\mathcal{T}^*}\left\{\mathcal{I}^*(S + L)\right\}\right),$$

$$\mathbf{G}^\perp\left(S', L'\right) = \left(\mathcal{P}_{\Omega^{*\perp}}\left\{\mathcal{I}^*\left(S' + L'\right)\right\}, \mathcal{P}_{\mathcal{T}^{*\perp}}\left\{\mathcal{I}^*\left(S' + L'\right)\right\}\right).$$

Assumption 10 *The adaptive irrepresentability condition is, as $n \to \infty$,*

$$h_{\rho_1,\rho_2,L^*,\Lambda_n}\left(\mathbf{G}^\perp\mathbf{G}^{-1}(\operatorname{sign}(\mathbf{O}(S^*)), \rho_1 U_1^* U_1^{*T})\right) < 1,$$

where recall $\mathbf{O}(S^*) = S^* - \operatorname{diag}(S^*)$. The invertibility of $\mathbf{G}$ is ensured by Lemma 25 in Appendix B.2.1.

Remark 11 *The adaptive irrepresentability assumption generalizes the strong representability condition proposed by Zhao and Yu (2006) and Ravikumar et al. (2011) for the pursuit of not only rank consistency but also diagonal-block structure consistency of L^*. If $\mathrm{sign}(\Lambda_n) = \mathrm{sign}(L_n^*)$, we say that the initial estimates are zero-consistent with rate r_n. In this case, $M_{n2} = 0$ and Assumption 10 reduces to assumption A4 in Chen et al. (2016).*

Remark 12 *Following Chandrasekaran et al. (2012), in Appendix D we provide sufficient conditions for the existence of ρ_1 and ρ_2 in Assumption 10 (see Theorem 33). In particular, such ρ_1 and ρ_2 typically exist when each low-rank submatrix L_i^* corresponding to a diagonal block of L^* is nearly maximally incoherent and the graph S^* has bounded degree (Chandrasekaran et al., 2012).*

Now, we state the main theorem about the adaptive ℓ_1 penalized estimator in the second stage.

Theorem 13 *Under Assumptions 1-10, let tuning parameter $\gamma_n \asymp n^{-\frac{1}{2}+\eta}$ for some $\eta > 0$, and $\delta_n = \rho_1\gamma_n, \tau_n = \frac{\rho_2\gamma_n}{r_n^{1-\alpha}}$ with ρ_1, ρ_2 satisfying Assumption 10. Then, with probability tending to one, the optimization problem (6) obtains a unique solution $(\widehat{S}, \widehat{L})$ that is both estimation consistent and selection consistent in the following sense,*

$$\|\widehat{S} - S^*\|_\infty \lesssim_P n^{-\frac{1}{2}+2\eta}, \quad \|\widehat{L} - L^*\|_\infty \lesssim_P n^{-\frac{1}{2}+2\eta},$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\mathrm{sign}(\widehat{S}) = \mathrm{sign}(S^*)\right) = 1,$$

$$\lim_{n\to\infty} \mathbb{P}\left(\mathrm{sign}(\widehat{L}) = \mathrm{sign}\left(L^*\right), \mathrm{rank}(\widehat{L}) = \mathrm{rank}\left(L^*\right)\right) = 1.$$

Theorem 13 shows that the adaptive ℓ_1 penalized estimator in the second stage achieves both the rank consistency and diagonal-block structure consistency for the non-overlapping community part L^*, and sparsity consistency for the sparse structure S^* with high probability. Note that Theorem 13 does not explicitly require the sparsity of S^*, and when cross-community variables are dependent, it is possible that the proposed model remains valid under the identifiability (Assumption 3) and irrepresentability conditions (Assumption 10). From the proof of Theorem 13, we can also get some consistency results for the r leading eigenvectors $\widehat{U}_1$ of $\widehat{L}$.

Theorem 14 *Under the same assumptions as those in Theorem 13, there exists $P_n \in \mathcal{E}(r)$ such that*

$$\|\widehat{U}_1 P_n - U_1^*\|_\infty \lesssim_P n^{-\frac{1}{2}+\eta}, \quad \lim_{n\to\infty} \mathbb{P}\left(\left(\widehat{U}_1 P_n\right)_{G_{ij}^*} = 0, \forall i \neq j\right) = 1.$$

Remark 15 *This theorem states that with high probability we can identify all the zero blocks $(U_1^*)_{G_{ij}^*}, i \neq j$ in the eigenspace corresponding to the off-diagonal zero blocks of L^*. In contrast, the sparsity within the eigenspace $(U_1^*)_{G_{ii}^*}$ corresponding to the diagonal blocks of L^* cannot be identified in general. Note that this theorem is stronger than Theorem 13, it shows that we can even obtain the rank consistency for each diagonal block $L_i, i \in [m]$.*

When $m = r$, the sparsity in the eigenspace U_1^* is equivalent to the sparsity in L^*. This is a direct corollary of Theorem 14.

Corollary 16 *Suppose $m = r$, then under Assumptions 1 - 10, there exists $P_n \in \mathcal{E}(r)$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\operatorname{sign}(\widehat{U}_1 P_n) = \operatorname{sign}(U_1^*)\right) = 1,$$

for the r leading eigenvectors $\widehat{U}_1$ of $\widehat{L}$.

In the third stage, we performed K-means clustering on the (transformed) row vectors of $\widehat{L}$. Recall that $[p] = \mathcal{C}_1^* \cup \cdots \cup \mathcal{C}_m^*$ is the true community partition. We introduce the $p \times 1$ vector $\boldsymbol{t}$ of true labels such that

$$\boldsymbol{t}(i) = k \quad \text{if and only if} \quad i \in \mathcal{C}_k^*, \quad 1 \le i \le p, 1 \le k \le m.$$

For the clustering procedure (8), there is a (disjoint) partition $[p] = \widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_m$ with estimated labels

$$\hat{\boldsymbol{t}}(i) = k \quad \text{if and only if} \quad i \in \widehat{\mathcal{C}}_k, \quad 1 \le i \le p, 1 \le k \le m.$$

Then the proportion of mis-clustered nodes (also called Hamming error rate) which has been previously used as a loss function for community detection in blockmodels (Ma et al., 2020; Jin, 2015) is defined as

$$H(\hat{\boldsymbol{t}}, \pi(\boldsymbol{t})) \triangleq \operatorname*{argmin}_{\pi \in S_m} \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}(\hat{\boldsymbol{t}}(i) \neq \boldsymbol{t}(\pi(i))),$$

where $S_m = \{\pi : \pi \text{ is a permutation of the set } [m]\}$. The indicator function $\mathbb{1}(x \in \mathcal{X})$ equals 1 if $x \in \mathcal{X}$, and equals 0 otherwise.

Theorem 17 *Denote the rows of L^* by $\{l_i^*\}_{i=1}^p$, and define $\mu_j^* = \frac{1}{d_j} \sum_{i:\boldsymbol{t}(i)=j} l_i^*$ as the center of the j-th community. Let $c \triangleq \min_{j \neq j'} \frac{1}{2}\|\mu_j^* - \mu_{j'}^*\| > 0$ and $d_j \in [\frac{p}{\omega m}, \frac{\omega p}{m}], j = 1, \cdots, m$. Then for any $\widehat{L}$ and the respective estimated labels $\hat{\boldsymbol{t}}$, we have*

$$H(\hat{\boldsymbol{t}}, \pi(\boldsymbol{t})) \le \frac{C\omega}{pc^2} \left(\|\widehat{L} - L^*\|_F^2 + \sum_{i=1}^{p} \|l_i^* - \mu_{\boldsymbol{t}(i)}^*\|^2 \right),$$

where $C > 0$ is an absolute constant and $\omega \ge 1$ regularizes community size variability.

Remark 18 *This bound for Hamming error rate is similar to Proposition 19 in Ma et al. (2020) involving two components. The latter component does not depend on data and hence can be viewed as an oracle clustering error term that originates from the true latent cluster only. In contrast, the former component depends on the estimation error of L and hence can be viewed as a noise-induced error term. Both terms are dependent on c and ω, which respectively quantify the degree of separation between the true cluster centers and the balance between communities. Under the conditions of Theorem 13, we can further obtain that $H(\hat{\boldsymbol{t}}, \pi(\boldsymbol{t})) \le \frac{C\omega}{pc^2} \sum_{i=1}^{p} \|l_i^* - \mu_{\boldsymbol{t}(i)}^*\|^2 + O_P(n^{-1+4\eta})$. When $\{l_i^*\}$ has small variations within the same community, the Hamming error rate will be zero exactly with high probability.*

In general, the oracle clustering error term cannot be eliminated. However, when each community has one dimension, performing K-means on $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$ will be more suitable than on $\widehat{L}$ directly. Denote the labels estimated from the clustering procedure (8) based on $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$ by $\tilde{\boldsymbol{t}}$. Define $\mathrm{cor}(\mathrm{abs}(x), \mathrm{abs}(y)) = \frac{\sum_{i=1}^{p}(|x_i|-\bar{x})(|y_i|-\bar{y})}{\sqrt{\sum_{i=1}^{p}(|x_i|-\bar{x})^2}\sqrt{\sum_{i=1}^{p}(|y_i|-\bar{y})^2}}$ with $\bar{x} = \frac{\sum_i |x_i|}{p}, \bar{y} = \frac{\sum_i |y_i|}{p}$. The following corollary reveals the rationale behind the transformation $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$.

Corollary 19 *When $m = r$, denote $U_{1k}^* = (u_{k1}, \cdots, u_{kd_k})$ as the eigenvector of L_k^*. For any $i \in \mathcal{C}_k^*, j \in \mathcal{C}_l^*, k, l = 1, \cdots, m$, we have*

$$\mathrm{cor}(\mathrm{abs}(l_i^*), \mathrm{abs}(l_j^*)) = \begin{cases} 1 & \text{if } k = l \\ \frac{-\bar{u}_k \bar{u}_l}{\sqrt{\widetilde{u}_k - \bar{u}_k^2}\sqrt{\widetilde{u}_l - \bar{u}_l^2}} & \text{if } k \neq l \end{cases},$$

where $\bar{u}_k = \frac{1}{p}\sum_{h=1}^{d_k} |u_{kh}|$ and $\widetilde{u}_k = \frac{1}{p}\sum_{h=1}^{d_k} u_{kh}^2$. Furthermore, assume that $d_j \in [\frac{p}{\omega m}, \frac{\omega p}{m}], j = 1, \cdots, m$ and conditions of Theorem 13 hold. Then

$$\lim_{n\to\infty} \mathbb{P}\left(H(\tilde{\boldsymbol{t}}, \pi(\boldsymbol{t})) = 0\right) = 1.$$

This corollary follows directly from Theorems 17 and 13, noting that the oracle clustering error term is zero. In practice, before the third stage, we obtain a consistent estimate $\widehat{r}$ for the rank of L^*. If it is known that each community is rank-1, i.e., $m = \widehat{r}$, K-means is applied to $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$; otherwise, it is applied directly to $\widehat{L}$.

Remark 20 *Note that when clustering is performed directly on $\widehat{L}$ (or even on L^*), the oracle term is strictly positive unless the within-community row features are identical.*

To make this explicit, consider the rank-one within-community setting $m = r$ where each diagonal block admits $L_k^* = \sigma_k u_k u_k^T$ with $\|u_k\|_2 = 1$. For any $i \in \mathcal{C}_k^*$, the i-th row satisfies $l_i^* = \sigma_k u_{k,i} u_k^T$ (with zeros outside the k-th block), and hence the within-community dispersion is

$$\sum_{i\in\mathcal{C}_k^*} \|l_i^* - \mu_k^*\|^2 = \sigma_k^2 \sum_{i\in\mathcal{C}_k^*} (u_{k,i} - \bar{u}_k)^2, \qquad \bar{u}_k = \frac{1}{d_k}\sum_{i\in\mathcal{C}_k^*} u_{k,i}.$$

Therefore, unless $u_{k,i}$ is constant over $i \in \mathcal{C}_k^*$, the oracle term does not vanish. Moreover, if there exists $\upsilon_0 > 0$ such that $\min_{k\in[m]} \frac{1}{d_k}\sum_{i\in\mathcal{C}_k^*}(u_{k,i} - \bar{u}_k)^2 \geq \upsilon_0$ and $\sigma_{\min} = \min_k \sigma_k > 0$, then

$$\sum_{i=1}^{p} \|l_i^* - \mu_{\boldsymbol{t}(i)}^*\|^2 = \sum_{k=1}^{m} \sigma_k^2 \sum_{i\in\mathcal{C}_k^*} (u_{k,i} - \bar{u}_k)^2 \;\geq\; \sigma_{\min}^2 \sum_{k=1}^{m} d_k\, \upsilon_0 = \sigma_{\min}^2\, p\, \upsilon_0,$$

which is bounded away from zero as $n \to \infty$.

5. Simulation Study

We conduct simulation studies to investigate the performance of the proposed method. We present the results for the adaptive ℓ_1 penalized estimation in the second stage and K-means clustering in the third stage, respectively.

5.1 Second Stage: Adaptive ℓ_1 Penalized Estimation

We investigate the performance of the adaptive ℓ_1 penalized estimation method in the second stage from two perspectives described in Section 2. Throughout this section, we fix $p = 45, r = 3$ and take $n = 1000, 2000, 4000, 8000$. For both graphical models with non-overlapping latent communities and grouped latent variables, we first generate two covariates C_1, C_2 ($q = 2$) with $C_1, C_2 \sim N(0, 1)$ independently and let $B_{ij} \sim \text{Uniform}(0.5, 1)$, for $i = 1, \cdots, q$, and $j = 1, \cdots, p$. Then for the precision matrix of R, we consider the following two scenarios separately.

1. **Latent community graphical models**: For non-overlapping community graph L in Section 2.1, we take $m = 2$, i.e., there exist two latent communities: the first community is of rank 2 and the second is of rank 1. Let nodes $N_1 = \{1, 2, \cdots, 25\}$ and $N_2 = \{26, \cdots, 45\}$ belong to the first community and the second community respectively, i.e., we suppose $\boldsymbol{a}_i = (\boldsymbol{a}_{i,N_1}, \boldsymbol{a}_{i,N_2}), i = 1, 2, 3$. In this example, we set $\boldsymbol{a}_{1,N_1} = 3\boldsymbol{u}_1, \boldsymbol{a}_{2,N_1} = 2.5\boldsymbol{u}_2, a_{3,N_1} = 0$ and $\boldsymbol{a}_{1,N_2} = 0, \boldsymbol{a}_{2,N_2} = 0, \boldsymbol{a}_{3,N_2} = 2\boldsymbol{u}_3$, where $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^{25}, \boldsymbol{u}_3 \in \mathbb{R}^{20}$ are unit vectors and $\boldsymbol{u}_1, \boldsymbol{u}_2$ are orthogonal to each other. For the sparse structure S, we set $s_{ii} = 5$, for $i = 1, \cdots p$. To encode additional within-community sparse connections, we set

$$s_{ji} = s_{ij} \sim \text{Bernoulli}(\pi) * \text{Uniform}\left((-2, -1.5) \cup (1.5, 2)\right),$$

   independently, where $\pi = 0.01$, for $i \in N_1, j \in N_2$. In addition, to record information within communities, we set $s_{(i+2),i} = s_{i,(i+2)} \sim \text{Uniform}\left((-2, -1.5) \cup (1.5, 2)\right), i \in \{26, \cdots, 35\}$.

2. **Grouped latent variable graphical models**: Suppose $m = 3$, i.e. there are three groups of latent variables, each of which has only one latent variable H_i, for $i = 1, 2, 3$. To meet the conditional independence assumption (4), we let nodes $N_1 = \{1, 2, \cdots, 15\}$, $N_2 = \{16, \cdots, 30\}$ and $N_3 = \{31, \cdots, 45\}$, and suppose $(\Theta_{O,H_1})_{N_1} = 3.5\boldsymbol{u}_1, (\Theta_{O,H_2})_{N_2} = 3\boldsymbol{u}_2, (\Theta_{O,H_3})_{N_3} = 2.5\boldsymbol{u}_3$ and $(\Theta_{O,H_1})_{N_2 \cup N_3} = 0, (\Theta_{O,H_2})_{N_1 \cup N_3} = 0, (\Theta_{O,H_3})_{N_1 \cup N_2} = 0$, where $\boldsymbol{u}_1 \in \mathbb{R}^{15}, \boldsymbol{u}_2 \in \mathbb{R}^{15}, \boldsymbol{u}_3 \in \mathbb{R}^{15}$ are three unit vectors. Denote $\Theta_O = (\theta^O_{ij})$, and $\Theta_H = (\theta^H_{ij})$. For Θ_O, we generate $\theta^O_{i,(i+2)} = \theta^O_{(i+2),i} \sim \text{Uniform}\left((-2, -1.5) \cup (1.5, 2)\right)$ independently within N_1, $i = 1, \cdots, 15$. For the diagonal, we set $\theta^O_{ii} = 5, i = 1, \cdots p$ and $\theta^H_{jj} = 3$, $j = 1, \cdots, r$.

The proposed method will be compared to three other competing methods. First, we obtain the estimator by solving (11), where we adopt the BIC criterion used in Chen et al. (2016) to select the tuning parameters. In what follows, we refer to it as the estimator from the latent variable Gaussian graphical model (LVGGM). Next, we extend the LVGGM method by adopting the hard thresholding criterion $\widehat{L}^{\text{HT-LVGGM}} = \widehat{L}^{\text{LVGGM}} \cdot 1_{\{|\widehat{L}^{\text{LVGGM}}| > a_n\}}$, and refer to it as the estimator from the hard-thresholding latent variable Gaussian graphical model (HT-LVGGM). Note that the hard thresholding step will produce sparsity (we will set $a_n \asymp \frac{1}{\sqrt{n}}$ in practice), and we also set $\widehat{S}^{\text{HT-LVGGM}}$ same as $\widehat{S}^{\text{LVGGM}}$. Third, the nonadaptive penalized maximum likelihood estimation method (NonAPMLE) is a special case of the proposed adaptive ℓ_1 penalized estimation, where we take $W_n = \mathbf{1}_p \mathbf{1}_p^T$.

We evaluate the CV-based tuning parameter selection via five criteria. Let $(\widehat{S}, \widehat{L})$ be the estimates of the selected model defined as in (6). Here TR_L stands for the proportion of true rank estimation of L, which equals one when the rank is correctly estimated,

$$\mathrm{TR}_L = 1_{\left\{\mathrm{rank}(\widehat{L})=\mathrm{rank}(L^*)\right\}}.$$

The criterion TP_L evaluates the proportion of true discoveries of non-zero elements in the network structure estimation of L (the closer to one the better),

$$\mathrm{TP}_L = \frac{|\{(k,l) : k \le l, \hat{l}_{kl} \neq 0, \text{ and } l^*_{kl} \neq 0\}|}{|\{(k,l) : k \le l, l^*_{kl} \neq 0\}|}.$$

We use FP_L to represent the proportion of false discoveries of zeros in off-diagonal blocks of L (the closer to zero the better),

$$\mathrm{FP}_L = \frac{|\{(k,l) : k \le l, \hat{l}_{kl} \neq 0, \text{ and } l^*_{kl} = 0\}|}{|\{(k,l) : k \le l, l^*_{kl} = 0\}|}.$$

Similarly, TP_S evaluates the proportion of true discoveries of non-zero elements in the network structure estimation of S (the closer to one the better),

$$\mathrm{TP}_S = \frac{|\left\{(i,j) : i < j, \hat{s}_{ij} \neq 0, \text{ and } s^*_{ij} \neq 0\right\}|}{\left|\left\{(i,j) : i < j, s^*_{ij} \neq 0\right\}\right|}.$$

The criterion FP_S evaluates the proportion of false discoveries of zeros in S (the closer to zero the better),

$$\mathrm{FP}_S = \frac{|\left\{(i,j) : i < j, \hat{s}_{ij} \neq 0, \text{ and } s^*_{ij} = 0\right\}|}{\left|\left\{(i,j) : i < j, s^*_{ij} = 0\right\}\right|}.$$

For each sample size, we generate 100 replications to obtain the mean and standard deviation of the five evaluation criteria above for LVGGM, HT-LVGGM, NonAPMLE as well as the proposed method. The results for graphical models with non-overlapping latent communities and grouped latent variables are presented in Tables 1 and 2 respectively.

From the numerical results, we see that for sufficiently large sample sizes, LVGGM, NonAPMLE, and the proposed method can accurately identify the rank of L. However, for the small sample size, NonAPMLE fails to identify rank, while the proposed method performs comparably well as LVGGM. Also, the proposed method efficiently identifies both non-zero blocks and zero blocks in L, indicating the community structure in the graph. In comparison, LVGGM and NonAPMLE struggle to recognize zero blocks. Though HT-LVGGM can identify the community structure, its accuracy in identifying the rank of L is almost zero. Moreover, all four methods correctly identify the zero and non-zero elements in S asymptotically, demonstrating their ability to correctly identify the sparse structure. When the sparse graph contains edges both between and within communities, the proposed method performs slightly worse compared to other methods, necessitating more samples to improve accuracy (cf. Table 1). However, when the sparse graph has edges only within communities, all four methods demonstrate nearly identical performance in accurately identifying all zero and non-zero elements (cf. Table 2).

5.2 Third Stage: K-means Clustering

Now we design experiments to evaluate the Hamming error rates of K-means procedure based on both $\widehat{L}$ and $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$ estimated from the four methods in the second stage. For illustration, we adopt the similar data generation mechanism designed in Section 5.1 for grouped latent variable graphical models among two scenarios with the following specifications:

1. $(\Theta_{O,H_1})_{N_1} = a\boldsymbol{u}_1, (\Theta_{O,H_2})_{N_2} = (a-0.1)\boldsymbol{u}_2, (\Theta_{O,H_3})_{N_3} = (a-0.2)\boldsymbol{u}_3$, where $u_{1i}, u_{2i}, u_{3i} \sim$ Uniform$(1,2)$ independently, and for comparison a is taken as $1.5, 2.0, 2.5, 3.0, 3.5$ respectively. These configurations reflect the maximum eigenvalues of the community part L when the variation within each community is small. In this set of experiments, there are three clusters (communities), namely N_1, N_2, N_3. Moreover, for each value of a, 100 replications are generated to obtain the mean and standard deviations of Hamming error rates of K-means based on four methods.

2. $(\Theta_{O,H_1})_{N_1} = 3.6\boldsymbol{u}_1, (\Theta_{O,H_2})_{N_2} = 3.3\boldsymbol{u}_2, (\Theta_{O,H_3})_{N_3} = 3\boldsymbol{u}_3$, $u_{1i} \sim$ Normal$(1,1), u_{2i} \sim$ Normal$(2,1), u_{3i} \sim$ Normal$(-1,1)$ independently. In this scenario, the variation within each community is larger than in the first scenario, and we set sample sizes as $1000, 2000, 3000, and 4000$ respectively to evaluate the performance of different clustering methods. For each sample size, 100 replications are generated to obtain the mean and standard deviations of Hamming error rates of K-means based on four methods.

To facilitate a consistent comparison across methods, we remove the zero rows produced by the proposed adaptive ℓ_1 estimator before performing clustering for all four methods. The empirical findings for each of these methods, employing either $\widehat{L}$ or $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$, are comprehensively depicted in Tables 3-4 and Figures 1-2 for both scenarios.

In the first scenario, based on both $\widehat{L}$ and $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$, when the strength of the maximum eigenvalue of the community part L is strong and the variation within each community is small, K-means under all four methods can correctly identify members within each community. Moreover, as the parameter a decreases, K-means clustering leveraging $\widehat{L}$ estimated by the proposed method attains the lowest Hamming error rate, thereby surpassing the alternative methods (cf. Table 3 and Figure 1). In contrast, the performance of K-means clustering remains largely consistent across all four methods when utilizing $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$, irrespective of the eigenvalue magnitudes of L. It is important to note, however, that in situations where the maximum eigenvalue of L is relatively weak, the resultant clustering error tied to $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$ markedly exceeds that associated with $\widehat{L}$. This discrepancy can be attributed to the heightened sensitivity of K-means clustering to significant estimation errors in $\widehat{L}$ estimated from the second stage when a is small.

In the second scenario, K-means based on $\widehat{L}$ from the adaptive ℓ_1 estimator yields smaller error rates than LVGGM as the sample size increases. Because there is non-negligible variation within each community, the oracle error term in Theorem 17 does not vanish. When clustering is based on $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$, the Hamming error rates of all methods converge to zero, and the proposed method continues to outperform LVGGM (see Table 4 and Figure 2).

Table 1: The average and standard deviation of different evaluation criteria are reported for latent community graphical models. Two covariates ($q = 2$) are incorporated and the sample size is $n = 1000, 2000, 4000, 8000$ respectively. The rank r, number of communities m, and number of nodes p are fixed as 3, 2, and 45 respectively. A 5-fold CV is adopted to select tuning parameters for NonAPMLE and the proposed method. For each sample size, 100 replications are performed. Criterion TR_L (the closer to one the better), TP_L (the closer to one the better), FP_L (the closer to zero the better), TP_S (the closer to one the better) and FP_S (the closer to zero the better) are used.

Sample Size	Criteria	LVGGM	HT-LVGGM	NonAPMLE	Proposed
$n = 1000$	TR_L	0.960 (0.197)	0.000 (0.000)	0.010 (0.100)	0.960 (0.197)
	TP_L	1.000 (0.000)	0.884 (0.036)	1.000 (0.000)	0.921 (0.040)
	FP_L	1.000 (0.000)	0.018 (0.016)	0.999 (0.002)	0.053 (0.025)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.003 (0.003)	0.003 (0.003)	0.027 (0.005)	0.045 (0.006)
$n = 2000$	TR_L	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	TP_L	1.000 (0.018)	0.937 (0.000)	1.000 (0.000)	0.930 (0.030)
	FP_L	1.000 (0.000)	0.011 (0.010)	0.996 (0.003)	0.020 (0.012)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.001)	0.000 (0.001)	0.005 (0.002)	0.014 (0.003)
$n = 4000$	TR_L	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	TP_L	1.000 (0.000)	0.954 (0.010)	1.000 (0.000)	0.965 (0.014)
	FP_L	1.000 (0.000)	0.003 (0.005)	0.996 (0.003)	0.015 (0.011)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.000)	0.000 (0.000)	0.001 (0.001)	0.003 (0.001)
$n = 8000$	TR_L	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	TP_L	1.000 (0.000)	0.960 (0.010)	1.000 (0.000)	0.984 (0.011)
	FP_L	1.000 (0.000)	0.001 (0.002)	0.997 (0.003)	0.014 (0.010)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.000)	0.000 (0.000)	0.001 (0.001)	0.001 (0.001)

Table 2: The average and standard deviation of different evaluation criteria are reported for the grouped latent variable graphical models. Two covariates ($q = 2$) are incorporated and the sample size is $n = 1000, 2000, 4000, 8000$ respectively. The number of latent variables r, number of communities m, and number of observed variables p are fixed as 3, 3, and 45 respectively. A 5-fold CV is adopted to select tuning parameters for NonAPMLE and the proposed method. For each sample size, 100 replications are performed. Criterion TR_L (the closer to one the better), TP_L (the closer to one the better), FP_L (the closer to zero the better), TP_S (the closer to one the better) and FP_S (the closer to zero the better) are used.

Sample Size	Criteria	LVGGM	HT-LVGGM	NonAPMLE	Proposed
$n = 1000$	TR_L	1.000 (0.000)	0.000 (0.000)	0.040 (0.197)	0.910 (0.288)
	TP_L	1.000 (0.000)	0.805 (0.043)	1.000 (0.001)	0.884 (0.057)
	FP_L	1.000 (0.000)	0.021 (0.012)	0.995 (0.003)	0.086 (0.055)
	TP_S	0.999 (0.013)	0.999 (0.013)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.001 (0.004)	0.001 (0.004)	0.000 (0.000)	0.000 (0.000)
$n = 2000$	TR_L	1.000 (0.000)	0.000 (0.000)	0.440 (0.499)	0.910 (0.288)
	TP_L	1.000 (0.000)	0.825 (0.041)	1.000 (0.001)	0.928 (0.038)
	FP_L	1.000 (0.000)	0.011 (0.001)	0.996 (0.002)	0.078 (0.041)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$n = 4000$	TR_L	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	0.990 (0.100)
	TP_L	1.000 (0.000)	0.849 (0.031)	1.000 (0.000)	0.952 (0.031)
	FP_L	1.000 (0.000)	0.008 (0.006)	0.997 (0.002)	0.068 (0.029)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$n = 8000$	TR_L	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	TP_L	1.000 (0.000)	0.864 (0.024)	1.000 (0.000)	0.963 (0.028)
	FP_L	1.000 (0.000)	0.006 (0.004)	0.996 (0.003)	0.057 (0.016)
	TP_S	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	FP_S	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

Table 3: When clustering based on $\widehat{L}$ and Cor(abs($\widehat{L}$)), the mean and standard deviation for Hamming error rates under different values of a (representing the strengths of eigenvalues of latent community part L) are reported for the grouped latent variable graphical models. Two covariates are incorporated and the sample size is fixed as $n = 1000$. The number of latent variables r, number of communities m, and number of observed variables p are fixed as 3, 3, and 45 respectively. For each value of a, 100 replications are performed.

Clustering based	a	LVGGM	HT-LVGGM	NonAPMLE	Proposed
$\widehat{L}$	1.5	0.157 (0.071)	0.153 (0.064)	0.126 (0.043)	0.121 (0.029)
	2.0	0.155 (0.071)	0.161 (0.071)	0.128 (0.045)	0.116 (0.028)
	2.5	0.088 (0.049)	0.086 (0.058)	0.095 (0.041)	0.099 (0.042)
	3.0	0.007 (0.015)	0.007 (0.015)	0.004 (0.010)	0.008 (0.016)
	3.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Cor(abs($\widehat{L}$))	1.5	0.530 (0.062)	0.513 (0.059)	0.521 (0.058)	0.537 (0.060)
	2.0	0.436 (0.088)	0.426 (0.096)	0.408 (0.091)	0.469 (0.078)
	2.5	0.087 (0.080)	0.065 (0.078)	0.067 (0.044)	0.086 (0.075)
	3.0	0.005 (0.011)	0.004 (0.010)	0.004 (0.009)	0.003 (0.007)
	3.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

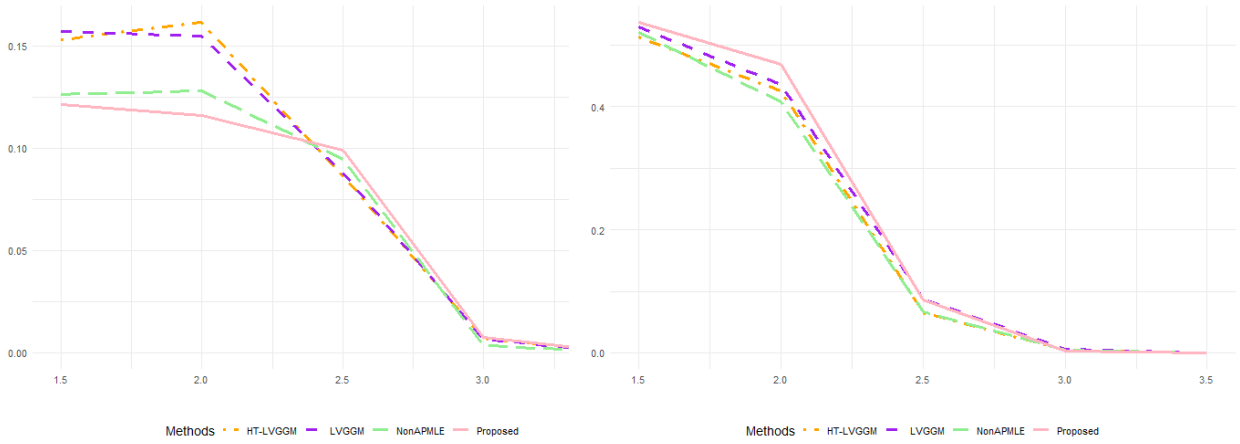

Figure 1: Comparison of Hamming error rates for HT-LVGGM, LVGGM, NonAPMLE and the proposed method under different values of a (representing the strengths of eigenvalues of latent community part L). Left panel: clustering based on $\widehat{L}$. Right panel: clustering based on Cor(abs($\widehat{L}$)). x-axis: different values of a. y-axis: Hamming error rates.

Table 4: When clustering based on $\widehat{L}$ and Cor(abs($\widehat{L}$)), the mean and standard deviation for Hamming error rates under different sample size are reported for the grouped latent variable graphical models. Two covariates are incorporated and the sample size is set as $1000, 2000, 3000, 4000$ respectively. The number of latent variables r, number of communities m, and number of observed variables p are fixed as 3, 3, and 45 respectively. For each sample size, 100 replications are performed.

Clustering based	Sample size n	LVGGM	HT-LVGGM	NonAPMLE	Proposed
$\widehat{L}$	1000	0.225 (0.027)	0.237 (0.020)	0.167 (0.031)	0.181 (0.031)
	2000	0.228 (0.022)	0.239 (0.014)	0.167 (0.029)	0.174 (0.033)
	3000	0.231 (0.017)	0.235 (0.015)	0.171 (0.025)	0.172 (0.025)
	4000	0.233 (0.015)	0.234 (0.014)	0.170 (0.021)	0.172 (0.023)
Cor(abs($\widehat{L}$))	1000	0.080 (0.032)	0.104 (0.033)	0.072 (0.034)	0.088 (0.044)
	2000	0.067 (0.030)	0.084 (0.030)	0.052 (0.027)	0.055 (0.028)
	3000	0.060 (0.030)	0.064 (0.030)	0.049 (0.025)	0.053 (0.028)
	4000	0.051 (0.029)	0.055 (0.028)	0.050 (0.023)	0.047 (0.024)

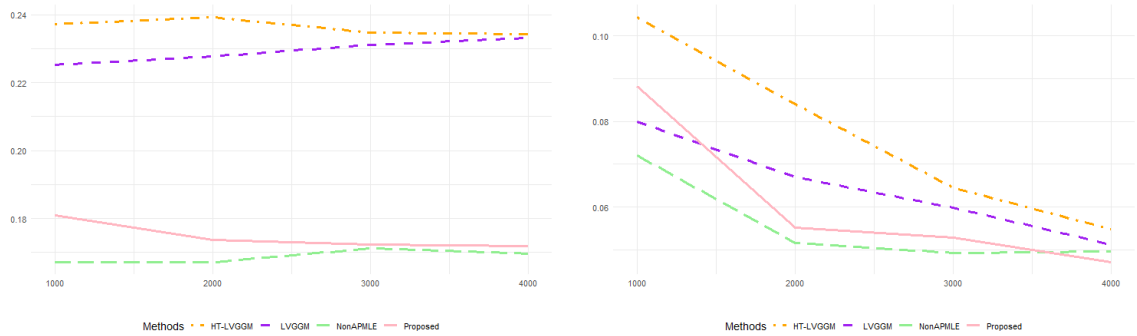

Figure 2: Comparison of Hamming error rates for HT-LVGGM, LVGGM, NonAPMLE and the proposed method under different sample sizes. Left panel: clustering based on $\widehat{L}$. Right panel: clustering based on Cor(abs($\widehat{L}$)). x-axis: different values of sample size. y-axis: Hamming error rates.

6. Real Data Application

We apply the proposed method to stock return data. We analyze 45 stocks from the energy, financial, and healthcare sectors, with 15 stocks from each sector. They are associated with public health and safety, environmental protection, and societal stability. We collect their adjusted closing prices between January 2014 and December 2023. These data are publicly available on the Yahoo Finance website and are bundled in standard software

packages such as the R package `quantmod`. To eliminate possible serial dependence, we compute the stock returns based on prices from alternating days, which lead to a sample of $n = 1253$ observations for each stock. In addition, to model the market effect, we use the S&P 500 index as our covariate. Then in the second stage, we apply the proposed adaptive ℓ_1 penalized estimation method to the grouped latent variable graphical models defined in Section 2.2 to learn the sparse graph S and the non-overlapping latent community graph L. For comparison, we use LVGGM as the benchmark model.

After accounting for the market effect revealed by the S&P 500 index, both LVGGM and the proposed method estimate the rank of the non-overlapping latent community graph L as $r = 3$, indicating three latent variables. This coincides with the three different sectors we have determined beforehand. Additionally, Figure 3 depicts the heatmaps of L estimated by both methods. From the perspective of community detection, the LVGGM method struggles to accurately identify the number of communities, indicating that it does not fully reveal the inter-dependencies among different sectors. In contrast, the proposed method significantly identifies the independent relationships among the three sectors.

Figure 4 shows the connectivity graphs for L estimated by both methods. The graph estimated by LVGGM is almost fully connected, whereas the proposed method demonstrates the independence among the three sectors (green represents the healthcare sector, pink represents the financial sector, and blue represents the energy sector). Specifically, combining Figures 3(b) and 4(b), we find a few weak edges between the energy sector and the financial sector. We speculate that with an increase in sample size, these edges may disappear. In addition, the healthcare sector exhibits almost no edges connecting it to the energy and financial sectors. This indicates that, after removing the influence of the overall market effect, the development of these three sectors is relatively independent.

Due to the presence of small values in the diagonal blocks of the initial estimator $\widehat{L}^{LVGGM}$, the corresponding diagonal blocks in the estimated latent community graph obtained from the proposed method are compressed to 0 (cf. Figure 3(b)). This explains the absence of the following nodes in Figure 4(b): (1) VLO (Valero Energy Corporation) in the energy sector; (2) V (Visa Inc.), MA (Mastercard Incorporated), and SPGI (S&P Global Inc.) companies in the financial sector; and (3) UNH (UnitedHealth Group Incorporated), TMO (Thermo Fisher Scientific Inc.), and MDT (Medtronic plc) companies in the healthcare sector. The absence of these nodes suggests that the corresponding sector effects are weak for these stocks and may be shrunk to zero in finite samples.

With regard to the estimation of the sparse graph S, we observe from Figures 5 and 6 that both methods identify similar values of correlations, aligning with the results from simulation studies. In particular, Table 2 shows both LVGGM and the proposed method can accurately identify zero and non-zero elements of S when the sample size is around 1000, and S contains only a small number of edges within the community. For instance, PSX (Phillips 66), VLO, and MPC (Marathon Petroleum Corporation) are among the major companies in the energy and petroleum industry and are involved in the downstream sector of the oil and gas industry, especially for refining and marketing. Due to their important roles in the production and distribution of petroleum products, even after removing the impact of the overall market and sectors, these three companies remain closely interconnected as is shown in Figure 6. A similar interpretation applies to the relationship between KMI (Kinder Morgan, Inc.) and WMB (Williams Companies, Inc.) as well. Also, as mentioned earlier,

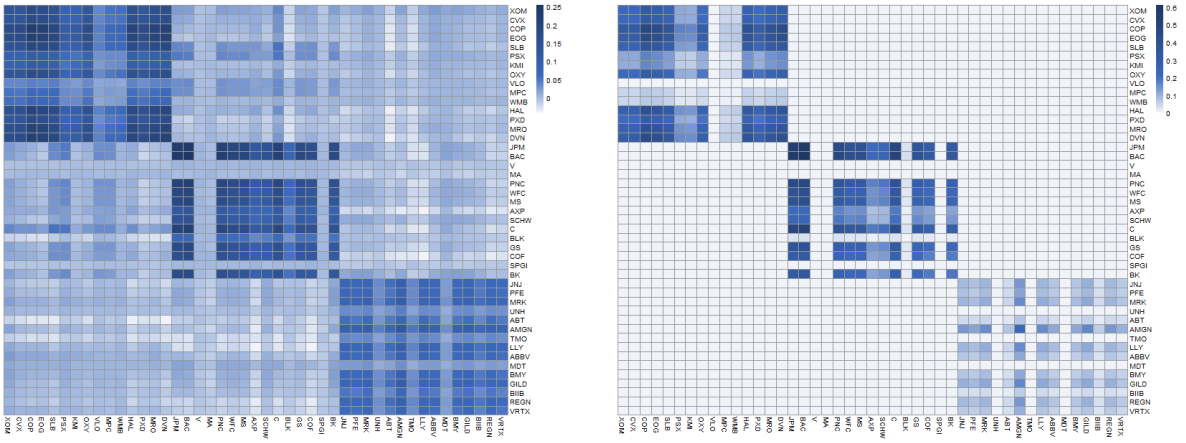

Figure 3: Heatmaps of latent community graph estimated by LVGGM method (left panel) and proposed method (right panel) respectively. The top 15 companies on the vertical axis are from the energy sector, the middle 15 are from the financial sector, and the bottom 15 are from the healthcare sector.

the financial services sector has little influence on the stock prices of V and MA. However, as observed from Figures 5 and 6, they still exhibit a strong correlation. This might be attributed to the fact that both companies are engaged in payment services, sharing similar business profiles and investor groups.

In the third stage, we apply K-means clustering to the outputs of both the proposed method and LVGGM. Because sector labels are available for these 45 stocks, we treat the true labels and the number of clusters as known and set $m = 3$. Owing to the shrinkage effect of the proposed method, 7 rows of $\widehat{L}$ are estimated as zero; therefore, clustering is performed on the remaining 38 stocks for both methods. When clustering is applied directly to $\widehat{L}$, the Hamming error rates are 0.079 for LVGGM and 0.132 for the proposed method. When clustering is based on $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$, both methods achieve zero Hamming error. This suggests that, for this dataset, clustering based on $\mathrm{Cor}(\mathrm{abs}(\widehat{L}))$ is more reliable than clustering based directly on $\widehat{L}$.

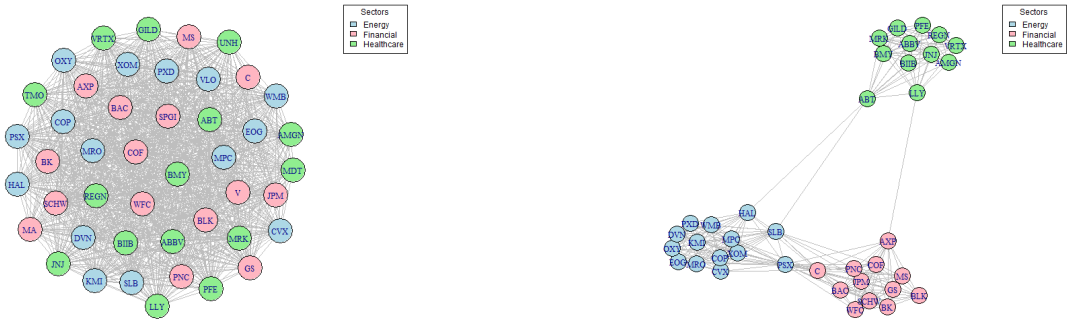

Figure 4: Latent community graphs estimated by LVGGM method (left panel) and the proposed method (right panel) respectively. Green represents the healthcare sector, pink represents the financial sector, and blue represents the energy sector.

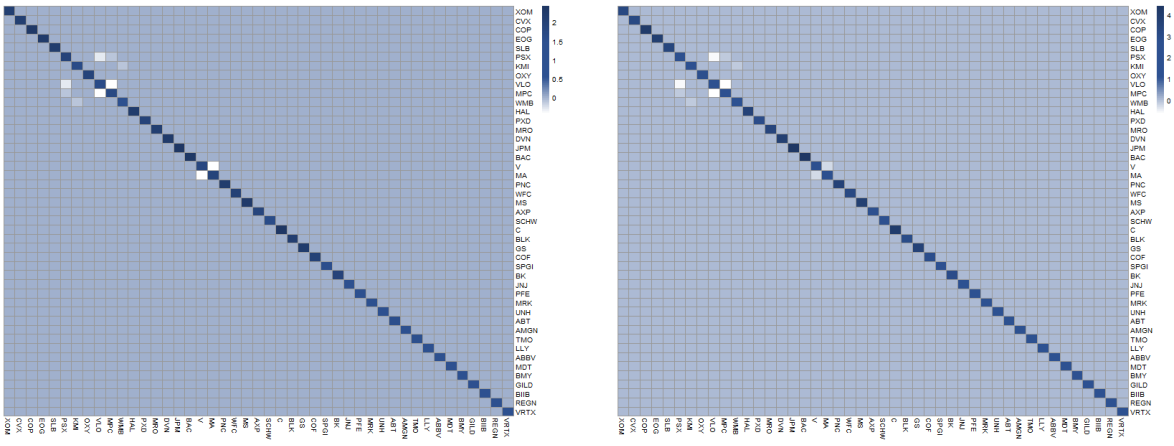

Figure 5: Heatmaps of sparse graph estimated by LVGGM method (left panel) and proposed method (right panel) respectively. The top 15 companies on the vertical axis are from the energy sector, the middle 15 are from the financial sector, and the bottom 15 are from the healthcare sector.

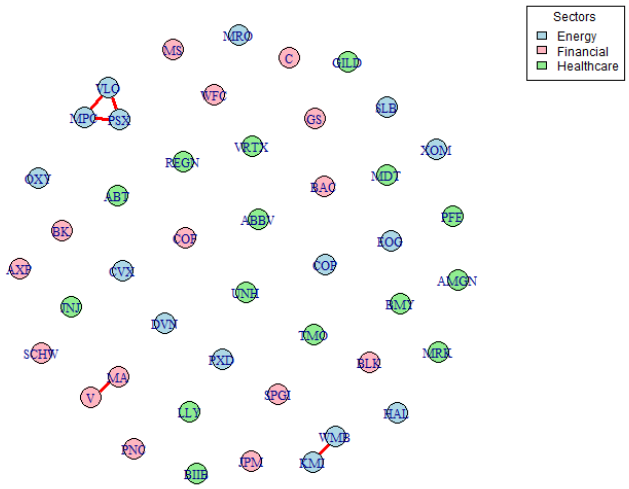

Figure 6: Sparse graph estimated by both LVGGM and the proposed method. Green represents the healthcare sector, pink represents the financial sector, and blue represents the energy sector.

7. Conclusion

In this paper, we propose a graph decomposition framework in which the precision matrix is decomposed into a sparse component and low-rank diagonal blocks. Based on this decomposition, we develop a three-stage procedure to estimate sparse conditional dependencies and recover non-overlapping communities after adjusting for global covariate effects. We provide two modeling interpretations of the decomposition, establish identifiability, derive a generalized irrepresentability condition for the adaptive ℓ_1 estimator in the second stage, and obtain a clustering error bound for the third-stage K-means procedure. Simulation studies and the stock return application show that the proposed method effectively recovers both community structure and sparse conditional dependencies.

In addition, we work under a fixed-dimension asymptotic regime in order to isolate the main methodological challenges arising from the grouped latent structure and the subsequent community-recovery step. Extending the theory to high-dimensional settings, overlapping communities, unobserved global factors, and discrete data would be valuable directions for future research.

Acknowledgments and Disclosure of Funding

T. Wang is supported by the National Natural Science Foundation of China Grant 12301660 and the Science and Technology Commission of Shanghai Municipality Grant 23JC1400700.

Appendix A. Algorithm Details

To ensure convergence of the ADMM algorithm, an equivalent convex minimization problem with two blocks of variables and two separable functions has been proposed in the problem (10). Denote $Y_1 = (\Theta, S, L_1, L_2)$, $Y_2 = (\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1, \widetilde{L}_2)$, problem (10) is

$$\begin{array}{ll} \min & f(Y_1) + \psi(Y_2) \\ s.t. & Y_1 - Y_2 = 0 \end{array} \tag{13}$$

where $f(Y_1) \triangleq f_1(\Theta) + f_2(S) + f_3(L_1) + f_4(L_2)$ with

$$f_1(\Theta) = -\log\det(\Theta) + \mathrm{tr}\left(\widehat{\Sigma}\Theta\right), f_2(S) = \gamma_n \cdot \|S\|_1,$$
$$f_3(L_1) = \tau_n \sum_{i,j=1}^{p} w_{ij}|L_{1,ij}|, f_4(L_2) = \delta_n \|L_2\|_*,$$

and $\psi(Y_2) \triangleq \psi_1(\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1) + \psi_2(\widetilde{L}_1, \widetilde{L}_2)$ with

$$\psi_1(\widetilde{\Theta}, \widetilde{L}_1, \widetilde{S}) = \mathcal{I}\{\widetilde{\Theta} - \widetilde{L}_1 - \widetilde{S} = 0\}, \psi_2(\widetilde{L}_2) \quad = \mathcal{I}\{\widetilde{L}_1 - \widetilde{L}_2 = 0\}.$$

The indicator function is defined as

$$\mathcal{I}(X \in \mathcal{X}) \triangleq \begin{cases} 0, & \text{if } X \in \mathcal{X} \\ +\infty, & \text{otherwise} \end{cases}.$$

Define the augmented Lagrangian function as

$$\mathcal{L}_\mu(Y_1, Y_2, \Gamma) = f(Y_1) + \psi(Y_2) + \Gamma \cdot (Y_1 - Y_2) + \frac{1}{2\mu}\|Y_1 - Y_2\|_F^2,$$

where Γ is the Lagrange multiplier and $\mu > 0$ is the penalty parameter. Then the problem (13) can be separated as some solvable subproblems through the scaled ADMM algorithm:

$$\begin{cases} Y_1^{k+1} \triangleq \underset{Y_1}{\mathrm{argmin}}\, \mathcal{L}_\mu(Y_1, Y_2^k, \Gamma^k) = \underset{Y_1}{\mathrm{argmin}}\, f(Y_1) + \frac{1}{2\mu}\left\|Y_1 - Y_2^k + \Gamma^k\right\|_F^2 \\ Y_2^{k+1} \triangleq \underset{Y_2}{\mathrm{argmin}}\, \mathcal{L}_\mu(Y_1^{k+1}, Y_2, \Gamma^k) = \underset{Y_2}{\mathrm{argmin}}\, \psi(Y_2) + \frac{1}{2\mu}\left\|Y_1^{k+1} - Y_2 + \Gamma^k\right\|_F^2, \\ \Gamma^{k+1} \triangleq \Gamma^k + \left(Y_1^{k+1} - Y_2^{k+1}\right). \end{cases}$$

We first define the proximal mapping for function $p(X) : \mathbb{R}^{p\times q} \to \mathbb{R}$ as

$$\mathrm{prox}_{\lambda, p(\cdot)}(Z) = \underset{X}{\mathrm{argmin}}\, p(X) + \frac{1}{2\lambda}\|X - Z\|_F^2.$$

The proximal mapping for $p(X) = \|X\|_1$ is the soft thresholding operator,

$$\mathrm{prox}_{\lambda, \|\cdot\|}(Z) = \underset{X}{\mathrm{argmin}}\, \|X\|_1 + \frac{1}{2\lambda}\|X - Z\|_F^2 = \mathrm{sign}(Z)(Z - \lambda \mathbf{1}\mathbf{1}^T)_+,$$

where $(\cdot)_+ = \max(\cdot, 0)$. The proximal mapping for $p(X) = \|X\|_*, X \succeq 0$ is

$$\text{prox}_{\lambda,\|\cdot\|_*}(Z) = \underset{X\succeq 0}{\text{argmin}} \|X\|_* + \frac{1}{2\lambda}\|X - Z\|_F^2 = U\widetilde{\Sigma}U^T,$$

where $Z = U\Sigma U^T, \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p)$ and $\widetilde{\Sigma} = \text{diag}\left((\sigma_1 - \lambda)_+, \ldots, (\sigma_p - \lambda)_+\right)$. The proximal mapping for $p(X) = I(X \in \mathcal{X})$ is the projection mapping

$$\text{prox}_I(Z) = \underset{X\in\mathcal{X}}{\text{argmin}} \frac{1}{2}\|X - Z\|_F^2 = \mathcal{P}_{\mathcal{X}}(Z),$$

where $\mathcal{P}_{\mathcal{X}}(Z)$ means the projection of Z on set $\mathcal{X}$ with respect to the inner product " $\cdot$ ".

A.1 First Subproblem

We consider the solution for the first subproblem:

$$Y_1^{k+1} = \underset{Y_1}{\text{argmin}}\, f(Y_1) + \frac{1}{2\mu}\left\|Y_1 - Y_2^k + \Gamma^k\right\|_F^2.$$

Denote $Y_1^{k+1} = (\Theta^{k+1}, S^{k+1}, L_1^{k+1}, L_2^{k+1})$, and partition the matrix $\Gamma^k = (\Gamma_\Theta^k, \Gamma_S^k, \Gamma_{L_1}^k, \Gamma_{L_2}^k)$ into four blocks associated to Y_2^k. This problem can also be reduced to the following four subproblems:

A.1.1 Update Step for Θ^{k+1}

$$\Theta^{k+1} = \underset{\Theta\succeq 0}{\text{argmin}} - \log\det(\Theta) + \text{tr}\left(\widehat{\Sigma}\Theta\right) + \frac{1}{2\mu}\left\|\Theta - \widetilde{\Theta}^k + \Gamma_\Theta^k\right\|_F^2, \tag{14}$$

The first-order optimal condition is given by $-\Theta^{-1} + \widehat{\Sigma} + \frac{1}{\mu}(\Theta - \widetilde{\Theta}^k + \Gamma_\Theta^k) = 0$. Denote the eigenvalue decomposition for $\widetilde{\Theta}^k - \mu\widehat{\Sigma} - \Gamma_\Theta^k$ by $U\,\text{diag}(\sigma)U^T$, then it is easy to verify that

$$\Theta^{k+1} = U\,\text{diag}(\xi)U^T, \quad \xi_i = \frac{\sigma_i + \sqrt{\sigma_i^2 + 4\mu}}{2}$$

satisfies the optimal condition and thus is the solution to the problem (14). (Ma et al. (2013))

A.1.2 Update Step for S^{k+1}

$$\begin{aligned}\left(S^{k+1}\right)_{ij} &= \underset{S}{\text{argmin}} \|S\|_{\text{off},1} + \frac{1}{2\mu\gamma_n}\left\|S - \widetilde{S}^k + \Gamma_S^k\right\|_F^2 \\ &= \begin{cases} \text{prox}_{\mu\gamma_n,\|\cdot\|}\left(\widetilde{S}_{ij}^k - (\Gamma_S^k)_{ij}\right), i \neq j \\ \widetilde{S}_{ij}^k - \left(\Gamma_S^k\right)_{ij}, \quad i = j. \end{cases}\end{aligned}$$

A.1.3 Update Step for L_1^{k+1}

Let $S^+ = \{L : L \succeq 0, L = L^T\}$ be the positive semidefinite cone,

$$\begin{aligned} L_1^{k+1} &= \underset{L_1 \succeq 0}{\operatorname{argmin}} \sum_{i,j=1}^{p} w_{ij}|L_{1,ij}| + \frac{1}{2\mu\tau_n}\left(\left\|L_1 - \widetilde{L}_1^k + \Gamma_{L_1}^k\right\|_F^2\right) \\ &= \mathcal{P}_{S^+}\left(\left(\operatorname{prox}_{\mu\tau_n w_{ij},\|\cdot\|}\left((\widetilde{L}_1^k)_{ij} - (\Gamma_{L_1}^k)_{ij}\right)\right)_{ij}\right) \end{aligned}$$

A.1.4 Update Step for L_2^{k+1}

$$\begin{aligned} L_2^{k+1} &= \underset{L_2 \succeq 0}{\operatorname{argmin}} \|L_2\|_* + \frac{1}{2\mu\delta_n}\left(\left\|L_2 - \widetilde{L}_2^k + \Gamma_{L_2}^k\right\|_F^2\right) \\ &= \operatorname{prox}_{\mu\delta_n,\|\cdot\|_*}(\widetilde{L}_2^k - \Gamma_{L_2}^k) \end{aligned}$$

A.2 Second Subproblem

Partition the matrix $T^k := Y_1^{k+1} + \Gamma^k = (T_\Theta^k, T_S^k, T_{L_1}^k, T_{L_2}^k)$ into four blocks in the same form as $Y_2 = (\widetilde{\Theta}, \widetilde{L}_1, \widetilde{L}_2, \widetilde{S})$. The second subproblem can be reduced to

$$\begin{aligned} \min \quad & \frac{1}{2}\left\|(\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1, \widetilde{L}_2) - (T_\Theta^k, T_S^k, T_{L_1}^k, T_{L_2}^k)\right\|_F^2 \\ s.t. \quad & \widetilde{\Theta} - \widetilde{L}_1 - \widetilde{S} = 0, \quad \widetilde{L}_1 - \widetilde{L}_2 = 0. \end{aligned} \tag{15}$$

The Lagrangian function associated with this problem is

$$\mathcal{L}(Y_2) = \frac{1}{2}\left\|(\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1, \widetilde{L}_2) - (T_\Theta^k, T_S^k, T_{L_1}^k, T_{L_2}^k)\right\|_F^2 + \Psi_1 \cdot (\widetilde{\Theta} - \widetilde{L}_1 - \widetilde{S}) + \Psi_2 \cdot (\widetilde{L}_1 - \widetilde{L}_2),$$

where Ψ_1, Ψ_2 are the corresponding Lagrangian multipliers. The first-order optimal conditions are given by

$$(\widetilde{\Theta}, \widetilde{S}, \widetilde{L}_1, \widetilde{L}_2) - (T_\Theta^k, T_S^k, T_{L_1}^k, T_{L_2}^k) + (\Psi_1, -\Psi_1, -\Psi_1 + \Psi_2, -\Psi_2) = 0.$$

Then

$$\widetilde{\Theta} = T_\Theta^k - \Psi_1, \quad \widetilde{S} = T_S^k + \Psi_1, \quad \widetilde{L}_1 = T_{L_1}^k + \Psi_1 - \Psi_2, \quad \widetilde{L}_2 = T_{L_2}^k + \Psi_2.$$

Substituting them into the equality constraint of the problem (15) and we get the solution $Y_2 = (\widetilde{\Theta}^{k+1}, \widetilde{S}^{k+1}, \widetilde{L}_1^{k+1}, \widetilde{L}_2^{k+1})$ for the problem (15),

$$\begin{cases} \widetilde{\Theta}^{k+1} = \frac{3}{5}T_\Theta^k + \frac{2}{5}T_S^k + \frac{1}{5}T_{L_1}^k + \frac{1}{5}T_{L_2}^k \\ \widetilde{S}^{k+1} = \frac{2}{5}T_\Theta^k + \frac{3}{5}T_S^k - \frac{1}{5}T_{L_1}^k - \frac{1}{5}T_{L_2}^k \\ \widetilde{L}_1^{k+1} = \frac{1}{5}T_\Theta^k - \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \\ \widetilde{L}_2^{k+1} = \frac{1}{5}T_\Theta^k - \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \end{cases}.$$

A.3 Algorithm for Grouped Latent Variable Graphical Models

For the grouped latent variable graphical models mentioned in Section 2, the algorithm is a bit different from above due to -1 in the decomposition. The main difference is reflected in the solution of the second subproblem,

$$\begin{cases} \widetilde{\Theta}^{k+1} = \frac{3}{5}T_\Theta^k + \frac{2}{5}T_S^k - \frac{1}{5}T_{L_1}^k - \frac{1}{5}T_{L_2}^k \\ \widetilde{S}^{k+1} = \frac{2}{5}T_\Theta^k + \frac{3}{5}T_S^k + \frac{1}{5}T_{L_1}^k + \frac{1}{5}T_{L_2}^k \\ \widetilde{L}_1^{k+1} = -\frac{1}{5}T_\Theta^k + \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \\ \widetilde{L}_2^{k+1} = -\frac{1}{5}T_\Theta^k + \frac{1}{5}T_S^k + \frac{2}{5}T_{L_1}^k + \frac{2}{5}T_{L_2}^k \end{cases}.$$

Appendix B. Technical Details

B.1 Proof of Proposition

Proof [Proof of Proposition 2] (1). Denote $\mathcal{L}(L_i^*) = \{L_i \in \mathbb{S}^{d_i\times d_i} : rank(L_i) = rank(L_i^*)\}$ (they are manifolds, also submanifolds of $\mathbb{S}^{p\times p}$) and $\mathcal{M}(L^*) = \{L = \mathrm{diag}(L_1, \cdots, L_m) : rank(L_i) = rank(L_i^*), L_i \in \mathbb{S}^{d_i\times d_i}\}$. Since $\mathcal{M}(L^*) \cong \mathcal{L}(L_1^*) \times \cdots \times \mathcal{L}(L_m^*)$ ("$\cong$" means diffeomorphism), according to Example 1.13 in Lee (2012), $\mathcal{M}(L^*)$ is a product manifold. In addition, since $\mathcal{M}'(L^*) = \{L : rank(L) = rank(L^*), \mathrm{Supp}(L) = \mathrm{Supp}(L^*)\}$ is an open subset in $\mathcal{M}(L^*)$, according to Example 1.8 in Lee (2012), $\mathcal{M}'(L^*)$ is also an open submanifold of $\mathcal{M}(L^*)$. All $L \in \mathcal{M}'(L^*)$ form the smooth points in $\mathcal{LS}(L^*)$.
(2). Since $\mathcal{T}_1(L) = \{U_1Y^T + YU_1^T : Y \in \mathbb{R}^{p\times r}\}$ is the tangent space of $L \in \mathcal{M}(L^*)$ (Chandrasekaran et al., 2011) and $\mathcal{M}'(L^*)$ is also an open submanifold of $\mathcal{M}(L^*)$, then $\mathcal{T}(L) = \mathcal{T}_1(L) \cap \mathcal{T}_2^*$.
(3). Moreover, note that $\mathcal{V}(L)$ is orthogonal to $\mathcal{T}_2^{*\perp}$ and they are both the subspace of $\mathcal{T}^\perp(L)$. Hence, we only need to prove that $\dim(\mathcal{T}^\perp(L)) = \dim(\mathcal{V}(L)) + \dim(\mathcal{T}_2^{*\perp})$. In fact, from definition, we know $\dim(\mathcal{T}^\perp(L)) = \dim(\mathbb{S}^{p\times p}) - \dim(\mathcal{T}(L)) = \dim(\mathbb{S}^{p\times p}) - \sum_{i=1}^m \dim(\mathcal{T}_1(L_i))$, $\dim(\mathcal{V}(L)) = \sum_{i=1}^m [\dim(\mathbb{S}^{d_i\times d_i}) - \dim(\mathcal{T}_1(L_i))]$, and $\dim(\mathcal{T}_2^{*\perp}) = \sum_{i<j} \dim(\mathbb{R}^{d_i\times d_j})$ and $\dim(\mathbb{S}^{p\times p}) = \dim(\mathbb{S}^{d_i\times d_i}) + \sum_{i<j} \dim(\mathbb{R}^{d_i\times d_j})$, which leads to the conclusion. ■

B.2 Proof of Theorems

Proof [Proof of Theorem 4] According to Chandrasekaran et al. (2011), for every $i \in [m]$, (S_i^*, L_i^*) is locally identifiable with respect to $\mathcal{S}(s_i) \times \mathcal{L}(r_i)$ under Assumption 3, where $s_i \triangleq \|\mathbf{O}(S_i^*)\|_0$, i.e., there exists $\varepsilon > 0$ such that if it holds $S_i + L_i = S_i^* + L_i^*$ and $\|S_i - S_i^*\|_\infty \le \varepsilon, \|L_i - L_i^*\|_2 \le \varepsilon$ for $(S_i, L_i) \in \mathcal{S}(s_i) \times \mathcal{L}(r_i)$, then $S_i = S_i^*, L_i = L_i^*$. Hence for any $(S, L) \in \mathcal{S}(s_0) \times \mathcal{LS}(m, r)$ satisfying $\|S - S^*\|_\infty \le \varepsilon, \|L - L^*\|_2 \le \varepsilon$, if $S + L = S^* + L^*$, then $L = L^*$ (since $L = \mathrm{diag}(L_1, \cdots, L_m)$), which also implies $S = S^*$. This completes the proof. ■

Proof [Proof of Main Theorem 13] **Proof Strategy for Theorem 13**. The proof line is similar to Chen et al. (2016). We first provide a sketch of the proof for Theorem 13. We introduce several notations and definitions. Let the eigendecomposition of L^* be $L^* =$

$U^*D^*U^{*\top}$, such that U^* is a $p \times p$ orthogonal matrix and D^* is a $p \times p$ diagonal matrix whose first r diagonal elements are strictly positive. We write $U^* = [U_1^*, U_2^*]$ where U_1^* is the first r columns of U^*. Let D_1^* be the $r \times r$ diagonal matrix containing the nonzero diagonal elements of D^*. Define the localization set

$$\begin{aligned}\mathcal{M}_1 = \{(S, L) : S = S^* + \Delta_S, \|\Delta_S\|_\infty \le \gamma_n^{1-2\eta}, \Delta_S \text{ is symmetric,} \\ L = U\Sigma U^\top, \|U - U^*\|_\infty \le \gamma_n^{1-\eta}, U \text{ is a } p \times p \text{ orthogonal matrix,} \\ \|\Sigma - \Sigma^*\|_\infty \le \gamma_n^{1-2\eta}, \text{ and } \Sigma \text{ is a } p \times p \text{ diagonal matrix } \big\},\end{aligned}$$

and a subset

$$\begin{aligned}\mathcal{M}_2 = \{(S, L) : S = S^* + \Delta_S, \ \|\Delta_S\|_\infty \le \gamma_n^{1-2\eta}, \Delta_S \in \Omega^*, \\ L = U_1\Sigma_1 U_1^\top, \|U_1 - U_1^*\|_\infty \le \gamma_n^{1-\eta}, U_1 \in \mathbb{R}^{p\times r}, U_1^\top U_1 = I_r, (U_1)_{G_{ij}^*} = 0, i \ne j, \\ \|\Sigma_1 - \Sigma_1^*\|_\infty \le \gamma_n^{1-2\eta}, \text{ and } \Sigma_1 \text{ is a } r \times r \text{ diagonal matrix } \big\},\end{aligned}$$

where as discussed under Assumption 1, we choose U_1 owning the same zero-block pattern as $(U_1)_{G_{ij}^*} = 0, i \ne j$, and η being a positive constant that is sufficiently small. For each pair of $(S, L) \in \mathcal{M}_1, S$ is close to S^*. Moreover, the eigendecomposition of L and L^* are close to each other. As the sample size n grows large, the set $\mathcal{M}_1$ will tend to $\{(S^*, L^*)\}$. The set $\mathcal{M}_2$ is a subset of $\mathcal{M}_1$. For each pair of $(S, L) \in \mathcal{M}_2, S$ has the same sparsity pattern as S^*, and L has the same rank and block-sparsity pattern as L^* for sufficiently large n. The proof consists of three steps.

1. To be ready, we present some technical lemmas in Section B.2.1. In addition, we also provide some other useful lemmas interspersed with the proof of the main theorem. We later give the proof of all the lemmas in Section B.3.

2. We prove that with a probability converging to 1 the optimization problem (6) restricted to the subset $\mathcal{M}_2$ has a unique solution, which does not lie on the manifold boundary of $\mathcal{M}_2$. This part of the proof is presented in Step 1 of Section B.2.2.

3. We then show the unique solution restricted to $\mathcal{M}_2$ is also a solution to (6) on $\mathcal{M}_1$. It is further shown that with probability converging to 1, this solution is the unique solution to (6) restricted to $\mathcal{M}_1$. This part of the proof is presented in Step 2 of Section B.2.3.

The previous three steps together imply that the convex optimization problem (6) with the constraint $(S, L) \in \mathcal{M}_1$ has a unique solution that belongs to $\mathcal{M}_2$. Furthermore, this solution $(\widehat{S}, \widehat{L})$ is an interior point of $\mathcal{M}_1$ and thus is also the unique solution to the optimization problem (6) thanks to the convexity of the objective function. We conclude the proof for Theorem 13 by noticing that all $(S, L) \in \mathcal{M}_2$ converge to the true parameter (S^*, L^*) as $n \to \infty$.

B.2.1 Lemmas

We first establish some useful lemmas.

Lemma 21 *Let* $\mathbf{s}_{n1} = W_n \odot \operatorname{sign}(L^*)$. *Suppose Assumption 8 holds. Then,*

$$\|\mathbf{s}_{n1}\| = (1 + o_P(1))\, M_{n1} = o_P(r_n^{1-\alpha}).$$

Lemma 22 *Denote the eigendecomposition* $L = U_1 \Sigma_1 U_1^\top$, *and define the corresponding linear spaces* $\mathcal{T}(L)$ *and* $\mathcal{D}$ *as*

$$\mathcal{D} = \left\{ U_1 \Sigma_1' U_1^\top : \Sigma_1' \text{ is a } r \times r \text{ diagonal matrix} \right\},$$

and

$$\mathcal{T}(L) = \left\{ U_1 Y + Y^\top U_1^\top : Y \text{ is a } r \times p \text{ matrix} \right\}.$$

Then, the mappings $U_1 \to \mathcal{P}_{\mathcal{T}(L)}, U_1 \to \mathcal{P}_{\mathcal{T}^\perp(L)}$ *and* $U_1 \to \mathcal{P}_{\mathcal{D}}$ *are Lipschitz in* U_1. *That is, for all* $r \times r$ *symmetric matrix* M, *there exists a constant* κ *such that*

$$\begin{aligned} &\max\left\{ \left\|\mathcal{P}_{\mathcal{T}(L)} M - \mathcal{P}_{\mathcal{T}^*} M\right\|_\infty, \left\|\mathcal{P}_{\mathcal{T}^\perp(L)} M - \mathcal{P}_{T^{*\perp}} M\right\|_\infty, \left\|\mathcal{P}_{\mathcal{D}} M - \mathcal{P}_{\mathcal{D}^*} M\right\|_\infty \right\} \\ &\le \kappa \left\|U_1 - U_1^*\right\|_\infty \|M\|_\infty. \end{aligned}$$

Denote

$$g_{\rho_1,\rho_2,L,\widetilde{W}_n}(L') = \max\left\{ \frac{\|L_1\|_2}{\rho_1}, \frac{r_n^{1-\alpha}\|\widetilde{W}_n \odot L_2\|_\infty}{\rho_2} \right\},$$

where $L' = L_1 \oplus L_2$ is the unique orthogonal direct sum decomposition of L restricted to $\mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp} = \mathcal{T}^\perp(L)$, i.e., $L_1 = \mathcal{P}_{\mathcal{V}(L)}(L'), L_2 = \mathcal{P}_{\mathcal{T}_2^{*\perp}}(L')$. For convenience, without confusion sometimes we will abbreviate $\mathcal{V}(L)$ as $\mathcal{V}$. The following lemma states that this norm is well-defined.

Lemma 23 *Given any smooth point* L *of* $\mathcal{LS}(L^*)$ *with* $\mathcal{T}^\perp(L) = \mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp}$,

1. $g_{\rho_1,\rho_2,L,\widetilde{W}_n}(L)$ *is a norm for fixed* $\widetilde{W}_n, L$.
2. $\|(\mathcal{P}_{\mathcal{V}(L)} - \mathcal{P}_{\mathcal{V}^*})(M)\|_2 \lesssim \|L - L^*\|_2 \|M\|$.

Lemma 24 *The subgradient for* $\partial(\delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1)$ *has the following form*

$$\begin{aligned} \partial(\delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1) = &\{\delta_n UU^T + \tau_n W_n \odot \operatorname{sign}(L^*) + \delta_n F_1 + \tau_n W_n \odot F_2 : \\ &F_1 \in \mathcal{T}_1^\perp(L), F_2 \in \mathcal{T}_2^{*\perp}, \|F_1\|_\infty \le 1, \|F_2\|_2 \le 1\}. \end{aligned}$$

For any $Z \in \mathbb{R}^{p\times p}$, *let* $L = U\Sigma U^T$ *be the eigendecomposition of* L. *If*

$$\mathcal{P}_{\mathcal{T}(L)}(Z) = \delta_n UU^T + \tau_n \mathcal{P}_{\mathcal{T}(L)}\left(W_n \odot \operatorname{sign}(L)\right),$$

and

$$g_{\rho_1,\rho_2,L,\widetilde{W}_n}\left(\mathcal{P}_{\mathcal{T}^\perp(L)}(Z) - \tau_n \mathcal{P}_{\mathcal{T}^\perp(L)}\left(W_n \odot \operatorname{sign}(L)\right)\right) \le \gamma_n,$$

then $Z \in \partial\left(\delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1\right)$.

Lemma 25 *Under Assumption 3, the linear operator* $\mathbf{G}$ *is invertible over* $\Omega^* \times \mathcal{T}^*$.

Lemma 26 *Given any smooth point L of $\mathcal{LS}(L^*)$ with $\|L-L^*\|_2 \le d_0$ for sufficiently small $d_0 > 0$. For any $L' \in \mathcal{T}^{\perp}(L) = \mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp}$, it can be uniquely decomposed as $L' = L_1 \oplus L_2$. Then there exists some constants $\varepsilon_0 > 0, \varepsilon_1 > 0$ such that*

$$\|L'\|_F^2 = \|L_1\|_F^2 + \|L_2\|_F^2.$$

Moreover,

$$\max\left\{\|S\|_\infty, l_n\|L'\|_2\right\} \le h_{\rho_1,\rho_2,L,\Lambda_n}(S,L') \le \max\left\{\|S\|_\infty, u_n\|L'\|_2\right\},$$

where l_n, u_n are bounded as n tending to infinite under Assumption 8. (All the constants are specified in the proof.)

Remark 27 *This lemma states that the norm depending on L can be dominated by the norm $\max\{\|\cdot\|_\infty, \|\cdot\|_2\}$ uniformly.*

B.2.2 Step 1

Denote by $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ a solution to the optimization problem

$$\begin{aligned} &\min\left\{\ell_n(S+L) + \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1\right\} \\ &\text{subject to} \quad L \succeq 0, \quad S = S^T, \quad (S,L) \in \mathcal{M}_2. \end{aligned} \tag{16}$$

We write the eigendecomposition $\widehat{L}_{\mathcal{M}_2} = \widehat{U}_{1,\mathcal{M}_2}\widehat{\Sigma}_{1,\mathcal{M}_2}\widehat{U}_{1,\mathcal{M}_2}^\top$, where $\widehat{U}_{1,\mathcal{M}_2}^\top\widehat{U}_{1,\mathcal{M}_2} = I_r$, $\widehat{U}_{1,\mathcal{M}_2}$ has the same zero-block pattern as U_1^* and $\widehat{\Sigma}_{1,\mathcal{M}_2}$ is a $r \times r$ diagonal matrix. To establish that $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ does not lie on the manifold boundary of $\mathcal{M}_2$, it is sufficient to show that

$$\left\|\widehat{S}_{\mathcal{M}_2} - S^*\right\|_\infty <_P \gamma_n^{1-2\eta}, \tag{17}$$

$$\left\|\widehat{\Sigma}_{1,\mathcal{M}_2} - \Sigma_1^*\right\|_\infty <_P \gamma_n^{1-2\eta}, \tag{18}$$

$$\left\|\widehat{U}_{1,\mathcal{M}_2} - U_1^*\right\|_\infty <_P \gamma_n^{1-\eta}. \tag{19}$$

Lemma 28 *For $\|\widehat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty \le \gamma_n^{1-\eta}$, let*

$$\widehat{\mathcal{D}} = \left\{\widehat{U}_{1,\mathcal{M}_2}\Sigma_1'\widehat{U}_{1,\mathcal{M}_2}^\top : \Sigma_1' \text{ is a } r \times r \text{ diagonal matrix}\right\}$$

Consider the convex optimization problem

$$\min_{S\in\Omega^*, L\in\widehat{\mathcal{D}}} \left\{\ell_n(S+L) + \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1\right\}. \tag{20}$$

Then (20) has a unique solution with probability converging to 1. Denote the solution by $\left(\widehat{S}_{\widehat{\mathcal{D}}}, \widehat{L}_{\widehat{\mathcal{D}}}\right)$ and $\widehat{L}_{\widehat{\mathcal{D}}} = \widehat{U}_{1,\mathcal{M}_2}\widehat{\Sigma}_{1,\widehat{\mathcal{D}}}\widehat{U}_{1,\mathcal{M}_2}^\top$. In addition, there exists a constant $\kappa > 0$ such that $\left\|\widehat{S}_{\widehat{\mathcal{D}}} - S^*\right\|_\infty \le_P \kappa\gamma_n^{1-\eta}$ and $\left\|\widehat{\Sigma}_{1,\widehat{\mathcal{D}}} - \Sigma_1^*\right\|_\infty \le_P \kappa\gamma_n^{1-\eta}$.

According to the convexity of the objective function, a direct application of Lemma 28 is

$$\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right) = {}_P \left(\widehat{S}_{\widehat{\mathcal{D}}}, \widehat{L}_{\widehat{\mathcal{D}}}\right) \text{ and } \widehat{\Sigma}_{1,\mathcal{M}_2} = {}_P \widehat{\Sigma}_{1,\widehat{\mathcal{D}}}$$

Thus, (17) and (18) are proved. We show (19) by contradiction. If on the contrary $\|\widehat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty = \gamma_n^{1-\eta}$, then in what follows we will show that

$$\begin{aligned} &\ell_n\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}}\right) + \gamma_n \left\|\widehat{S}_{\widehat{\mathcal{D}}}\right\|_{\text{off},1} + \delta_n \left\|\widehat{L}_{\widehat{\mathcal{D}}}\right\|_* + \tau_n \|W_n \odot \widehat{L}_{\widehat{\mathcal{D}}}\|_1 \\ >{}_P &\ell_n\left(S^* + L^*\right) + \gamma_n \left\|S^*\right\|_{\text{off},1} + \delta_n \left\|L^*\right\|_* + \tau_n \|W_n \odot L^*\|_1, \end{aligned} \tag{21}$$

and thus a contradiction is reached. We start with the Taylor expansion of $\ell_n(S+L)$ around S^* and L^*. Since $\widehat{\Sigma} = \frac{1}{n}\widehat{\boldsymbol{R}}^T\widehat{\boldsymbol{R}} = \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}_n - \widetilde{\boldsymbol{C}}(\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}})^{-1}\widetilde{\boldsymbol{C}}^T)\widetilde{\boldsymbol{X}}$ satisfying that $\|\widehat{\Sigma} - \Sigma^*\|_\infty = O_P(\frac{1}{\sqrt{n}})$ for fixed $\boldsymbol{C}$ and $\Sigma^* = (S^* + L^*)^{-1}$, we let $\ell(\Theta) = \lim \mathbb{E}\ell_n(\Theta) = -\log\det(\Theta) + \text{tr}(\Sigma^*\Theta)$. Then we have

$$\ell_n\left(\Theta^* + \Delta\right) = \ell\left(\Theta^*\right) + \frac{1}{2} v(\Delta)^\top \mathcal{I}^* v(\Delta) + R_n(\Delta), \tag{22}$$

where $\Theta^* = S^* + L^*$, and the function $v : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{p^2} \times 1$ is a map that vectorizes a matrix. Moreover, $R_n(\Delta)$ is the remainder term satisfying

$$R_n(\Delta) = R_n\left(\mathbf{0}_{p\times p}\right) + O_P\left(\|\Delta\|_\infty^3\right) + O_P\left(\frac{\|\Delta\|_\infty}{\sqrt{n}}\right) \text{ as } \Delta \to 0, n \to \infty, \tag{23}$$

where $R_n\left(\mathbf{0}_{p\times p}\right) = \ell_n\left(\Theta^*\right) - \ell\left(\Theta^*\right), O_P\left(\|\Delta\|_\infty/\sqrt{n}\right)$ term corresponds to $v\left(\nabla\ell_n\left(\Theta^*\right)\right)^\top v(\Delta)$, and $O_P\left(\|\Delta\|_\infty^3\right)$ characterizes the remainder. Furthermore, from (22), as $\Delta \to 0$, the first derivative of $R_n(\Delta)$ satisfies

$$\begin{aligned} \nabla R_n(\Delta) &= \nabla\ell_n(\Theta^* + \Delta) - \mathcal{I}^* v(\Delta) \\ &= \nabla\ell_n(\Theta^* + \Delta) - \nabla\ell(\Theta^* + \Delta) + \nabla\ell(\Theta^* + \Delta) - \nabla\ell(\Theta^*) + \nabla\ell(\Theta^*) - \mathcal{I}^* v(\Delta) \\ &= O_P(\frac{1}{\sqrt{n}}) + O_P(\|\Delta\|_\infty^2) \text{ as } \Delta \to 0, \end{aligned} \tag{24}$$

where we used the fact $\nabla\ell(\Theta^*) = 0$. In addition, the second derivative satisfies

$$\nabla^2 R_n(\Delta) = \nabla^2\ell_n\left(\Theta^* + \Delta\right) - \mathcal{I}^* = O_P\left(\|\Delta\|_\infty\right). \tag{25}$$

We plug $\Delta = \widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - S^* - L^*$ into (22), then

$$\begin{aligned} \ell_n\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}}\right) - \ell_n\left(S^* + L^*\right) =& \frac{1}{2} v\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - (S^* + L^*)\right)^\top \mathcal{I}^* v\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - (S^* + L^*)\right) \\ &+ R_n\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - (S^* + L^*)\right) - R_n\left(\mathbf{0}_{p\times p}\right). \end{aligned} \tag{26}$$

We first establish a lower bound for $R_n\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - (S^* + L^*)\right) - R_n\left(\mathbf{0}_{p\times p}\right)$. According to Lemma 28, we have

$$\left\|\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}} - (S^* + L^*)\right\|_\infty \leq_P \kappa\gamma_n^{1-\eta},$$

with a possibly different κ. The above display and (23) yield

$$R_n\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right)-R_n\left(\mathbf{0}_{p\times p}\right)=O_P\left(\gamma_n^{3(1-\eta)}\right)+O_P\left(\frac{\gamma_n^{1-\eta}}{\sqrt{n}}\right). \tag{27}$$

We proceed to a lower bound for the term

$$\frac{1}{2}v\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right)^\top\mathcal{I}^*v\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right)$$

on the right-hand side of (26) with the aid of Lemma 29 and Lemma 30.

Lemma 29 *Under Assumption 3, there exists a positive constant ε such that*

$$\|S+L\|_\infty\geq\varepsilon\|L\|_\infty \text{ for all } (S,L)\in\Omega^*\times\mathcal{T}^*.$$

Lemma 30 *Let*

$$\Delta_L=U_1^*\Sigma_1^*\left(\widehat{U}_{1,\mathcal{M}_2}-U_1^*\right)^\top+\left(\widehat{U}_{1,\mathcal{M}_2}-U_1^*\right)\Sigma_1^*U_1^{*\top}+U_1^*\left(\widehat{\Sigma}_{1,\widehat{\mathcal{D}}}-\Sigma_1^*\right)U_1^{*\top},$$

then under Assumption 1 we have (i) $\Delta_L\in\mathcal{T}^*$. (ii) There exists positive constant ε such that $\|\Delta_L\|_\infty>_P\varepsilon\gamma_n^{1-\eta}$, for all $\left\|\widehat{U}_{1,\mathcal{M}_2}-U_1^*\right\|_\infty=\gamma_n^{1-\eta}$. (iii) $\left\|\widehat{L}_{\widehat{\mathcal{D}}}-L^*-\Delta_L\right\|_\infty\leq_P\kappa\gamma_n^{2(1-\eta)}$.

According to Lemma 29 and Lemma 30 (i) (iii) and noticing that $\widehat{S}_{\widehat{\mathcal{D}}}-S^*\in\Omega^*$, we have

$$\left\|\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right\|_\infty\geq_P\left\|\widehat{S}_{\widehat{\mathcal{D}}}-S^*+\Delta_L\right\|_\infty-\kappa\gamma_n^{2(1-\eta)}\geq_P\varepsilon\left\|\Delta_L\right\|_\infty-\kappa\gamma_n^{2(1-\eta)}.$$

According to Lemma 30 (ii), the above display further implies that

$$\left\|\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right\|_\infty>_P\varepsilon\gamma_n^{1-\eta},$$

with a possibly different ε. Since $\mathcal{I}^*$ is positive definite,

$$\begin{aligned}v\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right)^\top&\mathcal{I}^*v\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right)\\&>\varepsilon\left\|\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}-(S^*+L^*)\right\|_\infty^2\geq_P\varepsilon^2\gamma_n^{2(1-\eta)}.\end{aligned} \tag{28}$$

Hence, combing (26), (27) and (28) give

$$\ell_n\left(\widehat{S}_{\widehat{\mathcal{D}}}+\widehat{L}_{\widehat{\mathcal{D}}}\right)-\ell_n\left(S^*+L^*\right)>_P\frac{\varepsilon^2}{2}\gamma_n^{2(1-\eta)}. \tag{29}$$

We proceed to the regularization terms in (21). For the ℓ_1 penalty term for S, we have

$$\left\|\left(\widehat{S}_{\widehat{\mathcal{D}}}\right)\right\|_{\text{off},1}-\|(S^*)\|_{\text{off},1}=\text{sign}\left(\mathbf{O}\left(S^*\right)\right)\cdot\left(\widehat{S}_{\widehat{\mathcal{D}}}-S^*\right)=O_P\left(\gamma_n^{1-\eta}\right).$$

The second equality in the above display is due to Lemma 28. For the nuclear norm term and adaptive ℓ_1 for L, we have also from Lemma 28,

$$\left\|\widehat{L}_{\widehat{\mathcal{D}}}\right\|_* - \|L^*\|_* \geq -\left\|\widehat{L}_{\widehat{\mathcal{D}}} - L^*\right\|_* \geq -\kappa\gamma_n^{1-\eta},$$

and from Lemma 21 and Lemma 28,

$$\left\|W_n \odot \widehat{L}_{\widehat{\mathcal{D}}}\right\|_1 - \|W_n \odot L^*\|_1 = (W_n \odot \operatorname{sign}(L^*)) \cdot \left(\widehat{L}_{\widehat{\mathcal{D}}} - L^*\right) = o_P(r_n^{1-\alpha}\gamma_n^{1-\eta}).$$

Notice that $\delta_n = \rho_1\gamma_n$ and $\tau_n = \rho_2\frac{\gamma_n}{r_n^{1-\alpha}}$. Equations (28), (29) and the above three inequalities imply

$$\begin{aligned}&\ell_n\left(\widehat{S}_{\widehat{\mathcal{D}}} + \widehat{L}_{\widehat{\mathcal{D}}}\right) - \ell_n\left(S^* + L^*\right) + \gamma_n\left(\left\|\left(\widehat{S}_{\widehat{\mathcal{D}}}\right)\right\|_{\text{off},1} - \|(S^*)\|_{\text{off},1}\right) + \delta_n\left(\left\|\widehat{L}_{\widehat{\mathcal{D}}}\right\|_* - \|L^*\|_*\right)\\ &+ \tau_n\left(\left\|W_n \odot \widehat{L}_{\widehat{\mathcal{D}}}\right\|_1 - \|W_n \odot L^*\|_1\right) >_P \varepsilon\gamma_n^{2(1-\eta)} >_P 0,\end{aligned}$$

with a possibly different ε. Notice that from Lemma 28, $\left(\widehat{S}_{\widehat{\mathcal{D}}}, \widehat{L}_{\widehat{\mathcal{D}}}\right) =_P \left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$, so we obtain (21) by rearranging terms in the above inequality, and this contradicts the definition of $\widehat{S}_{\mathcal{M}_2}$ and $\widehat{L}_{\mathcal{M}_2}$. This completes the proof for (19). Thus, $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ is an interior point of $\mathcal{M}_2$. The uniqueness of the solution is obtained according to the following Lemma 31.

Lemma 31 *The solution to the optimization problem (16) is unique with a probability converging to 1. In addition,*

$$\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right) = (S^*, L^*) + \mathbf{G}^{-1}\left(\gamma_n \operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \delta_n U_1^* U_1^{*\top}\right) + o_P\left(\gamma_n\right),$$

as $n \to \infty$.

B.2.3 Step 2

In this section, we first show that $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ is a solution of the optimization problem

$$\begin{aligned}&\min\left\{\ell_n(S+L) + \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1\right\}\\ &\text{subject to} \quad L \succeq 0, \quad S = S^T, \quad (S, L) \in \mathcal{M}_1.\end{aligned} \tag{30}$$

To prove this, we will show that $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ satisfies the first order condition

$$\mathbf{0}_{p\times p} \in \partial_S H|_{\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)} \text{ and } \mathbf{0}_{p\times p} \in \partial_L H|_{\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)},$$

where the function H is the objective function

$$H(S, L) = \ell_n(S+L) + \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1$$

and $\partial_S H$ and $\partial_L H$ denotes the sub-differentials of H. We first derive an explicit expression of the first-order condition. The sub-differential concerning S is defined as

$$\partial_S H|_{\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)} = \left\{\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2} + \widehat{L}_{\mathcal{M}_2}\right) + \gamma_n\left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)\right) + \gamma_n M : \|M\|_\infty \leq 1 \text{ and } M \in \Omega^{*\perp}\right\},$$

where $\Omega^{*\perp}$ is the orthogonal complement space of Ω^* in the space of symmetric matrices. According to Example 2 of Watson (1992), the sub-differential with respect to L is

$$\partial_L H|_{(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2})} = \Big\{\nabla \ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)+\delta_n \widehat{U}_{1,\mathcal{M}_2}\widehat{U}_{1,\mathcal{M}_2}^\top+\tau_n W_n\odot \operatorname{sign}(L^*) \\ +\delta_n \widehat{U}_{2,\mathcal{M}_2}M_1\widehat{U}_{2,\mathcal{M}_2}^\top+\tau_n W_n\odot M_2 : \|M_1\|_2\le 1, \|M_2\|_\infty\le 1, M_2\in \mathcal{T}_2^{*\perp}\Big\}$$

where $\widehat{U}_{1,\mathcal{M}_2}$ has the same zero-block as U_1^*, $\widehat{U}_{2,\mathcal{M}_2}$ is a $p\times(p-r)$ matrix satisfying $\widehat{U}_{1,\mathcal{M}_2}^\top\widehat{U}_{2,\mathcal{M}_2}=\mathbf{0}_{r\times(p-r)}$, $\widehat{U}_{2,\mathcal{M}_2}^\top\widehat{U}_{2,\mathcal{M}_2}=I_{p-r}$ and $\left\|\widehat{U}_{2,\mathcal{M}_2}-U_2^*\right\|_\infty\le\gamma_n^{1-\eta}$. For some (S,L), if

$$\mathcal{P}_{\Omega^*}(S)=\mathbf{0}_{p\times p}, \mathcal{P}_{\Omega^{*\perp}}(S)=\mathbf{0}_{p\times p}, \mathcal{P}_{T(\widehat{L}_{\mathcal{M}_2})}(L)=\mathbf{0}_{p\times p} \text{ and } \mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}(L)=\mathbf{0}_{p\times p}, \tag{31}$$

then $S=\mathbf{0}_{p\times p}$ and $L=\mathbf{0}_{p\times p}$. Consequently, to prove (31), it suffices to show that

$$\mathcal{P}_{\Omega^*}\partial_S H|_{(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2})}=\mathbf{0}_{p\times p} \text{ and } \mathcal{P}_{\mathcal{T}(\widehat{L}_{\mathcal{M}_2})}\partial_L H\Big|_{(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2})}=\mathbf{0}_{p\times p}. \tag{32}$$

and

$$\mathbf{0}_{p\times p}\in \mathcal{P}_{\Omega^{*\perp}}\partial_S H|_{(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2})} \text{ and } \mathbf{0}_{p\times p}\in \mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\partial_L H\Big|_{(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2})}. \tag{33}$$

According to the definition of $\left(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2}\right)$, it is the solution to the optimization (16). In addition, according to the discussion in Step 1, $\left(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2}\right)$ does not lie on the boundary of $\mathcal{M}_2$. Therefore, it satisfies the first order condition of (16), which is equivalent to (32). Thus, to prove 31, it is sufficient to show (33). Lemma 24 establishes a sufficient expression for the above equation. Take gradient on both side of (22) to obtain

$$\nabla \ell_n\left(S^*+L^*+\Delta\right)=\mathcal{I}^* v(\Delta)+\nabla R_n(\Delta).$$

We plug $\Delta=\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}-S^*-L^*$ into the above equation to get

$$\nabla \ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\mathcal{I}^* v\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}-S^*-L^*\right)+\nabla R_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}-S^*-L^*\right). \tag{34}$$

According to Lemma 31,

$$\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}-S^*-L^*=\mathcal{A}\mathbf{G}^{-1}\left(\gamma_n \operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \delta_n U_1^* U_1^{*\top}\right)+o_P\left(\gamma_n\right),$$

where $\mathcal{A}$ is the adding operator of two matrices $\mathcal{A}(A,B)=A+B$. Combining this with (24), (34), and notice that $\delta_n=\rho_1\gamma_n$, we have

$$\nabla \ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\gamma_n\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \rho_1 U_1^* U_1^{*\top}\right)+o_P\left(\gamma_n\right),$$

and consequently,

$$\mathcal{P}_{\Omega^{*\perp}}\nabla \ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\gamma_n\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \rho_1 U_1^* U_1^{*\top}\right)+o_P\left(\gamma_n\right), \tag{35}$$

$$\mathcal{P}_{\mathcal{V}(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\gamma_n\mathcal{P}_{\mathcal{V}^*}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right). \quad (36)$$

$$\mathcal{P}_{\mathcal{T}_2^{*\perp}}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\gamma_n\mathcal{P}_{\mathcal{T}_2^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right). \quad (37)$$

$$\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)=\gamma_n\mathcal{P}_{\mathcal{T}^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right). \quad (38)$$

According to Lemma 24, it suffices to show that

$$\|\mathcal{P}_{\Omega^{*\perp}}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)\|_\infty\le\gamma_n, \quad (39)$$

$$g_{\rho_1,\rho_2,\widehat{L}_{\mathcal{M}_2},\widetilde{W}_n}\left(\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(L)\right)\right)\le\gamma_n. \quad (40)$$

The first inequality is easy to check according to (35) and Assumption 10. We emphasize the second part. Notice that from (36), (37), (38) and Lemma 21, we have

$$\begin{aligned}&\mathcal{P}_{\mathcal{V}(\widehat{L}_{\mathcal{M}_2})}\left(\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right)\\=&\gamma_n\mathcal{P}_{\mathcal{V}^*}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right)+o_P(r_n^{1-\alpha}\frac{\gamma_n}{r_n^{1-\alpha}})\\=&\gamma_n\mathcal{P}_{\mathcal{V}^*}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P(\gamma_n),\end{aligned} \quad (41)$$

and

$$\begin{aligned}&\mathcal{P}_{\mathcal{T}_2^{*\perp}}\left(\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right)\\=&\gamma_n\mathcal{P}_{\mathcal{T}_2^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right)+o_P(r_n^{1-\alpha}\frac{\gamma_n}{r_n^{1-\alpha}})\\=&\gamma_n\mathcal{P}_{\mathcal{T}_2^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P(\gamma_n).\end{aligned} \quad (42)$$

and

$$\begin{aligned}&\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right)\\=&\gamma_n\mathcal{P}_{T^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P\left(\gamma_n\right)+o_P(r_n^{1-\alpha}\frac{\gamma_n}{r_n^{1-\alpha}})\\=&\gamma_n\mathcal{P}_{T^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)+o_P(\gamma_n).\end{aligned} \quad (43)$$

Then according to $\max\{a,b+c\}\le\max\{a,b\}+c$,

$$\begin{aligned}&g_{\rho_1,\rho_2,\widehat{L}_{\mathcal{M}_2},\widetilde{W}_n}\left(\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right)\\\le&g_{\rho_1,\rho_2,\widehat{L}_{\mathcal{M}_2},\Lambda_n}\left(\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right)\\&+\frac{r_n^{1-\alpha}\max_{(i,j)\notin J_1}|(\widetilde{W}_n)_{ij}-(\Lambda_n)_{ij}|}{\rho_2}\left\|\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\nabla\ell_n\left(\widehat{S}_{\mathcal{M}_2}+\widehat{L}_{\mathcal{M}_2}\right)-\tau_n\mathcal{P}_{\mathcal{T}^\perp(\widehat{L}_{\mathcal{M}_2})}\left(W_n\odot\text{sign}(\widehat{L}_{\mathcal{M}_2})\right)\right\|_F\\\triangleq&I+II.\end{aligned}$$

For I, by (41), (42), Assumption 10 and Lemma 26, we have

$$\begin{aligned}I\le&\gamma_ng_{\rho_1,\rho_2,L^*,\Lambda_n}\left(\mathcal{P}_{\mathcal{T}^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)\right)+\left(\frac{1}{\rho_1}+\frac{r_n^{1-\alpha}M_{n2}}{\rho_2}\right)o_P\left(\gamma_n\right)\\\le&\gamma_ng_{\rho_1,\rho_2,L^*,\Lambda_n}\left(\mathcal{P}_{\mathcal{T}^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\text{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right)\right)+o_P(\gamma_n)\\<&\gamma_n+o_P(\gamma_n),\end{aligned}$$

For II, from Assumption 8 and (43), we have

$$II = \frac{r_n^{1-\alpha} \max_{(i,j)\notin J_1} |(\widetilde{W}_n)_{ij} - (\Lambda_n)_{ij}|}{\rho_2} \left\| \gamma_n \mathcal{P}_{\mathcal{T}^{*\perp}} \mathcal{I}^* \mathcal{A} \mathbf{G}^{-1} \left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \rho_1 U_1^* U_1^{*\top} \right) + o_P\left(\gamma_n\right) \right\|_F$$
$$\leq r_n^{1-\alpha} O_P(r_n^{-1}) O_P(\gamma_n) = o_P(\gamma_n),$$

which completes the proof for equation (40). Then with probability tending to 1 as $n \to \infty$,

$$I + II \leq \gamma_n.$$

Thus far we have proved that with probability tending to 1, $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ is a solution to the optimization problem (30). We proceed to the proof of the uniqueness of the solution to (30). Because the objective function $H(S, L)$ is a convex function, it is sufficient to show the uniqueness of the solution in a neighborhood of $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$. Let $d_n = e^{-n} \ll \gamma_n$, we choose a small neighborhood as follows:

$$\begin{aligned}
\mathcal{N} = \Big\{ & (S, L) : S = S_{\Omega^*} + S_{\Omega^{*\perp}}, \left\| S_{\Omega^*} - \widehat{S}_{\mathcal{M}_2} \right\|_\infty < d_n, \|S_{\Omega^{*\perp}}\|_\infty < d_n, S_{\Omega^*} \in \Omega^*, S_{\Omega^{*\perp}} \in \Omega^{*\perp}, \\
& L \text{ has the following eigendecomposition :} \\
& L = [U_1, U_2] \begin{bmatrix} \Sigma_1 & \mathbf{0}_{p\times(p-r)} \\ \mathbf{0}_{(p-r)\times p} & \Sigma_2 \end{bmatrix} [U_1, U_2]^\top, U_1 = U_{11} + U_{11}^\perp, \\
& (U_{11})_{G_{ij}^*} = 0, \text{ for } i \neq j, \quad (U_{11}^\perp)_{G_{ij}^*} = 0, \text{ for } i = j, \quad i, j \in [m], \\
& \left\| U_{11} - \widehat{U}_{1,\mathcal{M}_2} \right\|_\infty < d_n, \left\| U_{11}^\perp \right\|_\infty < d_n, \left\| U_2 - \widehat{U}_{2,\mathcal{M}_2} \right\|_\infty < d_n, \\
& \left\| \Sigma_1 - \widehat{\Sigma}_{1,\mathcal{M}_2} \right\|_\infty < d_n, \|\Sigma_2\|_\infty < d_n. \Big\}.
\end{aligned}$$

The next lemma, together with the uniqueness of the solution to (16) established in Lemma 31, guarantees that $\left(\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2}\right)$ is the unique solution in $\mathcal{N}$.

Lemma 32 *For all $\left(\tilde{S}, \tilde{L}\right) \in \mathcal{N}$, if $\left(\tilde{S}, \tilde{L}\right)$ is a solution to* (30)*, then $\left(\tilde{S}, \tilde{L}\right) = (\widehat{S}_{\mathcal{M}_2}, \widehat{L}_{\mathcal{M}_2})$ with probability tending to 1.*

■

Proof [Proof of Theorem 17]

This proof is similar to Proposition 19 in Ma et al. (2020). Denote $\hat{\mu}_j = \frac{1}{|\{i:\hat{\boldsymbol{t}}(i)=j\}|} \sum_{i:\hat{\boldsymbol{t}}(i)=j} \hat{l}_i$, then from defintion of equation (8),

$$\sum_{j=1}^{m} \sum_{i:\hat{\boldsymbol{t}}(i)=j} \|\hat{l}_i - \hat{\mu}_j\|^2 \leq \sum_{j=1}^{m} \sum_{i:\hat{\boldsymbol{t}}(i)=j} \|\hat{l}_i - \mu_j^*\|^2.$$

Let $S = \{i : \|\hat{\mu}_{\hat{\boldsymbol{t}}(i)} - \mu^*_{\boldsymbol{t}(i)}\| \geq c\}$, then

$$\frac{1}{p}|S| \leq \frac{1}{pc^2}\sum_{i\in S}\|\hat{\mu}_{\hat{\boldsymbol{t}}(i)} - \mu^*_{\boldsymbol{t}(i)}\|^2 \leq \frac{2}{pc^2}\left(\sum_{i=1}^{p}\|\hat{l}_i - \hat{\mu}_{\hat{\boldsymbol{t}}(i)}\|^2 + \|\hat{l}_i - \mu^*_{\boldsymbol{t}(i)}\|^2\right) \leq \frac{4}{pc^2}\sum_{i=1}^{p}\|\hat{l}_i - \mu^*_{\boldsymbol{t}(i)}\|^2$$
$$\leq \frac{4}{pc^2}\sum_{i=1}^{p}\left(\|\hat{l}_i - l^*_i\|^2 + \|\hat{l}_i - \mu^*_{\boldsymbol{t}(i)}\|^2\right) = \frac{4}{pc^2}\left(\|\widehat{L} - L^*\|_F^2 + \sum_{i=1}^{p}\|\hat{l}_i - \mu^*_{\boldsymbol{t}(i)}\|^2\right)$$

In addition, similar to the proof of Proposition 19 in Ma et al. (2020), we have $H(\hat{\boldsymbol{t}}, \pi(\boldsymbol{t})) \leq \frac{C\omega}{p}|S|$, which completes the proof. ■

Proof [Proof of Corollary 19] For any $x, y \neq 0 \in \mathbb{R}^p$, note that $\mathrm{cor}(x, y) = \frac{\sum_{i=1}^{p}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{p}(y_i-\bar{y})^2}}$ with $\bar{x} = \frac{\sum_i x_i}{p}, \bar{y} = \frac{\sum_i y_i}{p}$. Then for $k = l$,

$$\mathrm{cor}(\mathrm{abs}(l^*_i), \mathrm{abs}(l^*_j)) = \frac{\frac{1}{p}\sum_{h=1}^{d_k}|u_{ki}u_{kj}|u^2_{kh} - |u_{ki}u_{kj}|\bar{u}_k^2}{|u_{ki}u_{kj}|\left(\frac{1}{p}\sum_{h=1}^{d_k}u^2_{kh} - (\bar{u}_k)^2\right)} = 1.$$

For $k \neq l$, from the orthogonal property of l^*_k, l^*_l, it holds that

$$\mathrm{cor}(\mathrm{abs}(l^*_i), \mathrm{abs}(l^*_j)) = \frac{-|u_{ki}u_{kj}|\bar{u}_k\bar{u}_l}{|u_{ki}u_{kj}|\sqrt{\frac{1}{p}\sum_{h=1}^{d_k}u^2_{kh} - (\bar{u}_k)^2}\sqrt{\frac{1}{p}\sum_{h=1}^{d_l}u^2_{lh} - (\bar{u}_l)^2}}$$
$$= \frac{-\bar{u}_k\bar{u}_l}{\sqrt{\frac{1}{p}\sum_{h=1}^{d_k}u^2_{kh} - (\bar{u}_k)^2}\sqrt{\frac{1}{p}\sum_{h=1}^{d_l}u^2_{lh} - (\bar{u}_l)^2}}.$$

Furthermore, according to Theorem 13, continuity of $\mathrm{Cor}(\mathrm{abs}(\cdot))$ and continuous mapping theorem, it holds that $\|\mathrm{Cor}(\mathrm{abs}(\widehat{L})) - \mathrm{Cor}(\mathrm{abs}(L^*))\|_F = o_P(1)$. Combining the above equality and Theorem 17 completes the proof. ■

B.3 Proofs of Lemmas

Proof [Proof of Lemmas] The proof is similar to Lemma 2 in Huang et al. (2008). Since $M_{n1} = o\left(r_n^{1-\alpha}\right)$, and $\max_{(i,j)\in J_1}|(\widetilde{W}_n)_{ij}|/|(\Lambda_n)_{ij}| - 1| \leq M_{1n}O_P\left(1/r_n\right) = o_P(1)$ by the r_n-consistency of $(\widetilde{W}_n)_{ij}$. Thus, $\|\mathbf{s}_{n1}\| = (1 + o_P(1))\, M_{n1} = o_P(r_n^{1-\alpha})$. ■

Proof [Proof of Lemma 23] (1). The first part is obvious by triangular inequality and the uniqueness decomposition of the direct sum. (2). According to Lemma 22 and Proposition 2, we have

$$\|(\mathcal{P}_{\mathcal{V}(L)} - \mathcal{P}_{\mathcal{V}^*})(M)\|_2 \lesssim \|L - L^*\|_2\|M\|.$$

■

Proof [Proof of Lemma 24] The fist part follows from expressions of subgradients for $\|\cdot\|_1, \|\cdot\|_*$ in (Candes and Recht, 2013) and the additivity of the subgradient. Second, from the condition, we can rewrite $Z = \delta_n UV^T + \tau_n W_n \odot \text{sign}(L) + F$, where $F = F_1 + W_n \odot F_2, F_1 \in V \subseteq \mathcal{T}_1(^{\perp}(L), F_2 \in \mathcal{T}_2^{\perp}(L)$ and $g^L_{\rho_1,\rho_2,\widetilde{W}_n}(F) \leq \gamma_n$. It implies $\|F_1\|_2 \leq \delta_n$, $\|F_2\|_\infty \leq \tau_n$. Then $Z \in \partial(\delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1)$. ■

Proof [Proof of Lemma 25] The proof is omitted here by referring to Lemma 1 in Chen et al. (2016). ■

Proof [Proof of Lemma 26] The constants specified in this lemma are as follows,

$$l_n = \frac{1}{\rho_1 + \rho_2 \frac{p}{m_{n2} r_n^{1-\alpha}}}, \quad u_n = \frac{1}{\rho_1} + \frac{r_n^{1-\alpha} M_{n2}}{\rho_2},$$

where $m_{2n} \triangleq \min_{(i,j)\notin J_1}(\Lambda_n)_{ij}$ satisfying $m_{n2} r_n^{1-\alpha} \leq M_{n2} r_n^{1-\alpha} = O_P(1)$ as $n \to \infty$.

First, note that for any given smooth point $L \in \mathcal{LS}(m, r)$ with $\mathcal{T}^{\perp}(L) = \mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp}$, $L' \in \mathcal{T}^{\perp}(L)$ has the orthogonal direct sum decomposition $L' = L_1 \oplus L_2$ with $L_1 = \mathcal{P}_{\mathcal{V}(L)}(L'), L_2 = \mathcal{P}_{\mathcal{T}_2^{*\perp}}(L')$. Then

$$\|L'\|_F^2 = \|L_1 + L_2\|_F^2 = \|L_1\|_F^2 + \|L_2\|_F^2.$$

Second, we prove the norm inequality. The right part of the inequality is a corollary of the above inequality for L. For the left part, note that for any $a \in [0, 1]$,

$$\begin{aligned} g_{\rho_1,\rho_2,L,\Lambda_n}(L') &\geq \frac{a}{\rho_1}\|L_1\|_2 + \frac{(1-a) r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot L_2\|_\infty \\ &\geq \frac{a}{\rho_1}\|L_1\|_2 + \frac{(1-a) r_n^{1-\alpha} m_{n2}}{\rho_2 p}\|L_2\|_2. \end{aligned}$$

Take $a = \frac{m_{n2} r_n^{1-\alpha} \rho_1}{p\rho_2 + m_{n2} r_n^{1-\alpha} \rho_1}$, then $g^L_{\rho_1,\rho_2,\Lambda_n}(L') \geq l_n \|L\|_2$, which concludes the proof. ■

Proof [Proof of Lemma 28] We consider the first-order condition for the optimization problem (20). Notice that Ω^* and $\widehat{\mathcal{D}}$ are linear spaces, so the first order condition becomes

$$\mathbf{0}_{p\times p} \in \mathcal{P}_{\Omega^*} \partial_S H\big|_{(S,L)} \text{ and } \mathbf{0}_{p\times p} \in \mathcal{P}_{\widehat{\mathcal{D}}} \partial_L H\big|_{(S,L)},$$

where H is defined as

$$H(S, L) = \ell_n(S + L) + \gamma_n\|S\|_{\text{off},1} + \delta_n\|L\|_* + \tau_n\|W_n \odot L\|_1.$$

We will show that there is a unique $(S, L) \in \Omega^* \times \widehat{\mathcal{D}}$ satisfying the first order condition. Because of the convexity of the optimization problem (20), it suffices to show that with a probability converging to 1, there is a unique $(S, L) \in \mathcal{B}$ satisfying the first order condition, where

$$\mathcal{B} = \left\{(S, L) \in \Omega^* \times \widehat{\mathcal{D}} : \|S - S^*\|_\infty \leq \gamma_n^{1-\eta}, \text{ and } \|L - L^*\|_\infty \leq \gamma_n^{1-\eta}\right\}.$$

We simplify the first order condition for $(S, L) \in \mathcal{B}$. For the ℓ_1 penalty term for S, if $\|S - S^*\|_\infty \leq \gamma_n^{1-\eta}$ and $S \in \Omega^*$, then $\|S\|_{\text{off},1}$ is smooth on Ω^* and

$$\mathcal{P}_{\Omega^*} \partial_S \|S\|_{\text{off},1} = \text{sign}\left(\mathbf{O}\left(S^*\right)\right) \text{ for } S \in \Omega^*.$$

Similarly, for $L \in \widehat{\mathcal{D}}$ and $\|L - L^*\|_\infty \leq \gamma_n^{1-\eta}$, $\|L\|_*, \|W_n \odot L\|_1$ is smooth over the linear space $\widehat{\mathcal{D}}$ and

$$\mathcal{P}_{\widehat{\mathcal{D}}} \partial_L \|L\|_* = \widehat{U}_{1,\mathcal{M}_2} \widehat{U}_{1,\mathcal{M}_2}^\top, \text{ for } L \in \widehat{\mathcal{D}},$$

and

$$\mathcal{P}_{\widehat{\mathcal{D}}} \partial_L \|W_n \odot L\|_1 = \mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \text{sign}(L^*)).$$

Combining the above two equations with the $\nabla \ell_n$ term, we arrive at an equivalent form of the first order condition, that is, there exists $(S, L) \in \mathcal{B}$ satisfying

$$\mathcal{P}_{\Omega^*} \nabla \ell_n(S + L) + \gamma_n \text{sign}\left(\mathbf{O}\left(S^*\right)\right) = \mathbf{0}_{p \times p},$$
$$\mathcal{P}_{\widehat{\mathcal{D}}} \nabla \ell_n(S + L) + \delta_n \widehat{U}_{1,\mathcal{M}_2} \widehat{U}_{1,\mathcal{M}_2}^\top + \tau_n \mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \text{sign}(L^*)) = \mathbf{0}_{p \times p}.$$

We will show the existence and uniqueness of the solution to the above equations using the contraction mapping theorem. We first construct the contraction operator. Let $(S, L) = (S^* + \Delta_S, L^* + \Delta_L)$. We plug (34) into the above equations, and arrive at their equivalent ones

$$\begin{aligned} &\mathcal{P}_{\Omega^*} \mathcal{I}^*\left(\Delta_S + \Delta_L\right) + \mathcal{P}_{\Omega^*} \nabla R_n\left(\Delta_S + \Delta_L\right) + \gamma_n \text{sign}\left(\mathbf{O}\left(S^*\right)\right) = \mathbf{0}_{p \times p}, \\ &\mathcal{P}_{\widehat{\mathcal{D}}} \mathcal{I}^*\left(\Delta_S + \Delta_L\right) + \mathcal{P}_{\widehat{\mathcal{D}}} \nabla R_n\left(\Delta_S + \Delta_L\right) + \delta_n \widehat{U}_{1,\mathcal{M}_2} \widehat{U}_{1,\mathcal{M}_2}^\top + \tau_n \mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \text{sign}(L^*)) = \mathbf{0}_{p \times p}. \end{aligned} \tag{44}$$

We define an operator $\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}} : \Omega^* \times \left(\widehat{\mathcal{D}} - L^*\right) \to \Omega^* \times \widehat{\mathcal{D}}$,

$$\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}}\left(\Delta_S, \Delta_L\right) = \left(\mathcal{P}_{\Omega^*} \mathcal{I}^*\left(\Delta_S + \Delta_L\right), \mathcal{P}_{\widehat{\mathcal{D}}} \mathcal{I}^*\left(\Delta_S + \Delta_L\right)\right),$$

where the set $\widehat{\mathcal{D}} - L^* = \left\{L - L^* : L \in \widehat{\mathcal{D}}\right\}$. We further transform equation (44) to

$$\begin{aligned} &\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}}\left(\Delta_S, \Delta_L\right) + \left(\mathcal{P}_{\Omega^*} \nabla R_n\left(\Delta_S + \Delta_L\right), \mathcal{P}_{\widehat{\mathcal{D}}} \nabla R_n\left(\Delta_S + \Delta_L\right)\right) + \Big(\gamma_n \text{sign}\left(\mathbf{O}\left(S^*\right)\right), \delta_n \widehat{U}_{1,\mathcal{M}_2} \widehat{U}_{1,\mathcal{M}_2}^\top \\ &+ \tau_n \mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \text{sign}(L^*))\Big) = \left(\mathbf{0}_{p \times p}, \mathbf{0}_{p \times p}\right). \end{aligned} \tag{45}$$

Notice that the projection $\mathcal{P}_{\widehat{\mathcal{D}}}$ is uniquely determined by the matrix $\widehat{U}_{1,\mathcal{M}_2}$. Lemma 22 states that the mapping $\widehat{U}_{1,\mathcal{M}_2} \to \mathcal{P}_{\widehat{\mathcal{D}}}$ is Lipschitz. Under Assumption 3, similar to Lemma 25, we have that $\tilde{\mathbf{G}}_{\mathcal{D}^*}$ is invertible, where we define $\tilde{\mathbf{G}}_{\mathcal{D}^*} : \Omega^* \times \mathcal{D}^* \to \Omega^* \times \mathcal{D}^*$,

$$\tilde{\mathbf{G}}_{\mathcal{D}^*}\left(S', L'\right) = \left(\mathcal{P}_{\Omega^*}\left\{\mathcal{I}^*\left(S' + L'\right)\right\}, \mathcal{P}_{\mathcal{D}^*}\left\{\mathcal{I}^*\left(S' + L'\right)\right\}\right), \text{ for } S' \in \Omega^* \text{ and } L' \in \mathcal{D}^*,$$

and $\mathcal{D}^* = \left\{U_1^* \Sigma_1' U_1^{*\top} : \Sigma_1' \text{ is a } r \times r \text{ diagonal matrix }\right\}$ ($\mathcal{D}^* - L^* = \mathcal{D}^*$). According to the invertibility of $\tilde{\mathbf{G}}_{\mathcal{D}^*}$, Lemma 22 and the fact that $\left\|\widehat{U}_{1,\mathcal{M}_2} - U_1^*\right\|_\infty \leq \gamma_n^{1-\eta}$, we know that $\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}}$ is also invertible over $\Omega^* \times \left(\widehat{\mathcal{D}} - L^*\right)$ and is Lipschitz in $\widehat{U}_{1,\mathcal{M}_2}$ for sufficiently large n. We apply $\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}}^{-1}$ on both sides of (45) and transform it to a fixed point problem,

$$\left(\Delta_S, \Delta_L\right) = \mathbf{C}\left(\Delta_S, \Delta_L\right),$$

where the operator $\mathbf{C}$ is defined by

$$\begin{aligned}\mathbf{C}\left(\Delta_S, \Delta_L\right) \triangleq & -\tilde{\mathbf{G}}_{\widehat{\mathcal{D}}}^{-1}\left(\left(\mathcal{P}_{\Omega^*} \nabla R_n\left(\Delta_S+\Delta_L\right), \mathcal{P}_{\widehat{\mathcal{D}}} \nabla R_n\left(\Delta_S+\Delta_L\right)\right)\right. \\ & \left.+\left(\gamma_n \operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \tau_n \mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \operatorname{sign}(L^*))+\delta_n \widehat{U}_{1, \mathcal{M}_2} \widehat{U}_{1, \mathcal{M}_2}^{\top}\right)\right) .\end{aligned} \tag{46}$$

Define the set $\mathcal{B}^* = \mathcal{B} - (S^*, L^*) = \{(S - S^*, L - L^*) : (S, L) \in \mathcal{B}\}$. We will show that with a probability converging to 1, $\mathbf{C}$ is a contraction mapping over $\mathcal{B}^*$.

First, according to (24), $\delta_n = \rho_1 \gamma_n$ and the definition of set $\mathcal{B}$, for sufficiently large n, we have

$$\begin{aligned}&\left\|\left(\mathcal{P}_{\Omega^*} \nabla R_n\left(\Delta_S+\Delta_L\right)+\gamma_n \operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right), \mathcal{P}_{\widehat{\mathcal{D}}} \nabla R_n\left(\Delta_S+\Delta_L\right)+\delta_n \widehat{U}_{1, \mathcal{M}_2} \widehat{U}_{1, \mathcal{M}_2}^{\top}\right)\right\|_{\infty} \\ \leq & \kappa\left(\left\|\Delta_S\right\|+\left\|\Delta_L\right\|+\frac{1}{\sqrt{n}}+\gamma_n\right) \leq O_P\left(\gamma_n\right),\end{aligned} \tag{47}$$

with possibly a different κ. In addtion, according to $\tau_n = \rho_2 \frac{\gamma_n}{r_n^{1-\alpha}}$ and Lemma 21,

$$\begin{aligned}\tau_n\|\mathcal{P}_{\widehat{\mathcal{D}}}(W_n \odot \operatorname{sign}(L^*))\|_\infty &\leq \tau_n\|(\mathcal{P}_{\widehat{\mathcal{D}}} - \mathcal{P}_{\mathcal{D}^*})(W_n \odot \operatorname{sign}(L^*))\|_\infty + \tau_n\|\mathcal{P}_{\mathcal{D}^*}(W_n \odot \operatorname{sign}(L^*))\|_\infty \\ &\leq \kappa\left(\tau_n\|\widehat{U}_1 - U_1^*\|_\infty\|W_n \odot \operatorname{sign}(L^*)\|_\infty + \tau_n\|W_n \odot \operatorname{sign}(L^*)\|_\infty\right) \\ &= o_P((\gamma_n+1) r_n^{1-\alpha}\tau_n) = o_P(\gamma_n).\end{aligned}$$

Then similar to the reason in the proof of Lemma 8, (20) has a unique solution in $\mathcal{B}^*$ with a probability converging to 1, which concludes our proof. ■

Proof [Proof of Lemma 22, Lemma 29 and Lemma 30, Lemma 31] The proof is similar to Lemma 8, Lemma 3, Lemma 4, and Lemma 5 in Chen et al. (2016) with the aid of Assumption 1 and Lemma 21, which are thus omitted here. Details can be found in the appendix of Shi et al. (2024). ■

Proof [Proof of Lemma 32] For any solution $(\tilde{S}, \tilde{L}) \in \mathcal{N}$, from definition of $\mathcal{N}$, we can rewrite $\tilde{S} = \tilde{S}_{\mathcal{M}_2} + \tilde{S}_{\Omega^{*\perp}}, \quad \tilde{L} = \tilde{L}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{T}^\perp(\tilde{L}_{\mathcal{M}_2})}$, where

$$\tilde{S}_{\mathcal{M}_2} \in \Omega^*, \quad \tilde{S}_{\Omega^{*\perp}} \in \Omega^{*\perp}, \quad \tilde{L}_{\mathcal{M}_2} = \tilde{U}_{1,\mathcal{M}_2}\tilde{\Sigma}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^T \in \mathcal{T}(\tilde{L}_{\mathcal{M}_2}), \quad \tilde{L}_{\mathcal{T}^\perp(\tilde{L}_{\mathcal{M}_2})} \in \mathcal{T}^\perp(\tilde{L}_{\mathcal{M}_2}),$$

$$\|\tilde{S}_{\mathcal{M}_2}\|_\infty \leq O(d_n), \quad \|\tilde{L}_{\mathcal{T}^\perp(\tilde{L}_{\mathcal{M}_2})}\|_\infty \leq O(d_n).$$

Notice that $\left(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}\right) \in \mathcal{M}_2$, we have the sub-differentials of $H(S, L)$ at $\left(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}\right)$,

$$\partial_S H|_{(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})} = \left\{\nabla \ell_n\left(\tilde{S}_{\mathcal{M}_2}+\tilde{L}_{\mathcal{M}_2}\right)+\gamma_n \operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)+\gamma_n M_1 : \|M_1\|_\infty \leq 1, \text{ and } M_1 \in \Omega^{*\perp}\right\},$$

$$\begin{aligned}\partial_L H|_{(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})} = & \left\{\nabla \ell_n\left(\tilde{S}_{\mathcal{M}_2}+\tilde{L}_{\mathcal{M}_2}\right)+\delta_n \tilde{U}_{1, \mathcal{M}_2} \tilde{U}_{1, \mathcal{M}_2}^{\top}+\tau_n W_n \odot \operatorname{sign}(L^*)\right. \\ & \left.+\delta_n M_2+\tau_n W_n \odot M_3 : \|M_2\|_2 \leq 1, M_2 \in \mathcal{T}_1^{\perp}(\tilde{L}_{\mathcal{M}_2}), \|M_3\|_\infty \leq 1, M_3 \in \mathcal{T}_2^{*\perp}\right\}\end{aligned}$$

According to the definition of sub-differential, we have

$$\begin{aligned}
&H(\tilde{S},\tilde{L})-H\left(\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}\right)\\
\geq&\nabla\ell_n\left(\tilde{S}_{\mathcal{M}_2}+\tilde{L}_{\mathcal{M}_2}\right)\cdot\left(\tilde{S}_{\Omega^{*\perp}}+\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\right)+\left(\gamma_n\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)+\gamma_n M_1\right)\cdot\tilde{S}_{\Omega^{*\perp}}\\
&+\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\cdot\left(\delta_n\widehat{U}_{1,\mathcal{M}_2}\widehat{U}_{1,\mathcal{M}_2}^{\top}+\tau_n W_n\odot\operatorname{sign}(L^*)+\delta_n M_2+\tau_n W_n\odot M_3\right)\\
\geq&\mathcal{P}_{\Omega^{*\perp}}\left[\nabla\ell_n\left(\tilde{S}_{\mathcal{M}_2}+\tilde{L}_{\mathcal{M}_2}\right)\right]\cdot\tilde{S}_{\Omega^{*\perp}}+\mathcal{P}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\left[\nabla\ell_n\left(\tilde{S}_{\mathcal{M}_2}+\tilde{L}_{\mathcal{M}_2}\right)\right]\cdot\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\\
&+\left(\gamma_n\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)+\gamma_n M_1\right)\cdot\tilde{S}_{\Omega^{*\perp}}+\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\cdot\left(\delta_n\widehat{U}_{1,\mathcal{M}_2}\widehat{U}_{1,\mathcal{M}_2}^{\top}+\tau_n W_n\odot\operatorname{sign}(L^*)\right.\\
&\left.+\delta_n M_2+\tau_n W_n\odot M_3\right),
\end{aligned}$$

due to $\tilde{S}_{\Omega^{*\perp}}\in\Omega^{*\perp}$ and $\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\in\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})$. Denote

$$N_1\triangleq\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right),N_2\triangleq\mathcal{P}_{\mathcal{T}^{*\perp}}\mathcal{I}^*\mathcal{A}\mathbf{G}^{-1}\left(\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right),\rho_1U_1^*U_1^{*\top}\right),$$

where we can decompose $N_2=P\oplus Q\in\mathcal{T}^{*\perp}=\mathcal{V}^*\oplus\mathcal{T}_2^{*\perp},P\in\mathcal{V}^*,Q\in\mathcal{T}_2^{*\perp}$. Then, plug them into the above equation, we have

$$\begin{aligned}
&H(\tilde{S},\tilde{L})-H\left(\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}\right)\\
\geq&\left(\gamma_nN_1+o_P\left(\gamma_n\right)\right)\cdot\tilde{S}_{\Omega^{*\perp}}+\left(\gamma_n(P+Q)+o_P(\gamma_n)\right)\cdot\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}+\left(\gamma_n\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)+\gamma_nM_1\right)\cdot\tilde{S}_{\Omega^{*\perp}}\\
&+\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\cdot\left(\delta_n\tilde{U}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^{\top}+\tau_nW_n\odot\operatorname{sign}(L^*)+\delta_nM_2+\tau_nW_n\odot M_3\right)\\
=&\gamma_nN_1\cdot\tilde{S}_{\Omega^{*\perp}}+\gamma_n(P+Q)\cdot\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\\
&+\gamma_nM_1\cdot\tilde{S}_{\Omega^{*\perp}}+\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\cdot\left(\delta_n\tilde{U}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^{\top}+\tau_nW_n\odot\operatorname{sign}(L^*)+\delta_nM_2+\tau_nW_n\odot M_3\right)+o_P(\gamma_nD_n).
\end{aligned}\tag{48}$$

where we use the fact that $\operatorname{sign}\left(\mathbf{O}\left(S^*\right)\right)\cdot\tilde{S}_{\Omega^{*\perp}}=0$ and define $D_n=\max\{\|\tilde{S}_{\Omega^{*\perp}}\|_\infty,\|\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}\|_\infty\}$.

Denote $A=\tilde{S}_{\Omega^{*\perp}},B=\tilde{L}_{\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})}$, and let $\mathcal{P}_{\mathcal{V}}B=\tilde{U}\tilde{D}\tilde{U}^T$, where $\tilde{U}$ is the orthogonal basis in linear space V with $\mathcal{T}^{\perp}(\tilde{L}_{\mathcal{M}_2})=\mathcal{V}\oplus\mathcal{T}_2^{*\perp}$. Take $M_1=\operatorname{sign}(\tilde{S}_{\Omega^{*\perp}})$, $M_2=\tilde{U}\operatorname{sign}(\tilde{D})\tilde{U}^T$, $M_3=\operatorname{sign}(\mathcal{P}_{\mathcal{T}_2^{*\perp}}B)$, and let $C=\delta_n\tilde{U}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^{\top},D=\tau_nW_n\odot\operatorname{sign}(L^*),E=\delta_nM_2,F=\tau_nW_n\odot M_3$, we have

$$P\cdot B=P\cdot(\mathcal{P}_{\mathcal{V}^*}B),\quad Q\cdot B=Q\cdot(\mathcal{P}_{\mathcal{T}_2^{*\perp}}B).$$

and

$$M_1\cdot A=\|A\|_1,\quad B\cdot C=0,\quad B\cdot D=0,\quad B\cdot E=\delta_n\|\mathcal{P}_{\mathcal{V}}B\|_*,\quad B\cdot F=\tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1.$$

In addition, according to Assumption 8, Assumption 10, the above expression and dual inequality, we have

$$\begin{aligned}
&\gamma_n\left|N_1\cdot A\right|\leq\gamma_n\|N_1\|_\infty\|A\|_1<\gamma_n\|A\|_1,\\
&\gamma_n\left|P\cdot B\right|=\gamma_n\left|P\cdot\mathcal{P}_{\mathcal{V}^*}B\right|\leq\gamma_n\|P\|_2\|\mathcal{P}_{\mathcal{V}^*}B\|_*<\gamma_n\rho_1\|\mathcal{P}_{\mathcal{V}^*}B\|_*=\delta_n\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*,
\end{aligned}$$

and

$$\begin{aligned}\gamma_n|Q\cdot B| &= \gamma_n|(\widetilde{W}_n\odot Q)\cdot(W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B)| \le \gamma_n\|\widetilde{W}_n\odot Q\|_\infty\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1\\ &\le \gamma_n\left(\|(\widetilde{W}_n-\Lambda_n)\odot Q\|_\infty+\|\Lambda_n\odot Q\|_\infty\right)\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1\\ &< \gamma_n\left(O_P(\frac{1}{r_n})+\frac{\rho_2}{r_n^{1-\alpha}}\right)\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1 < (\tau_n+o_P(\tau_n))\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1\\ &<_P \tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1.\end{aligned}$$

i.e., there exists $0<c_1,c_2,c_3<1$ such that $\gamma_n\left|N_1\cdot A\right|\le c_1\gamma_n\left\|A\right\|_1,\gamma_n\left|P\cdot B\right|\le c_2\delta_n\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*,\gamma_n|Q\cdot B|\le c_3\tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1$. Therefore, combining above inequalities with (48) we get

$$\begin{aligned}H(\tilde{S},\tilde{L})-H\left(\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}\right) \ge_P&\gamma_n(1-c_1)\|A\|_1+\delta_n(\left\|\mathcal{P}_V B\right\|_*-c_2\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*)\\ &+(1-c_3)\tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1+o_P(D_n\gamma_n)\\ \ge_P&\gamma_n(1-c_1)\|A\|_1+(1-c_3)\tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1\\ &+\delta_n\left((1-c_2)\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*-c_2\left\|(\mathcal{P}_V-\mathcal{P}_{\mathcal{V}^*})B\right\|_*\right)+o_P(D_n\gamma_n)\\ \ge_P&\gamma_n(1-c_1)\|A\|_1+\delta_n(1-c_2)\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*\\ &+(1-c_3)\tau_n\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1+o_P(D_n\gamma_n)\\ >_P&0,\end{aligned}$$

provided $\|A\|_1>0$ or $\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*>0$ or $\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1>0$, where the third inequality is according to Lemma 23,

$$\delta_n\left\|(\mathcal{P}_{\mathcal{V}}-\mathcal{P}_{\mathcal{V}^*})B\right\|_*\lesssim\kappa\gamma_n\|\tilde{L}_{\mathcal{M}_2}-L^*\|\|B\|=o_P(D_n\gamma_n)$$

with possibly different κ. Moreover, according to Lemma 31

$$H(\tilde{S},\tilde{L})>_P H\left(\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}\right)>_P H\left(\widehat{S}_{\mathcal{M}_2},\widehat{L}_{\mathcal{M}_2}\right)$$

provided $\|A\|_1>0$ or $\left\|\mathcal{P}_{\mathcal{V}^*}B\right\|_*>0$ or $\|W_n\odot\mathcal{P}_{\mathcal{T}_2^{*\perp}}B\|_1>0$. Since $(\tilde{S},\tilde{L})$ is a solution to (30), the above statement implies $A=B=\mathbf{0}_{p\times p}$, i.e., $\tilde{S}_{\Omega^{*\perp}}=\tilde{L}_{\mathcal{T}^\perp(\tilde{L}_{\mathcal{M}_2})}=\mathbf{0}_{p\times p}$. Therefore, $\tilde{S}=_P\widehat{S}_{\mathcal{M}_2},\tilde{L}=_P\widehat{L}_{\mathcal{M}_2}$, and $(\tilde{S},\tilde{L})\in\mathcal{M}_2$, which completes the proof. ■

Appendix C. Supplementary experiment

In this section, we demonstrate that the locally identifiable neighborhood of the true model is sufficiently large through experiments. We focus on grouped latent variable graphical models (GMGLV) for illustration. Specifically, we design three scenarios to investigate the impact of different initial values of the proposed algorithm on the estimation results. These scenarios are: GMGLV with increased eigenvalues of L, GMGLV with increased sparsity of S, and GMGLV with increased sample size.

1. **GMGLV with increased eigenvalues of L:** The data generation mechanism is similar to that described in Section 5.1, with the key difference being the specification of $\Theta_{O,H}$. Specifically, we set $(\Theta_{O,H_1})_{N_1} = \lambda \boldsymbol{u}_1$, $(\Theta_{O,H_2})_{N_2} = (\lambda - 0.4)\boldsymbol{u}_2$, and $(\Theta_{O,H_3})_{N_3} = (\lambda - 1.1)\boldsymbol{u}_3$, while enforcing $(\Theta_{O,H_1})_{N_2 \cup N_3} = 0$, $(\Theta_{O,H_2})_{N_1 \cup N_3} = 0$, and $(\Theta_{O,H_3})_{N_1 \cup N_2} = 0$. Here, $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3 \in \mathbb{R}^{15}$ are unit vectors. We vary $\lambda \in \{4.0, 4.1, 4.2\}$ to reflect increasing eigenvalues of L.

2. **GMGLV with increased sparsity of S:** The data generation mechanism is similar to that described in Section 5.1, and we fix $\lambda = 4.2$ as in the first scenario. The key difference lies in the number of nonzero elements in Θ_O, denoted as s_0. In the first example, we generate Θ_O such that $\theta^O_{i,i+2} = \theta^O_{i+2,i} \sim \text{Uniform}\left((-2, -1.5) \cup (1.5, 2)\right)$ independently within N_1 for $i = 1, \ldots, 13$. In the second example, in addition to N_1, we also generate $\theta^O_{i,i+2} = \theta^O_{i+2,i} \sim \text{Uniform}\left((-1.8, -1.0) \cup (1.0, 1.8)\right)$ independently within N_2 for $i = 16, \ldots, 28$. In the third example, we further include N_3 by generating $\theta^O_{i,i+2} = \theta^O_{i+2,i} \sim \text{Uniform}\left((-1.6, -0.8) \cup (0.8, 1.6)\right)$ independently within N_3 for $i = 31, \ldots, 43$. We define the sparsity levels of these three examples as high ($s_0 = 26$), medium ($s_0 = 52$), and low ($s_0 = 78$), respectively.

3. **GMGLV with increased sample size:** The data generation mechanism is similar to that described in Section 5.1, and we fix $\lambda = 4.2$ as in the first scenario. We vary the sample size with $n = 500, 1000, 2000$ to evaluate the effect of increased sample size on the performance of the proposed method.

In addition, for each scenario, we design five different initial value settings to evaluate the algorithm's stability and its impact on the local identifiability of the parameters. Specifically, the initial values are generated as follows:

$$
\begin{aligned}
&S_{\text{pert}} \sim \mathcal{N}(0, I_{p\times p}), \quad S_{\text{pert}} = \frac{S_{\text{pert}} + S_{\text{pert}}^T}{2}, \\
&L_{\text{pert}} \sim \mathcal{N}(0, I_{p\times p}), \quad L_{\text{pert}} = \frac{L_{\text{pert}} + L_{\text{pert}}^T}{2}, \\
&S_{\text{pert}} \leftarrow \frac{S_{\text{pert}}}{\text{norm}_{\text{pert}}}, \quad L_{\text{pert}} \leftarrow \frac{L_{\text{pert}}}{\text{norm}_{\text{pert}}}, \\
&S_{\text{init}} = S^* + d \cdot \text{norm}_{\text{total}} \cdot S_{\text{pert}}, \\
&L_{\text{init}} = L^* + d \cdot \text{norm}_{\text{total}} \cdot L_{\text{pert}},
\end{aligned}
$$

where $\text{norm}_{\text{pert}} = \sqrt{\|S_{\text{pert}}\|_F^2 + \|L_{\text{pert}}\|_F^2}$ and $\text{norm}_{\text{total}} = \sqrt{\|S^*\|_F^2 + \|L^*\|_F^2}$. We set $d \in \{0.2, 0.5, 1.0, 2.0, 3.0\}$ to represent the distance between the initial value and the ground truth, and examine how this distance influences the convergence behavior of the proposed algorithm. In the presentation of results, in addition to the evaluation metrics such as $\text{TR}_L, \text{TP}_L, \text{FP}_L, \text{TP}_S, \text{FP}_S$ introduced in Section 5.1, we also include the total estimation error metric, which is defined as follows:

$$
\text{Err} = \frac{\sqrt{\|\widehat{S} - S^*\|_F^2 + \|\widehat{L} - L^*\|_F^2}}{\sqrt{\|S^*\|_F^2 + \|L^*\|_F^2}}.
$$

For each scenario and each initial value setting, we conduct 50 simulation runs and record the estimated mean values and standard errors. All the results are recorded in Tables 5-7. Tables 5-7 show that, for each set of experiments, as the initial values become increasingly distant from the ground truth, all evaluation metrics for the estimation of S and L exhibit only minor changes. This indicates that our algorithm is stable and not sensitive to initialization, and thus can reliably converge to a neighborhood near the true values. These results suggest that the locally identifiable neighborhood of the true model is large. Moreover, Tables 5-6 show that when the eigenvalues of L increase (the signal strengths of the nonzero elements in L and S become more comparable) and the sparsity level of S decreases, the estimation performance for both components tends to deteriorate. Nevertheless, the overall degradation remains modest. Meanwhile, Table 7 shows that as the sample size increases, the estimation performance for both L and S improves accordingly.

When the model is unidentifiable outside the neighborhood, the practical impact is that the estimated diagonal low-rank matrix may exhibit sparsity within its diagonal blocks, and the corresponding sparse matrix S may display some local low-rank features. Our simulation experiments show that when the sample size is small, the parameter estimates tend to be unstable. However, when the sample size is large, the algorithm converges more stably, and the parameters can be correctly estimated.

Table 5: Performance of the proposed method under GMGLV with increased eigenvalues of L (Mean (Std))

λ	d	TR_L	TP_L	FP_L	TP_S	FP_S	Err
4.0	0.2	1.000 (0.000)	0.947 (0.029)	0.024 (0.015)	1.000 (0.000)	0.081 (0.008)	0.109 (0.005)
	0.5	1.000 (0.000)	0.937 (0.027)	0.024 (0.013)	1.000 (0.000)	0.084 (0.007)	0.112 (0.006)
	1.0	1.000 (0.000)	0.951 (0.030)	0.027 (0.016)	1.000 (0.000)	0.083 (0.010)	0.110 (0.006)
	2.0	0.980 (0.141)	0.950 (0.026)	0.029 (0.012)	1.000 (0.000)	0.080 (0.007)	0.112 (0.005)
	3.0	1.000 (0.000)	0.953 (0.029)	0.024 (0.014)	1.000 (0.000)	0.082 (0.009)	0.113 (0.006)
4.1	0.2	1.000 (0.000)	0.918 (0.028)	0.006 (0.006)	1.000 (0.000)	0.110 (0.010)	0.117 (0.006)
	0.5	1.000 (0.000)	0.910 (0.024)	0.004 (0.005)	1.000 (0.000)	0.111 (0.009)	0.119 (0.005)
	1.0	1.000 (0.000)	0.918 (0.027)	0.006 (0.006)	1.000 (0.000)	0.109 (0.009)	0.117 (0.005)
	2.0	1.000 (0.000)	0.920 (0.023)	0.005 (0.005)	1.000 (0.000)	0.112 (0.008)	0.119 (0.006)
	3.0	1.000 (0.000)	0.919 (0.022)	0.006 (0.006)	1.000 (0.000)	0.109 (0.009)	0.120 (0.005)
4.2	0.2	0.980 (0.141)	0.909 (0.024)	0.004 (0.005)	1.000 (0.000)	0.110 (0.009)	0.119 (0.005)
	0.5	1.000 (0.000)	0.901 (0.018)	0.004 (0.005)	1.000 (0.000)	0.108 (0.008)	0.120 (0.006)
	1.0	1.000 (0.000)	0.903 (0.021)	0.005 (0.006)	1.000 (0.000)	0.108 (0.011)	0.120 (0.006)
	2.0	1.000 (0.000)	0.908 (0.023)	0.005 (0.006)	1.000 (0.000)	0.108 (0.010)	0.121 (0.006)
	3.0	1.000 (0.000)	0.907 (0.021)	0.005 (0.005)	1.000 (0.000)	0.108 (0.009)	0.120 (0.005)

Table 6: Performance of the proposed method under GMGLV with increased sparsity level of S (Mean (Std))

Sparsity	d	TR_L	TP_L	FP_L	TP_S	FP_S	Err
High	0.2	0.980 (0.141)	0.909 (0.024)	0.004 (0.005)	1.000 (0.000)	0.110 (0.009)	0.119 (0.005)
	0.5	1.000 (0.000)	0.901 (0.018)	0.004 (0.005)	1.000 (0.000)	0.108 (0.008)	0.120 (0.006)
	1.0	1.000 (0.000)	0.903 (0.021)	0.005 (0.006)	1.000 (0.000)	0.108 (0.011)	0.120 (0.006)
	2.0	1.000 (0.000)	0.908 (0.023)	0.005 (0.006)	1.000 (0.000)	0.108 (0.010)	0.121 (0.006)
	3.0	1.000 (0.000)	0.907 (0.021)	0.005 (0.005)	1.000 (0.000)	0.108 (0.009)	0.120 (0.005)
Medium	0.2	1.000 (0.000)	0.866 (0.033)	0.000 (0.000)	0.999 (0.005)	0.136 (0.010)	0.158 (0.005)
	0.5	1.000 (0.000)	0.848 (0.029)	0.000 (0.001)	1.000 (0.000)	0.134 (0.010)	0.160 (0.005)
	1.0	1.000 (0.000)	0.862 (0.037)	0.000 (0.001)	1.000 (0.000)	0.133 (0.008)	0.160 (0.005)
	2.0	1.000 (0.000)	0.863 (0.037)	0.000 (0.000)	0.998 (0.008)	0.134 (0.011)	0.160 (0.006)
	3.0	1.000 (0.000)	0.861 (0.033)	0.000 (0.000)	1.000 (0.000)	0.135 (0.008)	0.160 (0.004)
Low	0.2	1.000 (0.000)	0.854 (0.030)	0.000 (0.000)	0.977 (0.018)	0.156 (0.010)	0.194 (0.006)
	0.5	1.000 (0.000)	0.863 (0.033)	0.000 (0.001)	0.973 (0.020)	0.155 (0.012)	0.194 (0.006)
	1.0	1.000 (0.000)	0.852 (0.027)	0.000 (0.000)	0.975 (0.019)	0.155 (0.009)	0.193 (0.005)
	2.0	1.000 (0.000)	0.861 (0.029)	0.000 (0.000)	0.977 (0.018)	0.157 (0.009)	0.194 (0.006)
	3.0	1.000 (0.000)	0.860 (0.027)	0.000 (0.000)	0.974 (0.022)	0.157 (0.009)	0.194 (0.005)

Table 7: Performance of the proposed method under GMGLV with increased sample size (Mean (Std))

n	d	TR_L	TP_L	FP_L	TP_S	FP_S	Err
500	0.2	0.760 (0.431)	0.834 (0.077)	0.001 (0.003)	1.000 (0.000)	0.202 (0.011)	0.153 (0.009)
	0.5	0.740 (0.443)	0.829 (0.087)	0.001 (0.001)	1.000 (0.000)	0.196 (0.012)	0.154 (0.010)
	1.0	0.820 (0.388)	0.837 (0.079)	0.002 (0.004)	1.000 (0.000)	0.198 (0.012)	0.153 (0.011)
	2.0	0.720 (0.454)	0.813 (0.092)	0.002 (0.004)	1.000 (0.000)	0.199 (0.013)	0.155 (0.011)
	3.0	0.860 (0.351)	0.859 (0.075)	0.002 (0.003)	1.000 (0.000)	0.197 (0.011)	0.151 (0.010)
1000	0.2	0.980 (0.141)	0.909 (0.024)	0.004 (0.005)	1.000 (0.000)	0.110 (0.009)	0.119 (0.005)
	0.5	1.000 (0.000)	0.901 (0.018)	0.004 (0.005)	1.000 (0.000)	0.108 (0.008)	0.120 (0.006)
	1.0	1.000 (0.000)	0.903 (0.021)	0.005 (0.006)	1.000 (0.000)	0.108 (0.011)	0.120 (0.006)
	2.0	1.000 (0.000)	0.908 (0.023)	0.005 (0.006)	1.000 (0.000)	0.108 (0.010)	0.121 (0.006)
	3.0	1.000 (0.000)	0.907 (0.021)	0.005 (0.005)	1.000 (0.000)	0.108 (0.009)	0.120 (0.005)
2000	0.2	1.000 (0.000)	0.924 (0.022)	0.006 (0.005)	1.000 (0.000)	0.066 (0.006)	0.108 (0.004)
	0.5	1.000 (0.000)	0.929 (0.025)	0.005 (0.006)	1.000 (0.000)	0.067 (0.008)	0.109 (0.004)
	1.0	1.000 (0.000)	0.926 (0.023)	0.004 (0.005)	1.000 (0.000)	0.067 (0.008)	0.109 (0.004)
	2.0	1.000 (0.000)	0.930 (0.022)	0.009 (0.007)	1.000 (0.000)	0.068 (0.008)	0.109 (0.004)
	3.0	1.000 (0.000)	0.927 (0.020)	0.008 (0.006)	1.000 (0.000)	0.067 (0.008)	0.110 (0.005)

C.1 Direct K-means comparison

In this section, we compare the proposed method with the direct K-means approach, in which K-means clustering is applied directly to the first-stage residuals. In the simulation study, we consider sample sizes $n \in \{1000, 1500, 2000\}$. For each value of n, we set $s_1 \in \{25, 30, 40\}$, so that the number of nonzero conditional dependence entries across different communities increases. For each replication, we generate n observations with $q = 2$ covariates. Specifically, the covariate matrix $Z \in \mathbb{R}^{n\times 2}$ has i.i.d. standard normal entries,

with each column standardized. The regression coefficient matrix $B \in \mathbb{R}^{2\times 45}$ has i.i.d. entries drawn from Uniform$(0.5, 1.0)$. The precision matrices are generated as follows.

Grouped latent variable graphical models: Suppose $m = 3$, i.e. there are three groups of latent variables, each of which has only one latent variable H_i, for $i = 1, 2, 3$. To satisfy the conditional independence assumption (4), we let $N_1 = \{1, 2, \ldots, 15\}, N_2 = \{16, \ldots, 30\}, N_3 = \{31, \ldots, 45\}$. For the observed-latent interaction matrix Θ_{OH}, we generate the entries of $(\Theta_{O,H_1})_{N_1}$ independently from Uniform$(1.20, 1.35)$, the entries of $(\Theta_{O,H_2})_{N_2}$ independently from Uniform$(1.25, 1.40)$, and the entries of $(\Theta_{O,H_3})_{N_3}$ independently from Uniform$(1.30, 1.45)$, while $(\Theta_{O,H_1})_{N_2\cup N_3} = 0, (\Theta_{O,H_2})_{N_1\cup N_3} = 0, (\Theta_{O,H_3})_{N_1\cup N_2} = 0$. Moreover, we set $\Theta_H = \mathrm{diag}(4.0, 4.6, 5.2)$. Denote $\Theta_O = (\theta^O_{ij})$. For each block N_b, $b = 1, 2, 3$, among all $\binom{15}{2}$ within-block pairs we sample $45 - s_1$ pairs uniformly without replacement, and set the corresponding off-diagonal entries independently from Uniform$\big((-0.18, -0.08)\cup (0.08, 0.18)\big)$. For each block pair $(N_a, N_b) \in \{(N_1, N_2), (N_1, N_3), (N_2, N_3)\}$, among all 15×15 cross-block pairs we sample s_1 pairs uniformly without replacement, and set the corresponding off-diagonal entries independently from Uniform$\big((-3, -2) \cup (2, 3)\big)$. All remaining off-diagonal entries of Θ_O are set to zero. For the diagonal entries, we generate $\theta^O_{ii} \sim$ Uniform$(10, 12), i = 1, \ldots, 45$, independent from everything else.

The results in Table 8 reveal two clear patterns. First, as s_1 increases, the Hamming distances of both the direct and proposed methods increase overall, indicating that stronger cross-community sparse contamination makes recovery of the latent grouping structure more difficult. Second, the two methods respond quite differently when sample size increases. The direct method does not show substantial improvement as n increases: when $s_1 = 30$ and $s_1 = 40$, its average Hamming distance remains stable at levels of $0.45-0.47$ and $0.57-0.58$, respectively.

However, the proposed method exhibits more gains as n increases. When $s_1 = 30$, its average Hamming distance decreases from 0.444 at $n = 1000$ to 0.306 at $n = 2000$; for $s_1 = 40$, it decreases from 0.574 to 0.469. Even under the weaker contamination setting $s_1 = 25$, the proposed method changes from being slightly inferior to the direct method in smaller samples to achieving the best performance at $n = 2000$.

We conclude that the direct method clusters primarily based on the marginal covariance structure of the residuals, so that when the implicit objective deviates from the true grouped latent structure, increasing the sample size may reduce the random error but not eliminate the structural bias. The proposed method, however, can more effectively separate cross-community contamination from the latent block structure through the sparse-plus-low-rank decomposition, and therefore achieves more stable and substantial performance improvements as the sample size grows.

Table 8: Hamming distance comparison between the direct K-means method and the proposed method under different sample sizes (Mean (Std)). s_1 denotes the number of nonzero conditional dependence entries across different community pairs.

n	s_1	Direct K-means	Proposed
1000	25	0.276 (0.099)	0.375 (0.080)
	30	0.449 (0.070)	0.444 (0.083)
	40	0.572 (0.048)	0.574 (0.033)
1500	25	0.252 (0.081)	0.320 (0.074)
	30	0.459 (0.073)	0.341 (0.080)
	40	0.574 (0.033)	0.490 (0.052)
2000	25	0.259 (0.078)	0.200 (0.072)
	30	0.467 (0.065)	0.306 (0.088)
	40	0.584 (0.027)	0.469 (0.054)

Appendix D. Sufficient Condition for Irrepresentability Assumption 9

In this section, we provide sufficient conditions for the existence of the tuning parameters ρ_1 and ρ_2 in Assumption 10. Following Chandrasekaran et al. (2012), we use $\rho(\mathcal{T}, \mathcal{T}')$ to denote the distance between two subspaces. The quantities $\xi(\mathcal{T}(\cdot))$ and $\mu(\Omega(\cdot))$ quantify, respectively, the extent to which a low-rank matrix can resemble a sparse one and the extent to which a sparse matrix can resemble a low-rank one. We also introduce the Fisher information–based quantities α_Ω, δ_Ω, $\alpha_\mathcal{T}$, $\delta_\mathcal{T}$, β_Ω, and $\beta_\mathcal{T}$, which are used to verify the irrepresentable condition required for consistent model selection.

Assumption 3 guarantees the identifiability of S^* and L^* by requiring that Ω^* and $\mathcal{T}^*$ intersect transversally, i.e., $\Omega^* \cap \mathcal{T}^* = \{\mathbf{0}_{p\times p}\}$. To quantify the transversality of tangent spaces in a neighborhood of Ω^* and $\mathcal{T}^*$, we adopt the following quantities from Chandrasekaran et al. (2012). For any smooth point $L \in \mathcal{LS}(m, r)$, define

$$\xi(\mathcal{T}(L)) \triangleq \max\Big\{\|N\|_\infty : N = \mathrm{diag}(N_1, \cdots, N_m) \in \mathcal{T}(L),\ \|N_i\|_2 \le 1 \text{ for all } i \in [m]\Big\}.$$

For any smooth point $S \in \mathcal{S}(S^*)$, define

$$\mu\big(\Omega(S)\big) \triangleq \max\big\{\|N\|_2 : N \in \Omega(S),\ \|N\|_\infty = 1\big\}.$$

In particular, $\mu(\Omega^*) = \mu(\Omega(S^*))$. A small value of $\mu(\Omega^*)$ implies that the spectrum of matrices in $\Omega(S)$ is not overly concentrated.

Given two subspaces $\mathcal{T}$ and $\mathcal{T}'$ with the same dimension, we define

$$\rho(\mathcal{T}, \mathcal{T}') = \|\mathcal{P}_\mathcal{T} - \mathcal{P}_{\mathcal{T}'}\|_{2\to 2} = \max_{\|N\|_2 \le 1} \|\mathcal{P}_\mathcal{T}(N) - \mathcal{P}_{\mathcal{T}'}(N)\|_2$$

to measure the misalignment between them. Denote $\Omega^* = \Omega(S^*)$, $\mathcal{T}^* = \mathcal{T}(L^*)$, and $\mathcal{T}_{1i}^* = \mathcal{T}_1(L_i^*)$. In the analysis of the estimator for L^*, we focus on the following neighborhood:

$$\mathcal{N}(\mathcal{T}^*) \triangleq \left\{\mathcal{T}(L) : L = \mathrm{diag}(L_1, \cdots, L_m),\ \rho(\mathcal{T}_1(L_i), \mathcal{T}_{1i}^*) \le \frac{\xi(\mathcal{T}_{1i}^*)}{2},\ i \in [m]\right\}.$$

Next, we define the minimum gain of the operators $\mathcal{P}_{\Omega^*}\mathcal{I}^*\mathcal{P}_{\Omega^*}$ and $\mathcal{P}_{\mathcal{T}}\mathcal{I}^*\mathcal{P}_{\mathcal{T}}$ on the tangent spaces Ω^* and any $\mathcal{T} \in \mathcal{N}(\mathcal{T}^*)$:

$$\alpha_{\Omega^*} \triangleq \min_{M\in\Omega^*, \|M\|_\infty=1} \left\|\mathcal{P}_{\Omega^*}\mathcal{I}^*\mathcal{P}_{\Omega^*}(M)\right\|_\infty, \qquad \alpha_{\mathcal{T}^*} \triangleq \min_{\mathcal{T}\in\mathcal{N}(\mathcal{T}^*)} \min_{M\in\mathcal{T}, \|M\|_2=1} \left\|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*\mathcal{P}_{\mathcal{T}}(M)\right\|_2.$$

We also introduce the maximum effect of elements of Ω^* and $\mathcal{T}$ in the orthogonal directions $\Omega^{*\perp}$ and $\mathcal{T}^\perp$, respectively:

$$\delta_{\Omega^*} \triangleq \max_{M\in\Omega^*, \|M\|_\infty=1} \left\|\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*\mathcal{P}_{\Omega^*}(M)\right\|_\infty, \qquad \delta_{\mathcal{T}^*} \triangleq \max_{\mathcal{T}\in\mathcal{N}(\mathcal{T}^*)} \max_{M\in\mathcal{T}, \|M\|_2=1} \left\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*\mathcal{P}_{\mathcal{T}}(M)\right\|_2.$$

These four quantities describe how $\mathcal{I}^*$ behaves when restricted to the spaces Ω^* and $\mathcal{T}^*$ under their natural norms. Finally, we control the behavior of $\mathcal{I}^*$ on Ω^* in the spectral norm and on $\mathcal{T}^*$ in the ℓ_∞ norm:

$$\beta_{\Omega^*} \triangleq \max_{M\in\Omega^*, \|M\|_2=1} \left\|\mathcal{I}^*(M)\right\|_2, \qquad \beta_{\mathcal{T}^*} \triangleq \max_{\mathcal{T}\in\mathcal{N}(\mathcal{T}^*)} \max_{M\in\mathcal{T}, \|M\|_\infty=1} \left\|\mathcal{I}^*(M)\right\|_\infty.$$

We summarize the above quantities as follows:

$$\alpha^* \triangleq \min\left(\alpha_{\Omega^*}, \frac{1}{C+1}\alpha_{\mathcal{T}^*}\right), \delta_n \triangleq \max\left(\delta_{\Omega^*}, \sqrt{p}\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\delta_{\mathcal{T}^*}\right),$$

and $\beta^* \triangleq \max\left(\beta_{\Omega^*}, \beta_{\mathcal{T}^*}\right)$, where $C > 1$ is a prescribed constant.

For convenience, we also define

$$k_{\rho_1}(S, L) = \max\left\{\|S\|_\infty, \frac{\|L\|_2}{\rho_1}\right\},$$

and the dual norm associated with the penalty by

$$g_{\rho_1,\rho_2,L,\Lambda_n}(S, M) = \max\left\{\|S\|_\infty, \min_{M=M_1+M_2}\left\{\frac{\|M_1\|_2}{\rho_1} \vee \frac{r_n^{1-\alpha}\left\|\Lambda_n \odot M_2\right\|_\infty}{\rho_2}\right\}\right\}.$$

Note that $h_{\rho_1,\rho_2,L,\Lambda_n}$ is an upper bound for $g_{\rho_1,\rho_2,L,\Lambda_n}$.

Theorem 33 *For smooth points $S \in \mathcal{S}(s_0)$ and $L \in \mathcal{LS}(m, r)$, let Ω^* and $\mathcal{T}^*$ be the corresponding tangent spaces. Suppose that there exists a $v \in \left(0, \frac{1}{2}\right]$ such that*

$$\frac{\delta_n}{\alpha^*} \leq 1 - 2v$$

for all sufficiently large n, and that Assumption 1 holds with

$$\mu(\Omega^*)\xi(\mathcal{T}^*) \leq \frac{1}{3\sqrt{s}}\left(\frac{\alpha^*\tilde{v}}{C_n\beta^*}\right)^2,$$

where $s = \max\{d_1, \cdots, d_m\}$, $\tilde{v} = \frac{v}{2-v}$, and

$$C_n = \max\left\{\frac{3\sqrt{r}}{C+1},\ 4r\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\right\}.$$

Suppose further that the tuning parameters ρ_1 and ρ_2 satisfy

$$\rho_1 \in \left[\frac{\sqrt{s}C_n\beta^*\xi(\mathcal{T}^*)}{\alpha^*\tilde{v}}, \frac{\alpha^*\tilde{v}}{3C_n\beta^*\mu(\Omega^*)}\right], \qquad \frac{\rho_2}{\rho_1} \asymp \frac{r_n^{1-\alpha}}{pM_{n1}}.$$

Then the following conclusions hold for $\mathcal{Y} = \Omega^* \times \mathcal{T}$ with $\mathcal{T} \in \mathcal{N}(\mathcal{T}^*)$:

(1) The minimum gain of $\mathcal{I}^*$ restricted to $\Omega^* \oplus \mathcal{T}$ is bounded below, i.e., for all $(S, L') \in \mathcal{Y}$,

$$g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S, L')\right) \geq \frac{\alpha^*}{2} k_{\rho_1}(S, L').$$

(2) The effect of elements in $\mathcal{Y}$ on the orthogonal direction $\mathcal{Y}^\perp = \Omega^{*\perp} \times \mathcal{T}^\perp$ is bounded above:

$$\left\|\mathbf{G}^\perp\mathbf{G}^{-1}\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n} \to h_{\rho_1,\rho_2,L,\Lambda_n}} \leq 1 - v,$$

that is,

$$h_{\rho_1,\rho_2,L,\Lambda_n}(\mathbf{G}^\perp(S, L')) \leq (1 - v) g_{\rho_1,\rho_2,L,\Lambda_n}(\mathbf{G}(S, L')).$$

Remark 34 *1. The direct-sum property $\Omega^* \oplus \mathcal{T}$ in part (1) follows immediately from the lower bound in part (1) together with the injectivity of $\mathcal{I}^*$. Hence, Assumption 3 is satisfied in this setting. 2. The inequality in part (2) immediately implies Assumption 10 by taking $(S, L') = \mathbf{G}^{-1}(\text{sign}(\mathbf{O}(S^*)), \rho_1 U_1^* U_1^{*T})$. 3. By suitably choosing the weights W_n, we can ensure that $M_{n1}M_{n2} = O(1)$, so the requirement reduces to $\mu(\Omega^*)\xi(\mathcal{T}^*) \lesssim \frac{1}{\sqrt{s}}$. This condition can hold when each low-rank block L_i^* on the diagonal of L^* is sufficiently incoherent and the graph associated with S^* has bounded degree. In that case, Chandrasekaran et al. (2011, 2012) imply that $\xi(\mathcal{T}^*) \asymp \sqrt{\frac{r}{p}} \asymp \frac{1}{\sqrt{s}}, \mu(\Omega^*) = O(1)$.*

Proof Since $\mathcal{T} \in \mathcal{N}(\mathcal{T}^*)$, we have $\rho(\mathcal{T}, \mathcal{T}^*) \leq \frac{\xi(\mathcal{T}^*)}{2}$. Hence, by Lemma 3.1 and Proposition 3.3 in Chandrasekaran et al. (2012),

$$\xi(\mathcal{T}) \leq \frac{\xi(\mathcal{T}^*) + \rho(\mathcal{T}, \mathcal{T}^*)}{1 - \rho(\mathcal{T}, \mathcal{T}^*)} \leq 3\xi(\mathcal{T}^*).$$

(1) It suffices to show that

$$g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S, L')\right) \geq \frac{\alpha^*}{2} \max\left\{\|S\|_\infty, \frac{\|L'\|_2}{\rho_1}\right\} \triangleq \frac{\alpha^*}{2} k_{\rho_1}(S, L').$$

Equivalently,

$$\min_{k_{\rho_1}(S,L')=1} g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S, L')\right) \geq \frac{\alpha^*}{2}.$$

We consider two cases.

For case (i), let $\|S\|_\infty = 1$ and $\|L'\|_2 = \eta \le \rho_1$. Then

$$\begin{aligned}
\|\mathcal{P}_{\Omega^*}\mathcal{I}^*(S+L')\|_\infty &\ge \|\mathcal{P}_{\Omega^*}\mathcal{I}^*S\|_\infty - \|\mathcal{P}_{\Omega^*}\mathcal{I}^*L'\|_\infty \\
&\ge \alpha_{\Omega^*} - \|\mathcal{I}^*L'\|_\infty \\
&= \alpha_{\Omega^*} - \eta \left\| \mathcal{I}^*\left(\frac{L'}{\eta}\right)\right\|_\infty \\
&\ge \alpha_{\Omega^*} - \eta\,\beta^* \left\|\frac{L'}{\eta}\right\|_\infty \\
&\ge \alpha_{\Omega^*} - \rho_1\beta^*\xi(\mathcal{T}) \\
&\ge \alpha^* - \rho_1\beta^*\xi(\mathcal{T}).
\end{aligned}$$

Therefore,

$$g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S,L')\right) \ge \alpha^* - \rho_1\beta^*\xi(\mathcal{T}).$$

For case (ii), let $\|S\|_\infty = q \le 1$ and $\|L'\|_2 = \rho_1$. We first note that

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L')\|_2 &\ge \|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*L'\|_2 - \|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*S\|_2 \\
&\ge \rho_1\alpha_{\mathcal{T}^*} - 3\sqrt{r}\|\mathcal{I}^*S\|_2 \\
&= \rho_1\alpha_{\mathcal{T}^*} - 3\sqrt{r}\,q\left\|\mathcal{I}^*\left(\frac{S}{q}\right)\right\|_2 \\
&\ge \rho_1\alpha_{\mathcal{T}^*} - 3\sqrt{r}\,q\,\beta^*\left\|\frac{S}{q}\right\|_2 \\
&\ge \rho_1\alpha_{\mathcal{T}^*} - 3\sqrt{r}\beta^*\mu(\Omega^*).
\end{aligned}$$

It therefore remains to prove that

$$\min_{\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L')=M_1+M_2}\left\{\frac{\|M_1\|_2}{\rho_1} \vee \frac{r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot M_2\|_\infty\right\} \gtrsim \|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L')\|_2.$$

Let M_1^* and M_2^* be the corresponding minimizers, so that

$$\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L') = M_1^* + M_2^*.$$

Then for any $a \in (0,1)$,

$$\begin{aligned}
\frac{\|M_1^*\|_2}{\rho_1} \vee \frac{r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot M_2^*\|_\infty &\ge a\frac{\|M_1^*\|_2}{\rho_1} + \frac{(1-a)r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot M_2^*\|_\infty \\
&\ge \frac{a}{\rho_1}\|M_1^*\|_2 + \frac{\frac{r_n^{1-\alpha}}{M_{n1}}(1-a)}{p\rho_2}\|M_2^*\|_2.
\end{aligned}$$

Take

$$a = \frac{\frac{r_n^{1-\alpha}}{M_{n1}}\rho_1}{p\rho_2 + \frac{r_n^{1-\alpha}}{M_{n1}}\rho_1},$$

so that

$$\frac{a}{\rho_1} = \frac{\frac{r_n^{1-\alpha}}{M_{n1}}(1-a)}{p\rho_2}.$$

Then

$$\begin{aligned}\frac{\|M_1^*\|_2}{\rho_1} \vee \frac{r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot M_2^*\|_\infty &\geq \frac{\frac{r_n^{1-\alpha}}{M_{n1}}}{p\rho_2 + \frac{r_n^{1-\alpha}}{M_{n1}}\rho_1}\left(\|M_1^*\|_2 + \|M_2^*\|_2\right) \\ &\geq \frac{\frac{r_n^{1-\alpha}}{M_{n1}}}{p\rho_2 + \frac{r_n^{1-\alpha}}{M_{n1}}\rho_1}\|M_1^* + M_2^*\|_2 \\ &= \frac{\frac{r_n^{1-\alpha}}{M_{n1}}}{p\rho_2 + \frac{r_n^{1-\alpha}}{M_{n1}}\rho_1}\|\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L')\|_2.\end{aligned}$$

Hence,

$$\begin{aligned}g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S,L')\right) &\geq \min_{\mathcal{P}_{\mathcal{T}}\mathcal{I}^*(S+L')=M_1+M_2}\left\{\frac{\|M_1\|_2}{\rho_1} \vee \frac{r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot M_2\|_\infty\right\} \\ &\geq \frac{\frac{r_n^{1-\alpha}}{M_{n1}}}{p\rho_2 + \frac{r_n^{1-\alpha}}{M_{n1}}\rho_1}\left(\rho_1\alpha_{\mathcal{T}^*} - 3\sqrt{r}\beta^*\mu(\Omega^*)\right) \\ &= \frac{\frac{r_n^{1-\alpha}}{M_{n1}}}{p\frac{\rho_2}{\rho_1} + \frac{r_n^{1-\alpha}}{M_{n1}}}\left(\alpha_{\mathcal{T}^*} - \frac{3\sqrt{r}}{\rho_1}\beta^*\mu(\Omega^*)\right) \\ &\geq \alpha^* - \frac{3\sqrt{r}}{C+1}\frac{1}{\rho_1}\beta^*\mu(\Omega^*).\end{aligned}$$

Combining the two cases, and recalling that

$$C_n = \max\left\{\frac{3\sqrt{r}}{C+1},\ 4r\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\right\},$$

we obtain

$$\begin{aligned}g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}(S,L')\right) &\geq \alpha^* - \beta^*\max\left\{\rho_1\xi(\mathcal{T}), \frac{3\sqrt{r}}{C+1}\frac{1}{\rho_1}\mu(\Omega^*)\right\} \\ &\geq \alpha^* - C_n\beta^*\max\left\{3\rho_1\xi(\mathcal{T}^*), \frac{1}{\rho_1}\mu(\Omega^*)\right\} \\ &\geq \alpha^* - \frac{v}{2-v}\alpha^* \\ &= \frac{2-2v}{2-v}\alpha^* \\ &\geq \frac{1}{2}\alpha^*.\end{aligned}$$

The last inequality uses the condition $0 < v \leq \frac{1}{2}$.

(2) By definition of the operator norm,

$$\left\|\mathbf{G}^\perp\mathbf{G}^{-1}\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to h_{\rho_1,\rho_2,L,\Lambda_n}} \leq \left\|\mathbf{G}^\perp\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to h_{\rho_1,\rho_2,L,\Lambda_n}} \left\|\mathbf{G}^{-1}\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to g_{\rho_1,\rho_2,L,\Lambda_n}}.$$

Moreover, by the argument in part (1),

$$\begin{aligned}\left\|\mathbf{G}^{-1}\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to g_{\rho_1,\rho_2,L,\Lambda_n}} &= \sup_{(S,L')\in\mathcal{Y}} \frac{g_{\rho_1,\rho_2,L,\Lambda_n}\left(\mathbf{G}^{-1}(S,L')\right)}{g_{\rho_1,\rho_2,L,\Lambda_n}(S,L')} \\ &\le \frac{1}{\inf_{(S,L')\in\mathcal{Y}} \frac{g_{\rho_1,\rho_2,L,\Lambda_n}(\mathbf{G}(S,L'))}{k_{\rho_1}(S,L')}} \\ &\le \frac{1}{\alpha^* - C_n\beta^* \max\left\{3\rho_1\xi(\mathcal{T}^*), \frac{1}{\rho_1}\mu(\Omega^*)\right\}}.\end{aligned}$$

On the other hand, when $k_{\rho_1}(S,L') = 1$, i.e., either $\|S\|_\infty = 1$ and $\|L'\|_2 \le \rho_1$, or $\|S\|_\infty \le 1$ and $\|L'\|_2 = \rho_1$, we have

$$\begin{aligned}\|\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*(S+L')\|_\infty &\le \|\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*S\|_\infty + \|\mathcal{P}_{\Omega^{*\perp}}\mathcal{I}^*L'\|_\infty \\ &\le \delta_{\Omega^*} + \rho_1\beta^*\xi(\mathcal{T}),\end{aligned}$$

and

$$\begin{aligned}\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_2 &\le \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*S\|_2 + \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*L'\|_2 \\ &\le (3\sqrt{r}+1)\|\mathcal{I}^*S\|_2 + \rho_1\delta_{\mathcal{T}^*} \\ &\le 4\sqrt{r}\|\mathcal{I}^*S\|_2 + \rho_1\delta_{\mathcal{T}^*} \\ &\le 4\sqrt{r}\beta^*\mu(\Omega^*) + \rho_1\delta_{\mathcal{T}^*}.\end{aligned}$$

Now let

$$\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L') = N_1 \oplus N_2 \in \mathcal{V}(L) \oplus \mathcal{T}_2^{*\perp}$$

be the unique decomposition. By Lemma 26,

$$\begin{aligned}\max\left\{\frac{\|N_1\|_2}{\rho_1}, \frac{r_n^{1-\alpha}}{\rho_2}\|\Lambda_n \odot N_2\|_\infty\right\} &\le \max\left\{\frac{\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_F}{\varepsilon_0\rho_1}, \frac{r_n^{1-\alpha}M_{n2}}{\rho_2}\|N_2\|_\infty\right\} \\ &\le \max\left\{\frac{\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_F}{\varepsilon_0\rho_1}, \frac{r_n^{1-\alpha}M_{n2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_F}{\varepsilon_1\rho_2}\right\} \\ &\le \frac{1}{\rho_1}\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_F \\ &\le \frac{1}{\rho_1}\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\sqrt{rs}\,\|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{I}^*(S+L')\|_2 \\ &\le \max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\left(\frac{\sqrt{s}}{\rho_1}4r\beta^*\mu(\Omega^*) + \sqrt{rs}\delta_{\mathcal{T}^*}\right),\end{aligned}$$

where ε_0 and ε_1 are defined in Lemma 26. Therefore,

$$\begin{aligned}\|\mathbf{G}^\perp\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to h_{\rho_1,\rho_2,L,\Lambda_n}} &\le \max\left\{\delta_{\Omega^*}, \sqrt{rs}\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\delta_{\mathcal{T}^*}\right\} \\ &\quad + \beta^*\max\left\{\rho_1\xi(\mathcal{T}), \frac{\sqrt{s}}{\rho_1}4r\max\left\{\frac{1}{\varepsilon_0}, \frac{CpM_{n1}M_{n2}}{\varepsilon_1}\right\}\mu(\Omega^*)\right\} \\ &\le (1-2v)\alpha^* + C_n\beta^*\max\left\{\rho_1\xi(\mathcal{T}), \frac{\sqrt{s}}{\rho_1}\mu(\Omega^*)\right\}.\end{aligned}$$

Combining the above bounds, we conclude that

$$\begin{aligned}
\left\|\mathbf{G}^{\perp}\mathbf{G}^{-1}\right\|_{g_{\rho_1,\rho_2,L,\Lambda_n}\to h_{\rho_1,\rho_2,L,\Lambda_n}} &\leq \frac{(1-2v)\alpha^* + C_n\beta^* \max\left\{\rho_1\xi(\mathcal{T}), \frac{\sqrt{s}}{\rho_1}\mu(\Omega^*)\right\}}{\alpha^* - C_n\beta^* \max\left\{3\rho_1\xi(\mathcal{T}^*), \frac{1}{\rho_1}\mu(\Omega^*)\right\}} \\
&\leq \frac{(1-2v)\alpha^* + C_n\beta^* \max\left\{3\rho_1\xi(\mathcal{T}^*), \frac{\sqrt{s}}{\rho_1}\mu(\Omega^*)\right\}}{\alpha^* - C_n\beta^* \max\left\{3\rho_1\xi(\mathcal{T}^*), \frac{\sqrt{s}}{\rho_1}\mu(\Omega^*)\right\}} \\
&\leq \frac{(1-2v)\alpha^* + \tilde{v}\alpha^*}{\alpha^* - \tilde{v}\alpha^*} \\
&= \frac{(1-2v) + \frac{v}{2-v}}{1 - \frac{v}{2-v}} \\
&= 1 - v,
\end{aligned}$$

which completes the proof. ■

References

Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks1. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Emmanuel Candes and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

Emmanuel Candes, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2): 572–596, 2011.

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935 – 1967, 2012.

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. A fused latent and graphical model for multivariate binary data. *arXiv preprint arXiv:1606.08925*, 2016.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical LASSO for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397, 2014.

Gianluca De Nard, Olivier Ledoit, and Michael Wolf. Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*, 19(2):236–257, 2021.

Carson Eisenach, Florentina Bunea, Yang Ning, and Claudiu Dinicu. High-dimensional inference for cluster-based graphical models. *Journal of Machine Learning Research*, 21 (53), 2020.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521 – 541, 2009.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, 9(3):432–441, 12 2007. ISSN 1465-4644.

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

Jian Guo, Jie Cheng, Elizaveta Levina, George Michailidis, and Ji Zhu. Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics*, 9(2):821, 2015.

Botao Hao, Will Wei Sun, Yufeng Liu, and Guang Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18 (217):1–58, 2018.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618, 2008.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

John M Lee. *Smooth manifolds*. Springer, 2012.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Bing Li and Eftychia Solea. A nonparametric graphical model for functional data with application to brain networks based on fMRI. *Journal of the American Statistical Association*, 113(524):1637–1655, 2018.

Bing Li, Hyonho Chun, and Hongyu Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107 (497):152–167, 2012.

Xiaodong Li, Yudong Chen, and Jiaming Xu. Convex relaxation methods for community detection. *Statistical Science*, 36(1):2–15, 2021.

Cong Ma, Junwei Lu, and Han Liu. Inter-subject analysis: A partial Gaussian graphical model approach. *Journal of the American Statistical Association*, 116(534):746–755, 2021.

Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Computation*, 25(8):2172–2198, 2013.

Zhuang Ma, Zongming Ma, and Hongsong Yuan. Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21(1):86–152, 2020.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538 – 557, 2012.

Yabo Niu, Yang Ni, Debdeep Pati, and Bani K. Mallick. Covariate-assisted Bayesian graph learning for heterogeneous data. *Journal of the American Statistical Association*, 0(0): 1–15, 2023.

Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

Eugen Pircalabelu and Gerda Claeskens. Community-based group graphical LASSO. *Journal of Machine Learning Research*, 21(1):2406–2437, 2020.

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

Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low-rank matrices. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 51–58, 2012.

Dapeng Shi, Tiandong Wang, and Zhiliang Ying. Simultaneous identification of sparse structures and communities in heterogeneous graphical models. *arXiv preprint arXiv:2405.09841*, 2024.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

Kean Ming Tan, Daniela Witten, and Ali Shojaie. The cluster graphical LASSO for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, (85):23–36, 2015.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced-rank huber regression in high dimensions. *Journal of the American Statistical Association*, 118(544):2383–2393, 2023.

Davoud Ataee Tarzanagh and George Michailidis. Estimation of graphical models through structured norm minimization. *Journal of Machine Learning Research*, 18(1):1–48, 2018.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Lucina Q Uddin, Richard F Betzel, Jessica R Cohen, Jessica S Damoiseaux, Felipe De Brigard, Simon B Eickhoff, Alex Fornito, Caterina Gratton, Evan M Gordon, Angela R Laird, et al. Controversies and progress on standardization of large-scale brain network nomenclature. *Network Neuroscience*, 7(3):864–905, 2023.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yikai Wang and Ying Guo. Locus: A regularized blind source separation method with low-rank structure for investigating brain connectivity. *The Annals of Applied Statistics*, 17(2):1307, 2023.

Zeya Wang, Veerabhadran Baladandayuthapani, Ahmed O. Kaseb, Wenyi Wang Hesham M. Amin, Manal M. Hassan, and Jeffrey S. Morris. Bayesian edge regression in undirected graphical models to characterize interpatient heterogeneity in cancer. *Journal of the American Statistical Association*, 117(538):533–546, 2022.

G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl*, 170(1):33–45, 1992.

Ines Wilms and Jacob Bien. Tree-based node aggregation in sparse graphical models. *Journal of Machine Learning Research*, 23(243):1–36, 2022.

Angli Xue, Seyhan Yazar, Drew Neavin, and Joseph E Powell. Pitfalls and opportunities for applying latent variables in single-cell eqtl analyses. *Genome Biology*, 24(1):33, 2023.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014.

Ying Yu, Yuanbang Mai, Yuanting Zheng, and Leming Shi. Assessing and mitigating batch effects in large-scale omics studies. *Genome biology*, 25(1):254, 2024.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Xiao-Tong Yuan and Tong Zhang. Partial Gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687, 2014.

Jingfei Zhang and Yi Li. High-dimensional Gaussian graphical regression models with covariates. *Journal of the American Statistical Association*, 118(543):2088–2100, 2023.

Peng Zhao and Bin Yu. On model selection consistency of LASSO. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006.

Ruixuan Zhao, Haoran Zhang, and Junhui Wang. Identifiability and consistent estimation for Gaussian chain graph models. *Journal of the American Statistical Association*, 0(0): 1–12, 2024.

Lucija Žignić, Stjepan Begušić, and Zvonko Kostanjčar. Block-diagonal idiosyncratic covariance estimation in high-dimensional factor models for financial time series. *Journal of Computational Science*, 81:102348, 2024.

Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.