Title
Statistical Consistency of Structural Learning in Networks and Graphical Models

Author
Lou, Xingmei

Publication Date
2022

Peer reviewed|Thesis/dissertation

Statistical Consistency of Structural Learning in Networks and Graphical Models

By

XINGMEI LOU

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Xiaodong Li, Chair

Naoki Saito

Debashis Paul

Committee in Charge

2022

To my family.

Contents

Statistical Consistency of Structural Learning in Networks and Graphical Models

Abstract

This dissertation aims to address the statistical consistency for two classical structural learning problems, one is community detection in networks, and the other one is structural learning in probabilistic graphical models. The methods studied in this dissertation are straightforward and widely used. However, their statistical consistency has not been fully revealed, especially under more sophisticated conditions that emerge in modern data analysis.

Chapter 2 discusses my joint work with Professor Xiaodong Li and Lihua Lei. We study the hierarchy of communities in real-world networks under a generic stochastic block model, in which the connection probabilities are structured in a binary tree. Under such model, we prove the strong consistency of a standard recursive bi-partitioning algorithm under a wide range of model parameters, which include sparse networks with node degrees as small as $O(\log n)$. In addition, unlike most of existing work, our theory covers multi-scale networks where the connection probabilities may differ by orders of magnitude, which comprise an important class of models that are practically relevant but technically challenging to deal with. Finally we demonstrate the performance of our algorithm on synthetic data and real-world examples.

Chapter 3 discusses my work supervised by Professor Xiaodong Li and Professor Yu Hu. In this work, we are interested in the problem of learning the directed acyclic graph (DAG) when data are generated from a linear structural equation model (SEM) and the causal structure can be characterized by a polytree. Under the Gaussian polytree models, we study sufficient and necessary conditions on the sample sizes for the well-known Chow-Liu algorithm to exactly recover both the skeleton and the equivalence class of the polytree, which is uniquely represented by a CPDAG. We also consider extensions to the sub-Gaussian case, and then study the estimation of the inverse correlation matrix under such models. Our theoretical findings are illustrated by comprehensive numerical simulations, and experiments on benchmark data also demonstrate the robustness of polytree learning when the true graphical structures can only be approximated by polytrees.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Professor Xiaodong Li for his continuous help and support during the PhD program. His extensive knowledge, insightful advice and enthusiasm for research has inspired me deeply and guided me through this tough journey with fruitful accomplishment. Xiaodong is a great and supportive advisor who is always listening the students' voice and open to discuss different research topics and arouse interest in students. I would not have been able to work as an academic statistician and contact with a multitude of areas without his encouragement, trust and support. Most importantly, I have learned how to stay grounded and focused from him which will greatly benefit my future life. I would like to acknowledge the support from NSF CAREER Award DMS-184857 from my advisor Professor Xiaodong Li and NSF HDR TRIPODS grant CCF-1934568.

I would also like to extend my gratitude to my collaborators, Lihua Lei and Professor Yu Hu. I am extremely lucky to have Lihua as my collaborator in the first project during my PhD. He is always so sharp and capable to bring up brilliant ideas. Lihua sets a perfect role model for me as a graduate student in statistics. I also had great pleasure of working with Yu on my second research project. Without his broad knowledge and accurate intuition in graphical models, it would have been much more difficult for me to navigate through this area. I learned so many deep insights from the discussion between Xiaodong and Yu in our regular weekly meetings.

I would like to thank all the professors for serving on my qualifying exam and dissertation committees: Professor Naoki Saito, Professor James Sharpnack, Professor Can Le, Professor Debashis Paul, Professor Krishnakumar Balasubramanian and Professor Luis Rademacher. I am very grateful for their kind assistance, encouragement and valuable comments.

Next I would like to thank our department and PhD program, which I am very proud of. I am grateful to all the professors for offering wonderful courses, which helped me build a solid foundation of both statistical theory and applied skills. My thanks also go to all staff, especially Sarah Driver, who is always there helping me with numerous issues patiently. I am very thankful to my amazing peers, Yucheng Liu, Satarupa Bhattacharjee, Alvaro Eduardo Gajardo Cataldo and Rui Hu. My PhD journey was less intimidating and much more joyful because of your companion. I would also like to thank my excellent seniors, Zhenyu Wei, Lifeng Wei, Qin Ding, Yi-Wei Liu, Xiaodong Wang,

Jianing Fan, Siteng Hao and Junwen Yao. I learned a lot from our academic and career-planning discussions. I owe my special thanks to Shitong Wei and Tongyi Tang for providing me unwavering support and encouragement along the way.

Finally, I owe the most to my family. I could not have made any achievement without their unconditional love and support. I am brave enough to face the unknown future because I know my family is always standing behind me. I dedicate this dissertation to my parents and brother.

CHAPTER 1

Overview

Nowadays, graphs or networks are ubiquitous in statistical learning. We are going to consider two scenarios where graphs are playing an important role of understanding and explaining complex systems with statistical tools. Firstly, the data itself appears in the form of a graph, for example, social networks, where the nodes represent real entities and what we observe is whether any two nodes are connected by edges or not. Secondly, graphs can be used to compactly describe the data generating mechanism, where nodes represent random variables of interest and edges between nodes imply conditional dependence or cause-effect relationship between variables. In the first case, we focus on how to model and recover the hierarchically structured communities within a network, and study the statistical consistency of a recursive spectral method. In the second case, we revisit the classical Chow-Liu algorithm for learning a linear graphical model characterized by a polytree.

1.1. Consistency of Spectral Clustering on Hierarchical Stochastic Block Models

Community detection is an important problem in statistics, theoretical computer science and physics. A widely studied theoretical model in this area is the stochastic block model (Abbe, 2017). However, in real-world networks, the community structure is always hierarchical, which means large communities at coarse level can be further partitioned at finer scale. To properly model such hierarchy of communities in networks, we consider a general hierarchical stochastic block model that has been proposed in Clauset et al. (2008) and Balakrishnan et al. (2011), in which the within and between community connection probabilities are structured in a binary tree. Under such model, a standard recursive spectral bi-partitioning algorithm is dividing the network into two communities based on the Fiedler vector of the unnormalized graph Laplacian and repeating the split until a stopping rule indicates no further community structures.

Even though there has been extensive work on theoretical analysis of spectral methods for community detection in networks, the network models for theoretical analysis rarely encode the

hierarchy. On the other hand, the current theoretical literature tackling hierarchical community detection often has stringent conditions on the balance of the hierarchy, homogeneity of the connection probabilities and sparsity of the network. In Chapter 2, we show the weak and strong consistency of the recursive spectral bi-partitioning algorithm under a wider range of model parameters which permits sparse networks and multi-scale connection probabilities.

The flexibility of our model poses technical challenges for bounding the perturbation of the Fiedler vector of the unnormalized graph Laplacian under ℓ_2 and ℓ_∞ norms, which imply weak and strong consistency respectively. To overcome the challenges, we propose novel techniques, such as a novel decomposition of the graph Laplacian, and adapt the recently developed theory for entry-wise perturbation of eigenspace (Lei, 2019). Our theoretical findings are validated by various simulations under different hierarchies and different scales of the connection probabilities. We also demonstrate the performance of our method on recovering both the community membership and hierarchical structure with real-world examples.

This chapter is adapted from my joint work with Professor Xiaodong Li and Lihua Lei. The preprint was posted on ArXiv on April, 2020 (Lei et al., 2020).

1.2. Learning Linear Polytree Structural Equation Models

In the latter part of this dissertation, we consider another important problem, structural learning in probabilistic graphical models. Graphical models, especially with directed acyclic graphs(DAG), has regained its popularity because of the prevalence of causal inference in the past decade. Given the observations from a multivariate distribution, exactly recovering the underlying graph structure is an essential step for subsequent inference and prediction. In Chapter 3, we focus on the how to recover the skeleton and the equivalence class of the underlying graph, when data are generated from a linear structural equation model which encodes polytree causal structure. Polytree is a special kind of DAG without loops, and is tractable in inference with belief propagation.

The classical Chow-Liu algorithm (Chow and Liu, 1968) was used for structural learning of polytree graphical model since Rebane and Pearl (1987). Instead of applying Chow-Liu algorithm to the estimated mutual information as in Rebane and Pearl (1987), we apply it to the pairwise sample correlation in Chapter 3. When the exogenous variables in the linear structural equation model

follow Gaussian distribution, we provide sufficient conditions on the sample sizes to exactly recover the skeleton and the equivalence class of the polytree. On the other hand, necessary conditions on the required sample sizes for both skeleton and equivalence class recovery are also derived in terms of information-theoretic lower bounds, which match the respective sufficient conditions and thereby give a sharp characterization of the difficulty of these tasks. We extend the sufficient condition results to the sub-Gaussian case and provide the error bound for inverse correlation matrix estimation given that the equivalence class of the polytree is correctly recovered.

We compare our method with other structural learning approaches, namely the well-known PC algorithm and hill climbing algorithm, on comprehensive synthetic data. The simulation results agree with our theoretical findings. We also conduct such comparison on benchmark data where the true graph is actually not polytrees, which shows that the Chow-Liu algorithm is quite robust to the graphical structure in spite that it is tailored for polytrees. Thus this method can be universally used to provide a good initial guess for structural learning of graphical models under proper sparsity conditions.

This chapter is adapted from my joint work with Professor Xiaodong Li and Yu Hu. The preprint was posted on ArXiv on July, 2021 (Lou et al., 2021).

CHAPTER 2

Consistency of Spectral Clustering on Hierarchical Stochastic Block Models

2.1. Introduction

Community structures of real-world networks are typically hierarchical. In a coauthor network, it is not clear-cut whether all statisticians should be viewed as a single community — at a high level of granularity, they can be combined with mathematicians, physicists, computer scientists, and so on as quantitative scientists, while at a low level of granularity, they can be further split into finer groups based on research areas. When the desired level of granularity is unknown a priori, a hierarchy is a more informative representation of the relational information than a single partition of the network. For example, to design a cluster-wise randomized experiment, an A/B test designer can trade off the amount of interference (e.g., the number of edges between clusters) and effective sample size (e.g., the number of clusters) by searching over the hierarchy.

Agglomerative community detection algorithms (e.g. Girvan and Newman, 2002) are intrinsically hierarchical because they are able to produce a dendrogram characterizing the hierarchy of communities. However, bottom-up algorithms are sensitive to noise when amalgamating small clusters at the beginning of the run. As a consequence, theoretical guarantees are hard to come by for sparse and noisy networks. In contrast, divisive community detection algorithms, such as spectral clustering, has been proved to recover the community structure theoretically under various sparse network models (e.g. Abbe et al., 2015; Balakrishnan et al., 2011; Dasgupta et al., 2006; Jin, 2015; Lei and Rinaldo, 2015; Li et al., 2018; McSherry, 2001; Rohe et al., 2011). However, the network models for theoretical analysis rarely encode the hierarchy; often the communities are treated as logically separate units. Algorithmically, most divisive algorithms which have been analyzed in

theory are unable to produce a dendrogram. It is somewhat disappointing that hierarchical algorithms typically have no theoretical guarantees (under practically reasonable assumptions) while those justified in theory are often non-hierarchical algorithmically.

To mitigate this gap, one would need to consider a *hierarchical clustering algorithm* and study its statistical properties under a *hierarchical network model*. There have been attempts on this route, focusing on recursive divisive clustering algorithms, which recursively partition the network based on a top-down algorithm. A handful of such algorithms have been shown to recover the hierarchy under dense network models with average degree polynomial in n, where n is the number of nodes (Balakrishnan et al., 2011; Dasgupta, 2016; Lyzinski et al., 2016). However, real-world networks are typically much sparser. Dasgupta et al. (2006) analyzed a recursive spectral algorithm under a network model with average degree $O(\log^6 n)$. Apart from the artificial exponent, the algorithm involves multiple tuning parameters with no recommended default values, making it hard to implement in practice. Recently, Li et al. (2018) proposed the Binary Tree Stochastic Block Model (BTSBM) which encodes a binary hierarchy among primitive communities in the spirit of Clauset et al. (2008) and Balakrishnan et al. (2011). They analyzed a recursive spectral clustering algorithm, which splits the network into two clusters based on the signs of the components in an eigenvector of the adjacency matrix, under the BTSBM and showed it consistently recovers the hierarchy when the average degree scales as $O(\log^{2+\epsilon} n)$ for $\epsilon > 0$. The refined analysis by Lei (2019) brought it down to the critical regime with an $O(\log n)$ average degree. Despite being able to handle sparse networks, their analyses are based on a restrictive model which assumes a balanced hierarchy and strict layer-wise homogeneity in connection probabilities.

In this chapter, we analyze a Laplacian-based recursive bi-partitioning algorithm. It recursively splits the network into two based on the signs of the Fiedler vector (Fiedler, 1975), the eigenvector of the unnormalized Laplacian (formally defined in Section 2.2) corresponding to the second smallest eigenvalue. The procedure is repeated iteratively until a stopping rule indicates that there are no further communities in any subgraphs. Li et al. (2018) suggested various stopping rules that work reasonably well empirically and provided theoretical justification for certain ones. As shown by Li et al. (2018), to get K communities, this algorithm is computationally more efficient than the K-way

spectral clustering algorithm which splits the network into K communities at once, especially for a large K.

For the theoretical analysis, we consider a more general hierarchical SBM that has been proposed in Clauset et al. (2008) and Balakrishnan et al. (2011) to allow for an unbalanced hierarchy and heterogeneous connection probabilities. We prove that the proposed algorithm consistently recovers the hierarchy for sparse networks in the critical regime where the average degree scales as $O(\log n)$. Notably, as opposed to Li et al. (2018) and most works on non-hierarchical SBMs, we do not require all connection probabilities to be on the same scale; instead, we allow the connection probability between communities closer on the hierarchy to be orders of magnitude larger than that between communities farther apart. It makes our analysis more realistic since real-world networks are often multi-scale. Meanwhile, the theoretical analysis for clustering multi-scale networks becomes much more challenging. To highlight our main theoretical contribution, we will not investigate the performance of stopping rules in this chapter.

Our theory is built on (1) that the eigenvectors of the population Laplacian can identify the hierarchy; and (2) an entry-wise perturbation bound showing that the Fiedler vector of the observed Laplacian approximates the population version with a high accuracy. The first part generalizes the result of Balakrishnan et al. (2011) on the Fiedler vector to the entire eigen-structure. The second part rests on the recent development of $\ell_{2\to\infty}$ eigenvector perturbation theory (Abbe et al., 2017; Cape et al., 2019; Damle and Sun, 2020; Eldridge et al., 2017; Lei, 2019; Mao et al., 2017). The challenge is twofold: dependence between entries in the Laplacian and multi-scale connection probabilities. We tackle the first with the technique developed by Lei (2019), which, unlike most other perturbation bounds for random matrices, allows certain dependency structure among the entries. However, this technique alone is not enough to handle multi-scale networks. We overcome this challenge by introducing novel techniques, and we will elaborate on them in Section 2.3.

2.1.1. Notation. We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $\boldsymbol{e}_j$ to denote the j-th canonical basis where the j-th element equals to 1 and all other elements equal to 0 (with the dimension depending on the context). Vectors and matrices are boldfaced while scalars are not. We denote by $\boldsymbol{I}_n$ the $n \times n$ identity matrix and by $\mathbf{1}_n$ the $n \times 1$ column vector with all entries 1. For any vector $\boldsymbol{v}$, let $\|\boldsymbol{v}\|_p$ denote its ℓ_p norm. For any matrix $\boldsymbol{M}$, let $\boldsymbol{M}_k^T$ denote the k-th row of $\boldsymbol{M}$, $\|\boldsymbol{M}\|$ its spectral

norm, and $\|\boldsymbol{M}\|_{\mathrm{F}}$ its Frobenius norm. Further, we denote by $\lambda_1(\boldsymbol{M}), \ldots, \lambda_n(\boldsymbol{M})$ the eigenvalues of $\boldsymbol{M}$ in descending order, with $\boldsymbol{u}_1(\boldsymbol{M}), \ldots, \boldsymbol{u}_n(\boldsymbol{M})$ being the corresponding eigenvectors. For two sequences of real numbers $\{x_n\}$ and $\{y_n\}$, we write $x_n = o(y_n)$ or $y_n = \omega(x_n)$ if $\lim_{n\to\infty} \frac{x_n}{y_n} = 0$, $x_n = O(y_n)$ or $x_n \lesssim y_n$ if $|x_n| < C|y_n|$ for some constant C. Likewise, $x_n = \Omega(y_n)$ or $x_n \gtrsim y_n$ represents that there exists a constant C such that $|x_n| > C|y_n|$. Finally, we write $x_n \asymp y_n$ if $x_n \lesssim y_n$ and $y_n \lesssim x_n$.

2.2. A Hierarchical Stochastic Block Model

2.2.1. Model formulation. The *Stochastic Block Model* (SBM) proposed by Holland et al. (1983) has been widely used to study the empirical performance and theoretical properties of community detection methods. An SBM can be characterized by a vector $c = \{c_1, \ldots, c_n\} \in \{1, \ldots, K\}^n$ encoding the community membership of each node and a symmetric matrix $\boldsymbol{B} \in [0,1]^{K\times K}$ encoding the connection probabilities between communities. The upper triangular part of the adjacency matrix $\boldsymbol{A}$ has independent entries with $\boldsymbol{A}_{ij} \sim \text{Bernoulli}(\boldsymbol{B}_{c_i c_j})$ for any $i \le j$. By definition, the expected adjacency matrix, or equivalently the matrix of connection probabilities, can be represented as

$$\mathbb{E}[\boldsymbol{A}] \coloneqq \boldsymbol{P} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^\top - \text{diag}(\boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^\top) \in \mathbb{R}^{n\times n},$$

where $\boldsymbol{Z} \in \mathbb{R}^{n\times K}$ denotes the membership matrix with the i-th row vector $\boldsymbol{Z}_i = \boldsymbol{e}_{c_i}^\top$.

We consider a general Binary Tree Stochastic Block Model (BTSBM), which has been essentially proposed in Clauset et al. (2008) and Balakrishnan et al. (2011) with slightly different emphasis. Specifically, given a binary tree $\mathcal{T}$ with K leaf nodes, we represent each non-root node by a binary string recording the moves along the (unique) path from the root to that node, with 0 denoting a left move and 1 denoting a right move. For node s, let $|s|$ denote the depth of s and $(L(s), R(s))$ its two children nodes. For a pair of nodes s_1 and s_2, we denote by $\mathcal{A}(s_1, s_2)$ their lowest common ancestor. In such model, each node s on the binary tree $\mathcal{T}$ encodes two pieces of information. The first is a subset of units $\mathcal{G}_s \subset \{1, \ldots, n\}$, with $n_s = |\mathcal{G}_s|$. The model assumes that $\mathcal{G}_\emptyset = \{1, \ldots, n\}$ and $\{\mathcal{G}_{L(s)}, \mathcal{G}_{R(s)}\}$ forms a partition of $\mathcal{G}_s$, i.e., $\mathcal{G}_{L(s)} \cap \mathcal{G}_{R(s)} = \emptyset$ and $\mathcal{G}_{L(s)} \cup \mathcal{G}_{R(s)} = \mathcal{G}_s$. We refer to the network encoded by a leaf node as a *primitive community* and the network encoded by an internal node as a *mega-community*. The second piece of information is a connection probability

$p_s \in [0,1]$. For each pair of units $i \neq j$, the model assumes that the connection probability between this pair of nodes is

$$\boldsymbol{P}_{ij} = p_{\mathcal{A}(c(i),c(j))},$$

where $c(i)$ denotes the (unique) leaf node that contains i.

We illustrate in Figure 2.1 the above definitions by a toy example with $n = 8$ units and a binary tree $\mathcal{T}$ with $K = 5$ leaf nodes. The left panel shows the sub-networks encoded by each node. It can be read from the leaf nodes that $c(1) = c(7) = 00$, $c(2) = 10$, $c(3) = c(8) = 011$, $c(4) = 010$, and $c(5) = c(6) = 11$. The right panel shows the connection probabilities. Since $c(3) = 011, c(4) = 010$, and the lowest common ancestor of nodes 011 and 010 is 01, $\boldsymbol{P}_{34} = p_{01} = 0.04$. Similarly, $\boldsymbol{P}_{38} = p_{011} = 0.15, \boldsymbol{P}_{31} = p_0 = 0.03$ and $\boldsymbol{P}_{35} = p_\emptyset = 0.01$.

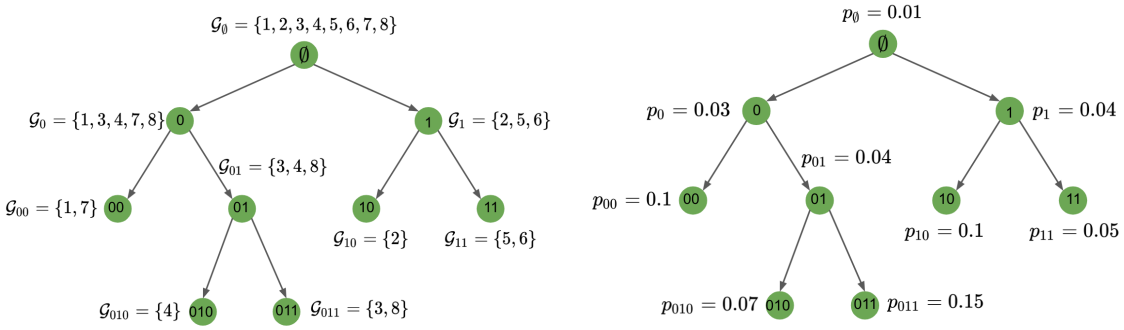


FIGURE 2.1. Illustration of a general BTSBM with $n = 8$ units and $K = 5$ primitive communities: (left) sub-network $\mathcal{G}_s$; (right) connection probability p_s.

Clearly, the hierarchical SBM defined above is a special case of SBM with $K = 5$ communities and the between-community connection probability matrix is

$$\boldsymbol{B} = \begin{bmatrix} p_{00} & p_0 & p_0 & p_\emptyset & p_\emptyset \\ p_0 & p_{010} & p_{01} & p_\emptyset & p_\emptyset \\ p_0 & p_{01} & p_{011} & p_\emptyset & p_\emptyset \\ p_\emptyset & p_\emptyset & p_\emptyset & p_{10} & p_1 \\ p_\emptyset & p_\emptyset & p_\emptyset & p_1 & p_{11} \end{bmatrix}.$$

Note that the BTSBM discussed in Li et al. (2018) is a restrictive special case of the model we study here in that $\mathcal{T}$ is assumed to be full and balanced with $p_s = p_{s'}$ and $n_s = n_{s'}$ whenever $|s| = |s'|$.

Throughout the chapter we focus on general BTSBMs that satisfy the weak assortativity.

Definition 2.2.1. A general BTSBM $(\mathcal{T}, \{p_s : s \in \mathcal{T}\}, c(\cdot))$ is weakly assortative if and only if $p_s < p_{s'}$ whenever s is the parent node of s'.

This is a natural property that is compatible with the intuitive explanations of hierarchies — nodes that are closer on the hierarchy are more likely to be connected. The example in Figure 2.1 is weakly assortative. Li et al. (2018) also discussed the disassortative setting, though it is less common than the assortative setting in practice.

We close this subsection with some notation that will be used repeatedly:

$$p^* = \max_{1\le i,j\le n} \boldsymbol{P}_{ij}, \quad \bar{p}^* = \max_{1\le i\le n} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{P}_{ij}, \quad \underline{p}_* = \min_{1\le i\le n} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{P}_{ij}. \tag{2.1}$$

By definition, p^* is the largest connection probability across the whole network, and $n\bar{p}^*$ ($n\underline{p}_*$) is the largest (smallest) expected degree. Obviously, $p^* \ge \bar{p}^* \ge \underline{p}_*$.

2.2.2. Population unnormalized graph Laplacian. The observed unnormalized graph Laplacian of a network is defined as

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A},$$

where $\boldsymbol{A} \in \{0,1\}^{n\times n}$ is the adjacency matrix and $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{A}\mathbf{1}_n)$ is the diagonal matrix whose diagonal entries are the node degrees. It is known that $\boldsymbol{L}$ is positive semidefinite. Let $\lambda_1 \ge \cdots \ge \lambda_{n-1} \ge \lambda_n = 0$ denote the eigenvalues of $\boldsymbol{L}$, and $\boldsymbol{u}_1, \ldots \boldsymbol{u}_{n-1}, \boldsymbol{u}_n$ the corresponding unit eigenvectors. Note that we always have $\boldsymbol{u}_n = \frac{1}{\sqrt{n}}\mathbf{1}_n$. In the case $\lambda_{n-1} = 0$, we choose $\boldsymbol{u}_{n-1}$ to be any eigenvector corresponding to $\lambda_{n-1} = 0$ which is orthogonal to $\mathbf{1}_n$.

As discussed in Section 2.1, our recursive spectral clustering algorithm splits the whole network into two based on the signs of the Fiedler eigenvector $\boldsymbol{u}_{n-1}$. Since $\boldsymbol{L}$ is an approximation of the population unnormalized graph Laplacian $\boldsymbol{L}^* = \mathbb{E}[\boldsymbol{L}]$, we shall study the eigenstructure of $\boldsymbol{L}^*$ as a stepping stone to prove the consistency of hierarchical clustering. Clearly, $\boldsymbol{L}^* = \mathrm{diag}(\boldsymbol{P}\mathbf{1}) - \boldsymbol{P}$. As with the sample version, the eigenvalues and unit eigenvectors of $\boldsymbol{L}^*$ are denoted as $\lambda_1^* \ge \cdots \ge \lambda_{n-1}^* \ge \lambda_n^* = 0$ and $\boldsymbol{u}_1^*, \ldots \boldsymbol{u}_{n-1}^*, \boldsymbol{u}_n^*$. Theorem 2.2.1 provides an elegant characterization of the eigenstructure of $\boldsymbol{L}^*$ under weak assortativity. The proof is relegated to Section 2.5.2.

THEOREM 2.2.1. *Under a weakly assortative general BTSBM, defined in Section 2.2.1,*

(1) $\lambda^*_{n-1} = np_\emptyset$ *with multiplicity* 1 *and the entries of the corresponding eigenvector obeys*

$$\boldsymbol{u}^*_{n-1,i} = \pm \begin{cases} \sqrt{n_1/(n_0 n)} & (i \in \mathcal{G}_0) \\ -\sqrt{n_0/(n_1 n)} & (i \in \mathcal{G}_1) \end{cases};$$

(2) $\lambda^*_{n-2} = \min\{n_1 p_1 + n_0 p_\emptyset, n_0 p_0 + n_1 p_\emptyset\}$*;*

(3) The number of eigenvalues, accounting for the multiplicity, that are strictly less than $n\underline{p}_*$ *is at most* K*, the number of leaf nodes in* $\mathcal{T}$*;*

(4) For any j *with* $\lambda^*_j < n\underline{p}_*$, $\|\boldsymbol{u}^*_j\|_\infty \le \sqrt{\xi/n}$ *where*

$$\xi := \max_s \frac{n}{n_s}. \tag{2.2}$$

Here recall that n_s *is the number of nodes in the subgragh encoded by the tree node* s*.*

Theorem 2.2.1 (1) implies that the entry signs of $\boldsymbol{u}^*_{n-1}$ encode the first split of the network. Theorem 2.2.1 (1) and (2) have been proved in Balakrishnan et al. (2011) and we include them here for completeness. In contrast, Theorem 2.2.1 (3) and (4) are new to the best of our knowledge. Although the eigenvalues and eigenvectors other than the Fiedler eigenpair appear to be algorithmically irrelevant, they are crucially useful in our theoretical analysis on the strong consistency of the clustering, especially for multi-scale networks whose connection probabilities have different scales. The conclusions of the eigen-structure in Theorem 2.2.1 also hold under a more general hierarchical SBM. We present the analogous theorem along with the proof in Appendix A.2.

2.3. Consistency of Recursive Spectral Clustering

2.3.1. Criteria for consistency. Recovering the hierarchy encoded by the tree is equivalent to recovering all mega-communities and primitive communities. When primitive communities are hard to be recovered, owing to either the small size or insufficient gap with others, it is still possible to recover some mega-communities at the top levels of the tree, yielding a partial hierarchy that is informative for downstream analysis. In either case, it is necessary to investigate the consistency of the first split, that is, whether the Fiedler vector of the unnormalized graph Laplacian can partition $\mathcal{G}_\emptyset$ accurately into $\mathcal{G}_0$ and $\mathcal{G}_1$.

We primarily focus on the strong consistency which requires the partition to be exactly correct with high probability. Without loss of generality, assume that $\boldsymbol{u}_{n-1,1} \geq 0$ and $\boldsymbol{u}^*_{n-1,1} \geq 0$. By Theorem 2.2.1, the strong consistency of the first split can be formally stated as

$$\mathbb{P}\left(\operatorname{sign}(\boldsymbol{u}_{n-1,i}) \neq \operatorname{sign}(\boldsymbol{u}^*_{n-1,i}), \text{ for some } i \in \{1,\ldots,n\}\right) = o(1), \tag{2.3}$$

and it can be easily extended to the consistent recovery of the whole hierarchy under the general BTSBM. Indeed, suppose we can identify a set of conditions under which the split at $\mathcal{G}_\emptyset$ is strongly consistent, replacing $\mathcal{G}_\emptyset$ with $\mathcal{G}_0$ yields the conditions for $\mathcal{G}_{00}$ and $\mathcal{G}_{01}$ to be exactly recovered, because the model for $\mathcal{G}_0$ is still a general BTSBM. Therefore, it is sufficient and necessary to investigate the strong consistency of the first split in order to establish the exact recovery of the full or partial hierarchy.

Another commonly studied criterion is the weak consistency which states that the misclustering error is asymptotically vanishing, i.e.,

$$\min_{a\in\{-1,1\}} \frac{1}{n}\sum_{i=1}^{n} I(\operatorname{sign}(\boldsymbol{u}_{n-1,i}) \cdot \operatorname{sign}(\boldsymbol{u}^*_{n-1,i}) = a) = o_{\mathbb{P}}(1). \tag{2.4}$$

In contrast to the strong consistency, the weak consistency of the first split does not carry over to lower splits under a general BTSBM, because the recovered $\mathcal{G}_0$ might involve units from $\mathcal{G}_1$ and hence the network model is no longer a general BTSBM. Nevertheless, we will still investigate the weak consistency since it relies on weaker conditions, and is also an important stepping stone to obtain strong consistency as we will explain later.

Both the strong and weak consistencies of the recursive spectral clustering under a general BTSBM can be viewed as extensions of the traditional consistency result for spectral clustering (e.g. Lei and Rinaldo, 2015) to a misspecified SBM that mistakenly assumes the number of clusters to be 2 while $K > 2$ in truth.

2.3.2. Main results. Intuitively, the clustering is consistent if $\boldsymbol{u}_{n-1}$ is close to $\boldsymbol{u}^*_{n-1}$. We will show the perturbation bound in ℓ_∞ and ℓ_2 norms under a weakly assortative general BTSBM, implying the strong and weak consistencies, respectively. The detailed proofs are deferred to Section 2.5.

THEOREM 2.3.1 (ℓ_∞ perturbation). *In the setting of Theorem 2.2.1, and further assume that $\xi = O(1)$, where ξ is defined in Equation (2.2). Then, for any fixed constant $r > 0$, there exists a constant C_{ℓ_∞} that only depends on r and ξ, such that*

$$\sqrt{n}\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_\infty < \min\{\sqrt{n_0/n_1}, \sqrt{n_1/n_0}\}$$

with probability at least $1 - (10K + 4)n^{-r}$, provided the following two conditions:

(2.5) *Density gap* $\quad \min\{n_0(p_0 - p_\emptyset), n_1(p_1 - p_\emptyset)\} \geq C_{\ell_\infty}\sqrt{(n_0p_0 + n_1p_1)\log n}$,

(2.6) *Degree variation* $\quad (n(\underline{p}_* - p_\emptyset))^4 \geq C_{\ell_\infty}(n\bar{p}^*)^3 \log n$,

where $\underline{p}_*$ and $\bar{p}^*$ are defined in (2.1).

REMARK 1 (Strong consistency). Recalling Theorem 2.2.1 that $\sqrt{n}\min_i|\boldsymbol{u}_{n-1,i}^*| = \min\{\sqrt{n_0/n_1}, \sqrt{n_1/n_0}\}$, Theorem 2.3.1 implies that the signs of the entries of $\boldsymbol{u}_{n-1}^*$ are preserved by $\boldsymbol{u}_{n-1}$ with high probability under conditions (2.5) and (2.6). Therefore, the conditions in Theorem 2.3.1 imply the strong consistency of the first split.

REMARK 2 (Sparse networks). Theorem 2.3.1 includes the case of sparse networks. In fact, conditions (2.5) and (2.6) can be simultaneously satisfied if $n_0 \asymp n_1$ and

$$\underline{p}_* \asymp \bar{p}^* \asymp (p_0 - p_\emptyset) \asymp (p_1 - p_\emptyset) = O(\log n/n),$$

in which case the expected degrees are on the order of $O(\log n)$.

REMARK 3 (Degree variation). Here we briefly discuss the degree variation condition (2.6). Consider the SBM with $K = 2$ again, where $n_0 \asymp n_1$, $p_\emptyset = O(\log n/n)$ and $p_0 = n^{-\gamma_0}$, $p_1 = n^{-\gamma_1}$ with constants γ_0, γ_1 satisfying $0 < \gamma_0 < \gamma_1 < 1$. Then the condition (2.5) is satisfied up to a constant if $\gamma_1 < \frac{\gamma_0+1}{2}$ and the condition (2.6) will hold up to a constant if $\gamma_1 < \frac{3\gamma_0+1}{4}$. We should admit these conditions may be improvable. For example, the two communities might be recoverable if we let $\gamma_0 \to 0$ while $\gamma_1 \to 1$.

REMARK 4 (Multi-scale networks). While most analyses of SBMs focus on the case where all connection probabilities are of the same order, our Theorem 2.3.1 can deal with multi-scale networks

where the maximum degree parameter $\bar{p}^* \gg p_\emptyset$. For example, when $p_s \asymp 1$ for every leaf node s, it is clear that the maximum and minimum degree parameters satisfy $\underline{p}_* \asymp \bar{p}^* \asymp 1$, and both the density gap condition (2.5) and the degree variation condition (2.6) hold for large n even when $p_\emptyset, p_0, p_1 \asymp \log n/n$. Our strong consistency result also guarantees adaptivity of graph Laplacian based spectral clustering to multi-scale networks with degree heterogeneity, e.g., $p_\emptyset, p_0, p_1 \asymp \log n/n$ and $\bar{p}^* = n^{-\gamma_0}$ and $\underline{p}_* = n^{-\gamma_1}$ with $0 < \gamma_1 < \frac{3\gamma_0+1}{4} < 1$.

As we will discuss in Section 2.3.3, the theoretical analysis for multi-scale networks is more challenging. In fact, to obtain the same bound in Theorem 2.3.1, the off-the-shelf ℓ_∞ perturbation bound by Lei (2019) on the Fiedler vector requires

$$n(\min\{p_0, p_1\} - p_\emptyset) \gtrsim \sqrt{n\bar{p}^* \log n}.$$

When $p_\emptyset, p_0, p_1 \asymp \log n/n$, $\bar{p}^*$ must be $O(\log n/n)$, excluding any multi-scale network. For example, consider the general BTSBM with $K = 4$, equal community sizes, and

$$\boldsymbol{B} = \begin{bmatrix} p_{00} & p_0 & p_\emptyset & p_\emptyset \\ p_0 & p_{01} & p_\emptyset & p_\emptyset \\ p_\emptyset & p_\emptyset & p_{10} & p_1 \\ p_\emptyset & p_\emptyset & p_1 & p_{11} \end{bmatrix}.$$

If we further assume $p_{00} = p_{01} = p_{10} = p_{11} = p^*$, then the maximum degree parameter satisfies $\bar{p}^* \asymp p^*$ due to weak assortativity, and thus the above eigengap condition implies $n \min\{p_0, p_1\} \gtrsim \sqrt{np^* \log n}$. It does not hold if $p_\emptyset, p_0, p_1 \asymp \log n/n$ but $p^* \gg \log n/n$. Note that p^* measures the connection probability within the primitive communities, the most connected groups on the hierarchy. It is hence disappointing and unrealistic to restrict p^* into the same order as $p_\emptyset$, p_0 and p_1. In Section 2.3.3, we will explain why multi-scale networks are challenging to work with in theory, and in Section 2.3.4 we will explain briefly how these challenges can be addressed.

The next result gives an ℓ_2 perturbation bound for the Fiedler vector.

THEOREM 2.3.2 (ℓ_2 perturbation). *Under the same setting of Theorem 2.2.1, for any fixed* $r, c > 0$,

$$\|\boldsymbol{u}_{n-1}\text{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_2 < c$$

with probability at least $1 - 2n^{-r}$*, provided that*

$$\min\{n_0(p_0 - p_\emptyset), n_1(p_1 - p_\emptyset)\} \geq C_{\ell_2}\sqrt{(n_0 p_0 + n_1 p_1)\log n} \tag{2.7}$$

where C_{ℓ_2} *is a sufficiently large constant that only depends on* r *and* c*.*

REMARK 5 (Beyond hierarchical SBM). While Theorem 2.3.2 is stated for general BTSBMs, the result holds for a much broader class of networks such that

$$\boldsymbol{P}_{ij} \begin{cases} = p_\emptyset & (i \in \mathcal{G}_0, j \in \mathcal{G}_1) \\ \geq p_0 & (i, j \in \mathcal{G}_0) \\ \geq p_1 & (i, j \in \mathcal{G}_1) \end{cases},$$

where $(p_\emptyset, p_0, p_1)$ satisfies the condition (2.7). The result will be stated formally in Appendix A.1.

REMARK 6 (Balancedness). Theorem 2.3.2 yields an bound on the misclustering error for the first split. Unlike the ℓ_∞ perturbation bound in Theorem 2.3.1, it does not require $\xi = O(1)$. However, the misclustering rate result may rely on certain balancedness. Assume $\boldsymbol{u}_{n-1}^T \boldsymbol{u}_{n-1}^* \geq 0$ without loss of generality and let $\mathcal{M} = \{i : \text{sign}(\boldsymbol{u}_{n-1,i}) \neq \text{sign}(\boldsymbol{u}_{n-1,i}^*)\}$. Obviously, the misclustering error is $|\mathcal{M}|/n$. By Theorem 2.2.1 (1), for each $i \in \mathcal{M}$, we have

$$|\boldsymbol{u}_{n-1,i} - \boldsymbol{u}_{n-1,i}^*| \geq \frac{1}{\sqrt{n}} \min\left\{\sqrt{\frac{n_1}{n_0}}, \sqrt{\frac{n_0}{n_1}}\right\} \geq \frac{1}{\sqrt{n\xi}},$$

where the last inequality uses the fact that $\min\{n_1/n_0, n_0/n_1\} \geq \min\{n_1, n_0\}/n \geq 1/\xi$. As a result,

$$\frac{|\mathcal{M}|}{n} \leq \xi \sum_{i \in \mathcal{M}} (\boldsymbol{u}_{n-1,i} - \boldsymbol{u}_{n-1,i}^*)^2 \leq \xi \|\boldsymbol{u}_{n-1} - \boldsymbol{u}_{n-1}^*\|_2^2 \leq \xi c^2.$$

When $\xi = O(1)$, Theorem 2.3.2 implies that ξc^2 can be arbitrarily small when C_{ℓ_2} is sufficiently large.

REMARK 7 (Relaxed degree variation condition). The density gap condition (2.7) is essentially the same as that for strong consistency, i.e., (2.5). However, the degree variation condition (2.6), which is required for strong consistency, is not required for weak consistency.

REMARK 8 ($O(\log n)$ degrees). For an SBM with $K = 2$ communities, the connection probabilities $p_\emptyset, p_0, p_1$ only need to be $\omega(1/n)$ for weak consistency (Abbe, 2017; Zhao et al., 2012).

Unfortunately, this cannot be achieved by spectral clustering based on the adjacency matrix or graph Laplacian (Krzakala et al., 2013), for which the condition (2.7) is required up to constants (e.g. Lei and Rinaldo, 2015). Nevertheless, we conjecture that the regularized spectral clustering, with appropriately chosen level of regularization, works in this regime. For example, one can consider removing the nodes whose degrees are greater than $C_0 np^*$ for some constant C_0. Le et al. (2017) proved that the adjacency matrix for the remaining graph has tighter spectral concentration around its population version. However, their theory does not directly apply to multi-scale hierarchical SBMs. Moreover, it is unclear how the truncation threshold should be chosen in practice since p^* is unknown. We leave this intriguing research question for future work.

2.3.3. Main challenge to handle multi-scale networks. A standard technique to obtain the ℓ_2 perturbation bound is via the Davis-Kahan $\sin\Theta$ Theorem (Lemma 2.5.3), which implies that

$$\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_2 \lesssim \frac{\|\boldsymbol{L}-\boldsymbol{L}^*\|}{\lambda_{n-2}^*-\lambda_{n-1}^*}.$$

Similarly, a straightforward application of the recently developed ℓ_∞ perturbation bound for unnormalized graph Laplacians (Lei, 2019) implies that, under additional regularity conditions and substantial simplifications,

$$\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_\infty \lesssim \frac{\|\boldsymbol{L}-\boldsymbol{L}^*\|}{\lambda_{n-2}^*-\lambda_{n-1}^*}\|\boldsymbol{u}_{n-1}^*\|_\infty \lesssim \frac{\|\boldsymbol{L}-\boldsymbol{L}^*\|}{\lambda_{n-2}^*-\lambda_{n-1}^*}\cdot\frac{1}{\sqrt{n}}$$

where the last inequality is implied by Theorem 2.2.1 (4). As a consequence, to obtain the $O(1/\sqrt{n})$ ℓ_∞ perturbation bound in Theorem 2.3.1 and the $O(1)$ ℓ_2 perturbation bound in Theorem 2.3.2 based on these techniques straightforwardly, it is required that

$$\lambda_{n-2}^* - \lambda_{n-1}^* \gtrsim \|\boldsymbol{L}-\boldsymbol{L}^*\|. \tag{2.8}$$

By Theorem 2.2.1 (1) and (2),

$$\lambda_{n-2}^* - \lambda_{n-1}^* = \min\{n_0(p_0-p_\emptyset), n_1(p_1-p_\emptyset)\}.$$

The best available matrix perturbation inequality (Lemma 2.5.1) shows that

$$\|\boldsymbol{L} - \boldsymbol{L}^*\| \lesssim \sqrt{n\bar{p}^* \log n}.$$

This bound cannot be improved when the average degrees of all nodes are the same in order, i.e., $\underline{p}_* \asymp \bar{p}^*$. Therefore, to guarantee (2.8), it requires that

$$\min\{n_0(p_0 - p_\emptyset), n_1(p_1 - p_\emptyset)\} \gtrsim \sqrt{n\bar{p}^* \log n}.$$

As discussed in Section 2.3.2, the above condition is overly stringent, illustrating that the standard techniques fail to handle multi-scale networks.

2.3.4. Proof ideas. To overcome the difficulty, we will introduce different novel techniques for the ℓ_∞ and ℓ_2 perturbation bounds. We start with the ℓ_∞ perturbation. As illustrated above, the main hurdle brought on by multi-scale networks is that the eigengap $\lambda^*_{n-2} - \lambda^*_{n-1}$ is local while the perturbation $\|\boldsymbol{L} - \boldsymbol{L}^*\|$ is global on the hierarchy. When $\bar{p}^* \gg \max\{p_0, p_1\}$, the eigengap for $\boldsymbol{u}_{n-1}$ is too small to yield a desirable eigenvector perturbation bound. Nevertheless, Theorem 2.2.1 (3) and (4) imply that there are at most K eigenvalues below $n\underline{p}_*$. By pigeonhole principle, there exists $j \le K$ such that

$$\lambda^*_{n-j+1} - \lambda^*_{n-j} \ge \frac{\lambda^*_{n-K} - \lambda^*_{n-1}}{K} \ge \frac{n(\underline{p}_* - p_\emptyset)}{K} \gtrsim n(\underline{p}_* - p_\emptyset),$$

where the last inequality uses the fact that $K \le \max_s(n/n_s) = \xi = O(1)$. Let $\boldsymbol{U}_j$ (resp. $\boldsymbol{U}^*_j$) be the $\mathbb{R}^{n\times j}$ matrix including eigenvectors $\{\boldsymbol{u}_{n-1}, \ldots, \boldsymbol{u}_{n-j}\}$ (resp. $\{\boldsymbol{u}^*_{n-1}, \ldots, \boldsymbol{u}^*_{n-j}\}$). Then the generic $\ell_{2\to\infty}$ bound proposed by Lei (2019), with substantial simplifications, implies that

$$\|\boldsymbol{U}_j\boldsymbol{O}_j - \boldsymbol{U}^*_j\|_{2\to\infty} \lesssim \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{\{n(\underline{p}_* - p_\emptyset)\}^2}\|\boldsymbol{U}^*_j\|_{2\to\infty},$$

where $\boldsymbol{O}_j \in \mathbb{R}^{j\times j}$ is an orthogonal matrix. The condition (2.6) and Theorem 2.2.1 (4) imply that

$$\|\boldsymbol{U}_j\boldsymbol{O}_j - \boldsymbol{U}^*_j\|_{2\to\infty} \lesssim \frac{1}{\sqrt{n}}.$$

Assuming $\boldsymbol{O}_j = I$, the above bound implies the desired ℓ_∞ perturbation bound for $\boldsymbol{u}_{n-1}$ since $\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_\infty \le \|\boldsymbol{U}_j - \boldsymbol{U}^*_j\|_{2\to\infty}$. This heuristic can be made rigorous by applying the ℓ_2

perturbation bound given in Theorem 2.3.2 and Davis-Kahan $\sin\Theta$ Theorem, which show that $\boldsymbol{O}_j \approx I$ in some appropriate sense. In sum, we deduce the ℓ_∞ perturbation bound for the Fiedler vector from a generic $\ell_{2\to\infty}$ perturbation bound for a larger eigenspace with a large eigengap. This also illustrates why we need the entire eigenstructure of $\boldsymbol{L}^*$ even if the algorithm merely uses the Fiedler vector.

As shown above, the ℓ_2 perturbation bound is key to establish the ℓ_∞ perturbation bound. As aforementioned, the direct application of Davis-Kahan $\sin\Theta$ Theorem fails because the matrix perturbation error $\|\boldsymbol{L}-\boldsymbol{L}^*\|$ is too large compared to the eigengap. However, instead of viewing $\boldsymbol{L}^*$ as the target and $\boldsymbol{L}-\boldsymbol{L}^*$ as the perturbation, we can replace $\boldsymbol{L}^*$ by any matrix $\tilde{\boldsymbol{L}}$ with $\boldsymbol{u}_{n-1}(\tilde{\boldsymbol{L}}) = \boldsymbol{u}^*_{n-1}$. The Davis-Kahan $\sin\Theta$ Theorem would imply a tighter bound if our selected $\tilde{\boldsymbol{L}}$ satisfies

$$\|\boldsymbol{L} - \tilde{\boldsymbol{L}}\| \ll \|\boldsymbol{L} - \boldsymbol{L}^*\|.$$

Typically, it is difficult to construct an explicit $\tilde{\boldsymbol{L}}$ without further structural assumptions on $\boldsymbol{L}^*$. However, we observe an intriguing property of unnormalized graph Laplacians that enables an easy construction of $\tilde{\boldsymbol{L}}$.

Lemma 2.3.3. *Let $\tilde{\boldsymbol{L}}$ be the unnormalized graph Laplacian of the pseudo-adjacency matrix $\widetilde{\boldsymbol{A}}$ that replaces the between-community edges by their common expectation $p_\emptyset$, i.e.,*

$$\widetilde{\boldsymbol{A}}_{ij} = \begin{cases} \boldsymbol{A}_{ij} & (i,j \in \mathcal{G}_0, \text{ or } i,j \in \mathcal{G}_1) \\ p_\emptyset & (\text{otherwise}) \end{cases}.$$

Under the same setting as in Theorem 2.3.2, if C_{ℓ_2} is a sufficiently large constant that only depends on r, with probability at least $1 - n^{-r}$,

$$\lambda_{n-1}(\tilde{\boldsymbol{L}}) = np_\emptyset \text{ with multiplicity } 1, \quad \text{and} \quad \boldsymbol{u}_{n-1,i}(\tilde{\boldsymbol{L}}) = \boldsymbol{u}^*_{n-1}.$$

By Lemma 2.3.3, $\tilde{\boldsymbol{L}}$ preserves the $(n-1)$-th eigenpair with high probability. Meanwhile, $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$ only differ in the between-mega-community entries which are only determined by $p_\emptyset$. When $p_\emptyset \ll \bar{p}^*$, it turns out that

$$\|\boldsymbol{L} - \tilde{\boldsymbol{L}}\| \lesssim \sqrt{np_\emptyset \log n} \ll \|\boldsymbol{L} - \boldsymbol{L}^*\| \asymp \sqrt{n\bar{p}^* \log n}.$$

To apply Davis-Kahan $\sin\Theta$ Theorem on the decomposition $\boldsymbol{L} = \tilde{\boldsymbol{L}} + (\boldsymbol{L} - \tilde{\boldsymbol{L}})$, we still need to bound the eigengap of $\tilde{\boldsymbol{L}}$. Since $\tilde{\boldsymbol{L}}$ is a graph Laplacian, it also preserves the n-th eigenvalue. Therefore, it remains to bound $\lambda_{n-2}(\tilde{\boldsymbol{L}})$ from below. Note that $\mathbb{E}[\tilde{\boldsymbol{L}}] = \boldsymbol{L}^*$. A natural lower bound can be obtained via Weyl's inequality:

$$\lambda_{n-2}(\tilde{\boldsymbol{L}}) \geq \lambda_{n-2}(\boldsymbol{L}^*) - \|\tilde{\boldsymbol{L}} - \boldsymbol{L}^*\|.$$

However, the multi-scale issue persists because $\lambda_{n-2}(\boldsymbol{L}^*)$ only involves $(p_\emptyset, p_0, p_1)$ while $\|\tilde{\boldsymbol{L}} - \boldsymbol{L}^*\|$ involves all connection probabilities.

The only exception is when the binary tree $\mathcal{T}$ only involves two leaf nodes 0 and 1, in which case the multi-scale issue disappears. As a result, it is sufficient to show that a deeper tree always increases $\lambda_{n-2}(\tilde{\boldsymbol{L}})$ compared to the simple 2-leaf tree. To this end, we can generate another graph $\widetilde{\boldsymbol{A}}'$ by dampening the entries $\widetilde{\boldsymbol{A}}$, with $\mathbb{E}[\widetilde{\boldsymbol{A}}']$ being the connection probability matrix corresponding to the 2-leaf tree. This can be achieved by multiplying $\widetilde{\boldsymbol{A}}_{ij}$ by a Bernoulli random variable with parameter p_s/P_{ij} for any $i, j \in \mathcal{G}_s$. Recalling that the unnormalized graph Laplacian becomes larger in the positive semidefinite ordering when the adjacency matrix increases entrywise (Lemma 2.5.2), we obtain that $\tilde{\boldsymbol{L}} \succeq \tilde{\boldsymbol{L}}'$, where $\tilde{\boldsymbol{L}}'$ is the graph Laplacian for $\widetilde{\boldsymbol{A}}'$. By Weyl's inequality,

$$\lambda_{n-2}(\tilde{\boldsymbol{L}}) \geq \lambda_{n-2}(\tilde{\boldsymbol{L}}') + \lambda_n(\tilde{\boldsymbol{L}} - \tilde{\boldsymbol{L}}') \geq \lambda_{n-2}(\tilde{\boldsymbol{L}}').$$

Therefore, we reduce the general BTSBM to the simple 2-leaf case and hence avoid the multi-scale issue completely.

2.3.5. Comparison with previous theoretical results. Although consistency of graph Laplacian-based spectral clustering under the general BTSBM has been studied by Balakrishnan et al. (2011), their regularity conditions only hold for dense networks. In particular, they consider weighted graphs with sub-Gaussian weights, including the Bernoulli weight in a network as a special case whose sub-Gaussian parameter is 1. According to Theorem 1 in Balakrishnan et al. (2011), to recover the first level of the hierarchy, they require $\gamma^4 \sqrt[4]{n/\log n} = \omega(1)$ where $\gamma = \min\{p_0, p_1\} - p_\emptyset$. As a result, the minimal expected degree $n\underline{p}_* = \Omega(n \min\{p_0, p_1\}) = \omega(n^{15/16})$. Therefore, their

strong consistency guarantee is only valid for very dense networks. In contrast, our Theorem 2.3.1 holds for sparse networks with $n\bar{p}^* = \Omega(\log n)$.

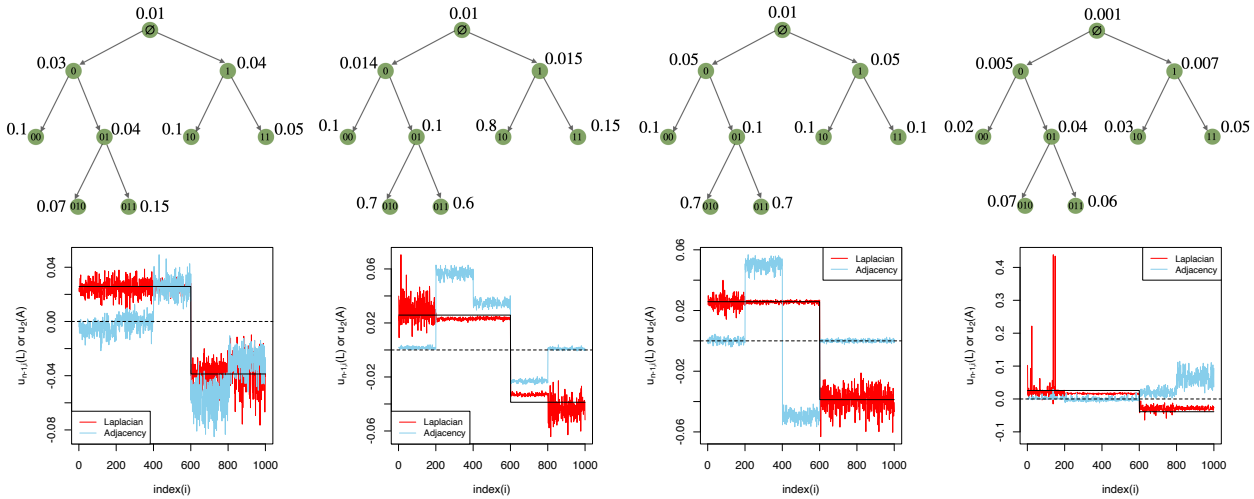
On the other hand, Li et al. (2018) and Lei (2019) derive the analogue of Theorem 2.3.1 for the adjacency matrix under a restrictive BTSBM where $\mathcal{T}$ is a full balanced binary tree and $(p_s, n_s) = (p_{s'}, n_{s'})$ whenever $|s| = |s'|$. Both analyses work for sparse networks; the former allows the expected degree to be $O(\log^{2+\epsilon} n)$ for $\epsilon > 0$ while the latter improves the dependence to the critical regime $O(\log n)$. A crucial property of the restrictive BTSBM is the strict homogeneity in the expected degrees, for which the population Laplacian has the same eigenvectors as the expected adjacency matrix. In addition, both works consider the traditional setting where all connection probabilities are of the same order and hence exclude multi-scale networks. Therefore, our analysis can be viewed as a substantial generalization.

2.4. Experiments

2.4.1. Synthetic networks. We generate synthetic networks from general BTSBMs with $\mathcal{T}$ being the simple binary tree presented in Figure 2.1 and different connection probabilities presented in the top panels of Figure 2.2. Each model has $n = 1000$ units and 200 units in each primitive cluster. The bottom panels of Figure 2.2 compares $\boldsymbol{u}_{n-1}(\boldsymbol{L})$, the Fiedler eigenvector of the unnormalized graph Laplacian, and $\boldsymbol{u}_{n-1}(\boldsymbol{L}^*)$, the population Fiedler vector. This can be viewed as an empirical check of the ℓ_∞ perturbation bound (Theorem 2.3.1) and the strong consistency of the first split. As a comparison, we also plot $\boldsymbol{u}_2(\boldsymbol{A})$, the eigenvector corresponding to the second largest eigenvalue of the adjacency matrix, which is considered in Li et al. (2018) and Lei (2019).

Figure 2.2a shows a setting where the connection probabilities are roughly of the same order. The signs of the Fiedler vector $\boldsymbol{u}_{n-1}(\boldsymbol{L})$ perfectly align with the mega-community memberships given by the first split, while $\boldsymbol{u}_2(\boldsymbol{A})$ messes up with the mega-community $\mathcal{G}_0$.

Figure 2.2b considers a highly multi-scale setting where $\bar{p}^* \gg \underline{p}_*$. This poses a potential threat to the degree variation condition (2.6) in Theorem 2.3.1. While the ℓ_∞ error of the Fiedler vector grows substantially compared to Figure 2.2a, the signs of the entries still perfectly identify the first split. An intriguing observation is the asymmetry of entrywise errors; whereas the overall ℓ_∞ perturbation error is sufficiently large to flip the sign of an entry, it is mainly contributed by



(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4

FIGURE 2.2. general BTSBMs (top) and the associated eigenvectors (bottom), including $\boldsymbol{u}_{n-1}(\boldsymbol{L}^*)$ (black), $\boldsymbol{u}_{n-1}(\boldsymbol{L})$ (red), and the $\boldsymbol{u}_2(\boldsymbol{A})$ (blue)

"outbound" deviations that have no effect on the sign while "inbound" deviations that pull entries to the other side of the axis are much smaller. This illustrates the potential suboptimality of our strategy to deduce the strong consistency from a small ℓ_∞ perturbation error. The phenomenon has been studied by Abbe et al. (2017) and Deng et al. (2020) for standard SBMs with $K = 2$ and by Lei (2019) for BTSBM. We leave the refined analysis for general BTSBM for future work.

Figure 2.2c examines a slight misspecification of the balanced BTSBM studied in Li et al. (2018) where the connection probabilities are identical within each level. Then $\boldsymbol{u}_2(\boldsymbol{A})$ fails to identify the first split while the Fiedler vector corresponding to the graph Laplacian works as desired. This illustrates the sensitivity of adjacency matrix-based spectral clustering to the model misspecification.

Figure 2.2d presents a setting with tiny connection probabilities, resulting in a very sparse network. In this case, the Fiedler vector exhibits a few spikes, thereby forcing the other values to be small. We observe this pattern frequently in repeated experiments. This suggests that the eigengap condition (2.5) fails to hold. In this case, it is not surprising that the performance, in terms of strong or weak consistency, degrades drastically. Our observation is in line with Krzakala et al. (2013) that spectral clustering based on the adjacency matrix or graph Laplacians cannot handle networks that are too sparse.

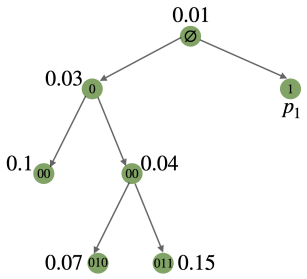
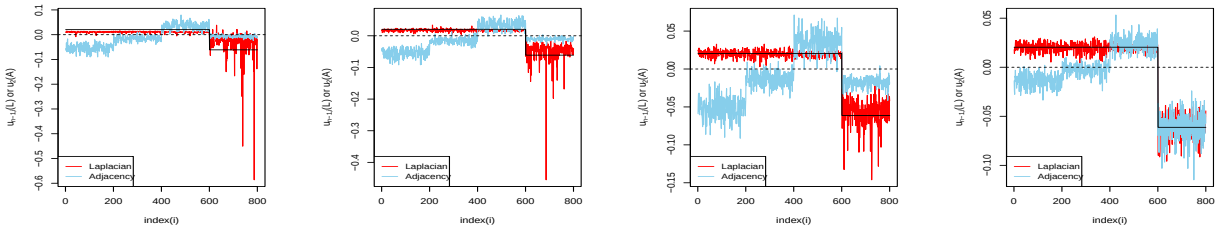


FIGURE 2.3. Unbalanced tree

We also consider the case when the binary tree in general BTSBM is more unbalanced. We generate synthetic networks from general BTSBMs with $\mathcal{T}$ being a slightly different binary tree with that in Figure 2.1, where we have three primitive clusters versus one primitive clusters in the first split. The tree structure and associated connection probabilities are presented in Figure 2.3. The number of nodes in each primitive cluster is still 200. We compare the aforementioned eigenvectors under different values of p_1 in Figure 2.4.



(a) $p_1 = 0.04$ (b) $p_1 = 0.06$ (c) $p_1 = 0.08$ (d) $p_1 = 0.1$

FIGURE 2.4. Eigenvectors including $\boldsymbol{u}_{n-1}(\boldsymbol{L}^*)$ (black), $\boldsymbol{u}_{n-1}(\boldsymbol{L})$ (red), and the $\boldsymbol{u}_2(\boldsymbol{A})$ (blue) under different values of p_1.

From Figure 2.4, we can observe that, as p_1 increases, the Fiedler vector is less spiky and the deviation of $\boldsymbol{u}_{n-1}(\boldsymbol{L})$ from $\boldsymbol{u}_{n-1}(\boldsymbol{L}^*)$ becomes smaller. In this case, n_1 is much smaller than n_0, rendering the eigengap condition (2.5) harder to satisfy. Nevertheless, as long as p_1 exceeds certain threshold, the eigengap condition becomes plausible and the Fiedler vector is able to perfectly identify the mega-communities $\mathcal{G}_0$ and $\mathcal{G}_1$.

2.4.2. Real-world networks. In this section, we compare recursive spectral bi-clustering algorithms based on the adjacency matrix $\boldsymbol{A}$, the unnormalized graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ and the normalized graph Laplacian $\boldsymbol{N} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ on seven real-world networks, summarized in

Table 2.1. All networks contain explicit information regarding the true community memberships, which we use to evaluate the performance of clustering algorithms; see the references in the second column of Table 2.1 for more details.

Dataset	Source	$\lvert V\rvert$	$\lvert E\rvert$	K	$d_{\min}$	$d_{\max}$	$\bar{d}$
Dolphins	Lusseau et al. (2003)	62	159	2	1	12	5.129
Karate	Zachary (1977)	34	78	2	1	17	4.588
Political books	Krebs (unpublished)	92	374	2	1	24	8.130
Political blogs	Adamic and Glance (2005)	1222	16714	2	1	351	27.355
UK faculty	Nepusz et al. (2008)	79	552	3	2	39	13.975
Football	Girvan and Newman (2002)	110	570	11	7	13	10.364
C. elegans	Jarrell et al. (2012)	229	1085	6	1	34	9.585

TABLE 2.1. Seven network datasets

We evaluate the performance of clustering algorithms via the completeness score (Rosenberg and Hirschberg, 2007), an external entropy-based cluster evaluation measure. A clustering result satisfies the completeness if all vertices that are members of a true (primitive) community reside in the same estimated community. Equivalently, each estimated community from a complete clustering must be the union of a subset of true (primitive) communities. Grouping all of the vertices into a single community is an extreme example of a complete clustering. The completeness score is designed to measure the proximity to completeness. Suppose the true communities are $V_1, \ldots, V_K$ and the estimated communities are $\hat{V}_1, \ldots, \hat{V}_{\hat{K}}$, where $\hat{K}$ might differ from K. The completeness score is defined as

$$c(\hat{V}, V) = \begin{cases} 1 & \text{if } H(\hat{V}) = 0 \\ 1 - \frac{H(\hat{V}|V)}{H(\hat{V})} & \text{otherwise} \end{cases}, \tag{2.9}$$

where $H(\hat{V}|V)$ is the conditional entropy of the estimated clusters given the true community assignments and $H(\hat{V})$ is the entropy of the estimated clusters, i.e.,

$$H(\hat{V}|V) = -\sum_{i=1}^{K}\sum_{j=1}^{\hat{K}} \frac{|V_i \cap \hat{V}_j|}{n} \log \frac{|V_i \cap \hat{V}_j|}{|\hat{V}_j|}, \quad H(\hat{V}) = -\sum_{j=1}^{\hat{K}} \frac{|\hat{V}_j|}{n} \log \frac{|\hat{V}_j|}{n}.$$

Clearly, the completeness score (2.9) takes value in $[0, 1]$ and a value 1 implies that the clustering is complete. This metric is invariant to label permutations and asymmetric in V and $\hat{V}$. The

asymmetry renders the completeness score a proper metric to evaluate the performance of recovering mega-communities.

Dataset	$\boldsymbol{A}$	$\boldsymbol{L}$	$\boldsymbol{N}$
Dolphins	0.470	**1**	0.883
Karate	**1**	0.840	0.840
Political books	0.823	**0.869**	**0.869**
Political blogs	**0.675**	0.007	0.012
UK faculty	0.765	0.908	**1**
Football	0.763	**0.802**	**0.802**
C. elegans	0.416	**0.939**	0.807

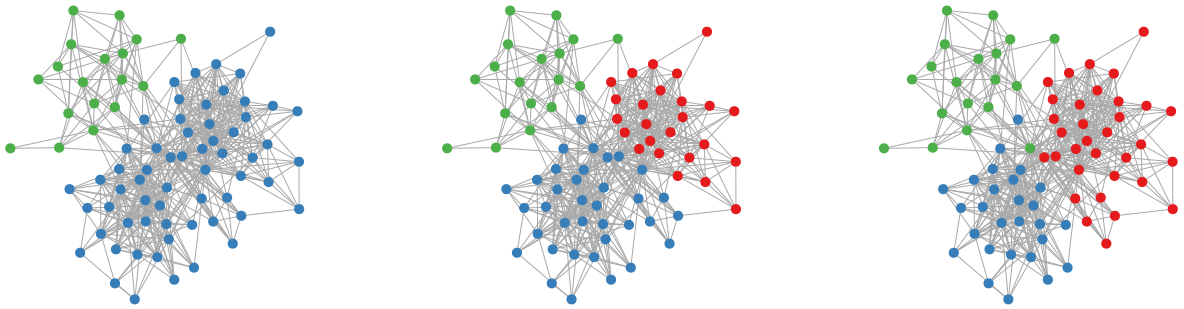
TABLE 2.2. Completeness scores for the first split

For each of the three recursive bi-partitioning algorithms, we compute the completeness score for the first split. The results are reported in Table 2.2. Notably, the unnormalized graph Laplacian-based algorithm performs well on all networks but Political blogs, which has substantially higher degree variation as shown in Table 2.1. This partly corroborates our theory that the degree variation plays an important role.

Next, we move deeper into the estimated hierarchy and evaluate the performance of (partial) hierarchy recovery. For illustration, we investigate three networks: UK faculty, Football, and C. elegans. We will examine the performance of the first few splits based on the completeness scores.

UK faculty is a personal friendship network of the academic staffs in a UK university, which consists of three separate schools. These three separate schools are treated as the true primitive communities; see Figure 2.5c. Figure 2.5a and 2.5b display the first and second splits given by the recursive bi-partitioning algorithm based on the Fiedler vector. The first split separates the green community from the others and achieves a high completeness score 0.908, suggesting that it captures two meaningful mega-communities in this network. Unsurprisingly, The second split has a lower completeness score because the connections within or between the red and blue vertices are similar and thus it is harder to distinguish them. Compared to the ground truth, our algorithm performs reasonably well in recovering both the primitive communities and the hierarchy.

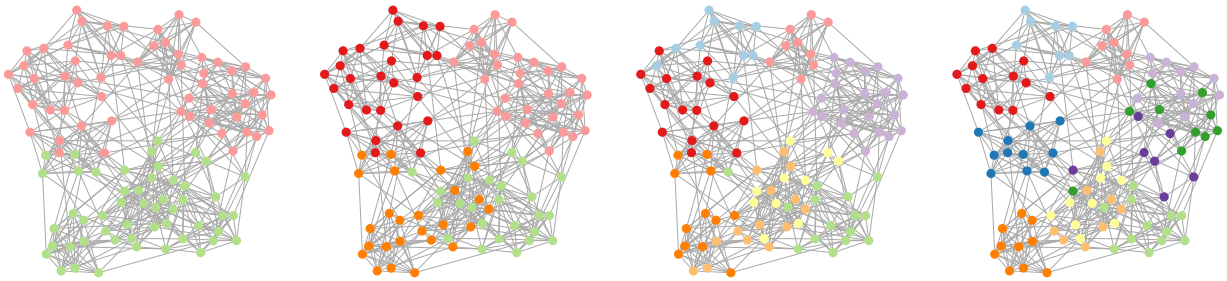
Football is a network of American college football teams during the regular season Fall 2000. The



(a) First split (0.908) (b) Second split (0.735) (c) Ground truth

FIGURE 2.5. Spectral recursive bi-partition on the UK faculty network, with the completeness scores in the parentheses

vertices represent teams and edges represent regular season games between the two teams they connect. The teams are divided into "conferences" containing around 8 to 12 teams each, which can be treated as the true primitive communities; see Figure 2.6d. Again, we apply the recursive bi-partitioning algorithm based on the Fiedler vector. Here, we build a balanced hierarchy of depth three without resorting to any stopping rule. Figure 2.6a - 2.6c show the first, second and third level of an estimated hierarchy, with 2, 4, and 8 resulting clusters respectively. Each level has a high completeness score, suggesting that the estimated hierarchy is meaningful.

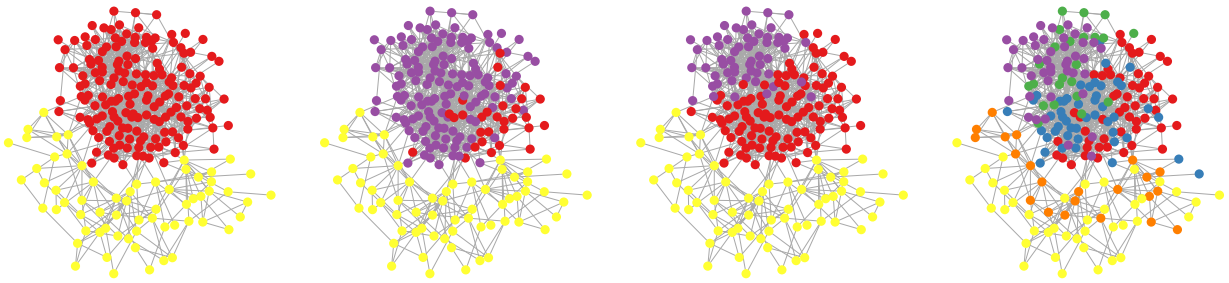


(a) First level (0.802) (b) Second level (0.828) (c) Third level (0.810) (d) Ground truth

FIGURE 2.6. Spectral recursive bi-partition on the Football network, with the completeness scores in the parentheses

C. elegans is a neural network consisting of gap junctional synaptic connections in the posterior nervous system of a single adult male of *Caenorhabditis elegans*, a primitive worm. The cells are grouped according to the modules and categories described in Jarrell et al. (2012). Specifically,

there are six types of cells: sensory neurons, interneurons, gender-shared neurons, command and motor neurons, gender-shared muscle cells, sex-specific muscle cells. We treat the cell types as the true primitive communities; see Figure 2.7d. Figure 2.7a shows the first split given by the spectral bi-partition algorithm based on the Fiedler vector. It performs well in the first split as suggested by the completeness score 0.939. Interestingly, the two mega-communities correspond to neurons and muscle cells precisely. Figure 2.7b shows the second split, which has a degraded performance. As a comparison, we also show the second split given by the normalized graph Laplacian in Figure 2.7c, which has a higher completeness score. They mainly differ in whether the blue nodes are merged with purple or red nodes.



(a) First split (0.939) (b) Second split (0.667) (c) Second split (0.752) (d) Ground truth

FIGURE 2.7. Spectral recursive bi-partition on the C. elegans network, with the completeness scores in the parentheses. (a) and (b) are based on the unnormalized graph Laplacian and (c) is based on the normalized graph Laplacian

2.5. Proofs of main results

2.5.1. Supporting lemmata.

Lemma 2.5.1 (Lemma 3.8 of Lei (2019)). *Let $\boldsymbol{L}$ denote the unnormalized graph Laplacian of a random unweighted graph with independent edges, and $\mathbb{E}(\boldsymbol{L}) = \boldsymbol{L}^*$. Then for any $r > 0$, there exists an absolute constant $C(r)$ that only depends on r, such that, with probability at least $1 - n^{-r}$,*

$$\|\boldsymbol{L} - \boldsymbol{L}^*\|_{\text{op}} \leq C(r)\sqrt{\left(\max_{1\leq i\leq n} L^*_{ii} + \log n\right)\log n}.$$

Lemma 2.5.2 (Lemma 10 of Balakrishnan et al. (2011)). *Let $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$ be two adjacency matrices with unnormalized Laplacians $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$, respectively. If $\widetilde{\boldsymbol{A}}_{ij} \geq \boldsymbol{A}_{ij}$ for any pair (i, j), then $\tilde{\boldsymbol{L}} - \boldsymbol{L}$ is positive semidefinite and $\lambda_k(\tilde{\boldsymbol{L}}) \geq \lambda_k(\boldsymbol{L})$ for all k.*

Lemma 2.5.3 (Theorem 2 of Yu et al. (2014), a variant of Davis-Kahan $\sin\Theta$ Theorem). *Let $\boldsymbol{A}, \boldsymbol{A}^* \in \mathbb{R}^{n\times n}$ be symmetric matrices. Fix positive integers s, d and let $\boldsymbol{U} = (\boldsymbol{u}_s(\boldsymbol{A}), \ldots, \boldsymbol{u}_{s+d-1}(\boldsymbol{A})) \in \mathbb{R}^{n\times d}$ and $\boldsymbol{U}^* = (\boldsymbol{u}_s(\boldsymbol{A}^*), \ldots, \boldsymbol{u}_{s+d-1}(\boldsymbol{A}^*)) \in \mathbb{R}^{n\times d}$. Further let $\Theta(\boldsymbol{U}, \boldsymbol{U}^*) \in \mathbb{R}^{d\times d}$ denote the principal angle matrix between the subspaces spanned by $\boldsymbol{U}$ and $\boldsymbol{U}^*$. Then*

$$\inf_{\substack{\boldsymbol{O}\in\mathbb{R}^{d\times d}\\ \boldsymbol{O}^T\boldsymbol{O}=I_d}} \|\boldsymbol{U}\boldsymbol{O} - \boldsymbol{U}^*\|_F = 2^{1/2} \|\sin\Theta(\boldsymbol{U}, \boldsymbol{U}^*)\|_F \leq \frac{2^{3/2} d^{1/2} \|\boldsymbol{A} - \boldsymbol{A}^*\|_{\mathrm{op}}}{\min\{\lambda_{s-1}(\boldsymbol{A}^*) - \lambda_s(\boldsymbol{A}^*), \lambda_{s+d-1}(\boldsymbol{A}^*) - \lambda_{s+d}(\boldsymbol{A}^*)\}}.$$

2.5.2. Proof of Theorem 2.2.1. We first state a more general result that characterizes the entire eigenstructure of $\boldsymbol{L}^*$ under the general BTSBM.

Lemma 2.5.4. *Denote*

$$g(s; \mathcal{T}) = \begin{cases} 1 & (s \text{ is an internal node}) \\ n_s - 1 & (s \text{ is a leaf node}) \end{cases}.$$

The eigenstructure of $\boldsymbol{L}^*$ has the following properties:

(1) $\lambda_n^* = 0$, $\boldsymbol{u}_n^* = \frac{1}{\sqrt{n}}[1, 1, \ldots, 1]^\top = \boldsymbol{1}_n/\sqrt{n}$, and $\lambda_{n-1}^* > 0$.

(2) For each node $s = b_1 b_2 \ldots b_{|s|}$, let $s_{(i)} = b_1 b_2 \ldots b_{|s|-i}$ and $s_{(|s|)} = \emptyset$. Then

$$\lambda^*(s; \mathcal{T}) \triangleq n_s p_s + \sum_{i=1}^{|s|} (n_{s_{(i)}} - n_{s_{(i-1)}}) p_{s_{(i)}}, \tag{2.10}$$

is an eigenvalue of $\boldsymbol{L}^*$ with multiplicity

$$\sum_{s': \lambda^*(s';\mathcal{T}) = \lambda^*(s;\mathcal{T})} g(s'; \mathcal{T}).$$

(3) The eigenspace corresponding to $\lambda^*(s;\mathcal{T})$ *is spanned by*

$$\bigcup_{s':\lambda^*(s';\mathcal{T})=\lambda^*(s;\mathcal{T})} \text{colspan}\left(\boldsymbol{U}(s';\mathcal{T})\right)$$

where $\boldsymbol{U}(s;\mathcal{T}) \in \mathbb{R}^{n\times g(s;\mathcal{T})}$ *such that*

- *if* s *is an internal node,*

$$\boldsymbol{U}_i(s;\mathcal{T}) = \begin{cases} \sqrt{n_{R(s)}/n_{L(s)}n_s} & i \in \mathcal{G}_{L(s)} \\ -\sqrt{n_{L(s)}/n_{R(s)}n_s} & i \in \mathcal{G}_{R(s)} \\ 0 & \textit{otherwise} \end{cases};$$

- *if* s *is a leaf node,* $\boldsymbol{U}_{\mathcal{G}_s^c}(s;\mathcal{T}) = \mathbf{0}_{(n-n_s)\times(n_s-1)}$ *and* $\boldsymbol{U}_{\mathcal{G}_s}(s;\mathcal{T}) \in \mathbb{R}^{n_s\times(n_s-1)}$ *is any orthogonal matrix with* $\boldsymbol{1}_{n_s}^\top \boldsymbol{U}_{\mathcal{G}_s}(s;\mathcal{T}) = \mathbf{0}^\top$.

Lemma 2.5.4 can be verified through simple algebra. By Lemma 2.5.4, it is straightforward to prove Theorem 2.2.1.

Proof of Theorem 2.2.1. (1) By Lemma 2.5.4, $np_\emptyset$ is an eigenvalue, corresponding to the root node and $g(\emptyset;\mathcal{T}) = 1$ since it is an internal node. Under weak assortativity, for any node $s \in \mathcal{T}$,

$$\lambda^*(s;\mathcal{T}) > n_s p_\emptyset + \sum_{i=1}^{|s|}(n_{s_{(i)}} - n_{s_{(i-1)}})p_\emptyset = np_\emptyset.$$

Therefore, $\lambda^*_{n-1} = np_\emptyset$ with multiplicity 1.

(2) By Lemma 2.5.4, $n_1p_1 + n_0p_\emptyset$ and $n_0p_0 + n_1p_\emptyset$ are both eigenvalues corresponding to node 1 and 0, respectively. For all other nodes $s \in \mathcal{T}$, it is easy to show that $\lambda^*(s;\mathcal{T}) > n_0p_0 + n_1p_\emptyset$ if s is a descendant of node 0, and $\lambda^*(s;\mathcal{T}) > n_1p_1 + n_0p_\emptyset$ if s is a descendant of node 1. Therefore, λ^*_{n-2} must be their minimum.

(3) For each leaf node s, it is not hard to see that

$$\lambda^*(s;\mathcal{T}) = \sum_{j=1}^{n} p_{ij}, \quad \forall i \text{ in the community } s.$$

By definition,

$$\lambda^*(s;\mathcal{T}) \geq n\underline{p}_*.$$

The number of eigenvalues that are at least $n\underline{p}_*$, accounting for multiplicity, is at least

$$\sum_{\text{leaf node } s} g(s;\mathcal{T}) = \sum_{\text{leaf node } s} (n_s - 1) = n - K.$$

(4) Based on the previous part in this proof, if the eigenvalue $\lambda_j^* < n\underline{p}_*$, then λ_j^* must correspond to an internal node. Then the part (3) of Lemma 2.5.4 implies that

$$\|\boldsymbol{u}_j^*\|_\infty = \frac{1}{\sqrt{n}} \max\left\{\sqrt{\frac{n_{R(s)}}{n_{L(s)}}}, \sqrt{\frac{n_{L(s)}}{n_{R(s)}}}\right\} \le \sqrt{\frac{\xi}{n}}.$$

□

2.5.3. Proof of Lemma 2.3.3. Let $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{A}}'$ be defined as in Section 2.3.4, i.e.,

$$\widetilde{\boldsymbol{A}}_{ij} = \begin{cases} \boldsymbol{A}_{ij} & (i,j \in \mathcal{G}_0, \text{ or } i,j \in \mathcal{G}_1) \\ p_\emptyset & (\text{otherwise}) \end{cases}, \quad \widetilde{\boldsymbol{A}}'_{ij} = \begin{cases} \boldsymbol{A}_{ij}\boldsymbol{B}_{ij} & (i,j \in \mathcal{G}_0, \text{ or } i,j \in \mathcal{G}_1) \\ p_\emptyset & (\text{otherwise}) \end{cases},$$

where $\{\boldsymbol{B}_{ij} : i < j\}$ are independent Bernoulli random variables that are independent of $\boldsymbol{A}$ with $\mathbb{E}[\boldsymbol{B}_{ij}] = p_0/\mathbb{E}[\boldsymbol{A}_{ij}]$ if $i,j \in \mathcal{G}_0$ and $\mathbb{E}[\boldsymbol{B}_{ij}] = p_1/\mathbb{E}[\boldsymbol{A}_{ij}]$ if $i,j \in \mathcal{G}_1$. Then $\widetilde{\boldsymbol{A}}'$ is the adjacency matrix given by a general BTSBM with a 2-leaf tree and parameters $(p_\emptyset, p_0, p_1)$. Further let $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{L}}'$ be their unnormalized graph Laplacians. Since $\widetilde{\boldsymbol{A}}_{ij} \ge \widetilde{\boldsymbol{A}}'_{ij}$ for all pairs (i,j), by Lemma 2.5.2,

$$\lambda_{n-2}(\tilde{\boldsymbol{L}}) \ge \lambda_{n-2}(\tilde{\boldsymbol{L}}'). \tag{2.11}$$

Note that

$$\max_{1\le i\le n} \mathbb{E}[\tilde{\boldsymbol{L}}'_{ii}] = \max\{n_0p_0 + n_1p_\emptyset, n_1p_1 + n_0p_\emptyset\} \le n_0p_0 + n_1p_1.$$

By Weyl's inequality and Lemma 2.5.1, with probability $1 - n^{-r}$,

$$\lambda_{n-2}(\tilde{\boldsymbol{L}}') \ge \lambda_{n-2}(\mathbb{E}[\tilde{\boldsymbol{L}}']) - \|\tilde{\boldsymbol{L}}' - \mathbb{E}[\tilde{\boldsymbol{L}}']\| \ge \lambda_{n-2}(\mathbb{E}[\tilde{\boldsymbol{L}}']) - C(r)\sqrt{(n_0p_0 + n_1p_1 + \log n)\log n}.$$

By Lemma 2.5.4,

$$\lambda_{n-2}(\mathbb{E}[\tilde{\boldsymbol{L}}']) = \min\{n_0p_0 + n_1p_\emptyset, n_1p_1 + n_0p_\emptyset\} = np_\emptyset + \min\{n_0(p_0 - p_\emptyset), n_1(p_1 - p_\emptyset)\}.$$

The condition (2.7) implies

$$n_0p_0 + n_1p_1 \geq C_{\ell_2}\sqrt{(n_0p_0 + n_1p_1)\log n} \Longrightarrow n_0p_0 + n_1p_1 \geq C_{\ell_2}^2 \log n. \tag{2.12}$$

When $C_{\ell_2} > \max\{1, 3C(r)\}$,

$$\begin{aligned}\lambda_{n-2}(\tilde{\boldsymbol{L}}') &\geq np_\emptyset + \left\{C_{\ell_2} - C(r)\sqrt{1 + \frac{1}{C_{\ell_2}^2}}\right\}\sqrt{(n_0p_0 + n_1p_1)\log n} \\ &\geq np_\emptyset + \frac{C_{\ell_2}}{2}\sqrt{(n_0p_0 + n_1p_1)\log n} \end{aligned} \tag{2.13}$$

By (2.11),

$$\lambda_{n-2}(\tilde{\boldsymbol{L}}) > np_\emptyset, \quad \text{with probability } 1 - n^{-r}. \tag{2.14}$$

On the other hand, for any $i \in \mathcal{G}_0$,

$$(\tilde{\boldsymbol{L}}\boldsymbol{u}_{n-1}^*)_i = \sum_{j=1}^{n} \tilde{\boldsymbol{L}}_{ij}\boldsymbol{u}_{n-1,j}^* = \sqrt{\frac{n_1}{n_0 n}}\sum_{j\in\mathcal{G}_0}\tilde{\boldsymbol{L}}_{ij} - \sqrt{\frac{n_0}{n_1 n}}\sum_{j\in\mathcal{G}_1}\tilde{\boldsymbol{L}}_{ij}.$$

By definition,

$$\sum_{j\in\mathcal{G}_0}\tilde{\boldsymbol{L}}_{ij} = \tilde{\boldsymbol{L}}_{ii} - \sum_{j\in\mathcal{G}_0\setminus\{i\}}\boldsymbol{A}_{ij} = \sum_{j\in\mathcal{G}_1\cup\{i\}}\boldsymbol{A}_{ij} = n_1p_\emptyset = -\sum_{j\in\mathcal{G}_1}\tilde{\boldsymbol{L}}_{ij}.$$

Thus,

$$(\tilde{\boldsymbol{L}}\boldsymbol{u}_{n-1}^*)_i = \left(\sqrt{\frac{n_1}{n_0 n}} + \sqrt{\frac{n_0}{n_1 n}}\right)n_1p_\emptyset = \sqrt{\frac{n_1}{n_0 n}}np_\emptyset = (np_\emptyset)\boldsymbol{u}_{n-1,i}^*.$$

Similarly, for any $i \in \mathcal{G}_1$,

$$(\tilde{\boldsymbol{L}}\boldsymbol{u}_{n-1}^*)_i = (np_\emptyset)\boldsymbol{u}_{n-1,i}^*.$$

As a result, $(np_\emptyset, \boldsymbol{u}_{n-1}^*)$ is an eigenpair of $\tilde{\boldsymbol{L}}$. Since $\tilde{\boldsymbol{L}}$ is an unnormalized Laplacian, 0 is always an eigenvalue. Therefore, on the event (2.14), which occurs with probability at least $1 - n^{-r}$, $\lambda_{n-1}(\tilde{\boldsymbol{L}}) = np_\emptyset$ with multiplicity 1 and $\boldsymbol{u}_{n-1}(\tilde{\boldsymbol{L}}) = \boldsymbol{u}_{n-1}^*$.

2.5.4. Proof of Theorem 2.3.2. Let $\mathcal{E}$ denote the event given by (2.13). As shown in the proof of Lemma 2.3.3,

$$\mathbb{P}(\mathcal{E}) \geq 1 - n^{-r},$$

if $C_{\ell_2} > \max\{1, 3C(r)\}$. On the event $\mathcal{E}$, by (2.11) and Lemma 2.3.3,

$$\lambda_{n-1}(\tilde{\boldsymbol{L}}) = np_\emptyset, \quad \lambda_{n-2}(\tilde{\boldsymbol{L}}) - \lambda_{n-1}(\tilde{\boldsymbol{L}}) \geq \frac{C_{\ell_2}}{2}\sqrt{(n_0p_0 + n_1p_1)\log n}, \quad \boldsymbol{u}_{n-1}(\tilde{\boldsymbol{L}}) = \boldsymbol{u}^*_{n-1}.$$

Fix any $\nu > 0$. Let

$$\boldsymbol{L}_\nu = \boldsymbol{L} + \nu\boldsymbol{J}, \quad \tilde{\boldsymbol{L}}_\nu = \tilde{\boldsymbol{L}} + \nu\boldsymbol{J}, \quad \text{where } \boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T.$$

Since both $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$ are unnormalized graph Laplacians, $\lambda_n(\boldsymbol{L}) = \lambda_n(\tilde{\boldsymbol{L}}) = 0, \boldsymbol{u}_n(\boldsymbol{L}) = \boldsymbol{u}_n(\tilde{\boldsymbol{L}}) = \mathbf{1}_n/\sqrt{n}$. In addition, for any $\nu > 0$,

$$\boldsymbol{u}_{n-1}(\boldsymbol{L}_\nu) = \boldsymbol{u}_{n-1}(\boldsymbol{L}), \quad \boldsymbol{u}_{n-1}(\tilde{\boldsymbol{L}}_\nu) = \boldsymbol{u}_{n-1}(\tilde{\boldsymbol{L}}),$$

and for $j \geq 1$,

$$\lambda_{n-j}(\boldsymbol{L}_\nu) = \lambda_{n-j}(\boldsymbol{L}) + \nu, \quad \lambda_{n-j}(\tilde{\boldsymbol{L}}_\nu) = \lambda_{n-j}(\tilde{\boldsymbol{L}}) + \nu.$$

On the event $\mathcal{E}$,

$$\min\{\lambda_{n-2}(\tilde{\boldsymbol{L}}) - \lambda_{n-1}(\tilde{\boldsymbol{L}}), \lambda_{n-1}(\tilde{\boldsymbol{L}}) - \lambda_n(\tilde{\boldsymbol{L}})\} \geq \min\left\{\frac{C_{\ell_2}}{2}\sqrt{(n_0p_0 + n_1p_1)\log n}, np_\emptyset + \nu\right\}. \tag{2.15}$$

By Davis-Kahan Theorem (Lemma 2.5.3),

$$\begin{aligned}\|\boldsymbol{u}_{n-1}\text{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}^*_{n-1}) - \boldsymbol{u}^*_{n-1}\|_2 &\leq \frac{2^{3/2}\|\boldsymbol{L}_\nu - \tilde{\boldsymbol{L}}_\nu\|}{\min\{\lambda_{n-2}(\tilde{\boldsymbol{L}}) - \lambda_{n-1}(\tilde{\boldsymbol{L}}), \lambda_{n-1}(\tilde{\boldsymbol{L}}) - \lambda_n(\tilde{\boldsymbol{L}})\}} \\ &\leq \frac{2^{3/2}\|\boldsymbol{L} - \tilde{\boldsymbol{L}}\|}{\min\left\{\frac{C_{\ell_2}}{2}\sqrt{(n_0p_0 + n_1p_1)\log n}, np_\emptyset + \nu\right\}}\end{aligned}$$

By definition, $\boldsymbol{L} - \tilde{\boldsymbol{L}}$ can be formulated as $\boldsymbol{L}_0 - \mathbb{E}[\boldsymbol{L}_0]$ where $\boldsymbol{L}_0$ is the unnormalized graph Laplacian for an adjacency matrix $\boldsymbol{A}_0$ with

$$\boldsymbol{A}_{0,ij} = \begin{cases} 0 & (i,j \in \mathcal{G}_0, \text{ or } i,j \in \mathcal{G}_1) \\ \boldsymbol{A}_{ij} & (\text{otherwise}) \end{cases}.$$

By Lemma 2.5.1, with probability at least $1 - n^{-r}$,

$$\|\boldsymbol{L} - \tilde{\boldsymbol{L}}\| \leq C(r)\sqrt{(np_\emptyset + \log n)\log n}.$$

Let $\nu = n\max\{p_1, p_0\}$. Then

$$\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_2 \leq \frac{2^{5/2}C(r)}{C_{\ell_2}}\frac{\sqrt{(np_\emptyset + \log n)\log n}}{\sqrt{(n_0p_0 + n_1p_1)\log n}}.$$

By weak assortativity and (2.12),

$$\sqrt{(np_\emptyset + \log n)\log n} \leq \sqrt{2(n_0p_0 + n_1p_1)\log n}.$$

Thus, with probability $1 - 2n^{-r}$,

$$\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_2 \leq \frac{8C(r)}{C_{\ell_2}}.$$

The proof is completed by setting $C_{\ell_2} = \max\{1, 8C(r)/c\}$ and replacing r with $r+1$.

2.5.5. Proof of Theorem 2.3.1. Without loss of generality we assume $\boldsymbol{u}_{n-1}^T\boldsymbol{u}_{n-1}^* \geq 0$. For each $j \in \{1, \ldots, n\}$ such that $\lambda_{n-j+1}^* < \lambda_{n-j}^*$, let $\boldsymbol{U}_j^* \in \mathbb{R}^{n\times j}$ and $\boldsymbol{U}_j \in \mathbb{R}^{n\times j}$ denote the eigenvector matrices $(\boldsymbol{u}_{n-1}^*, \ldots, \boldsymbol{u}_{n-j}^*)$ and $(\boldsymbol{u}_{n-1}, \ldots, \boldsymbol{u}_{n-j})$, respectively. Define

$$\boldsymbol{O}_j = \mathrm{sign}(\boldsymbol{U}_j^T\boldsymbol{U}_j^*),$$

where $\mathrm{sign}(\boldsymbol{M})$ denotes the matrix sign. Specifically, if $U\Sigma V^T$ is the singular value decomposition of $\boldsymbol{M}$, then $\mathrm{sign}(\boldsymbol{M}) = UV^T$. Since $\boldsymbol{O}_j \in \mathbb{R}^{j\times j}$ is an orthogonal matrix, we have

$$\sqrt{n}\|\boldsymbol{U}_j\boldsymbol{O}_j - \boldsymbol{U}_j^*\|_{2\to\infty} = \sqrt{n}\|\boldsymbol{U}_j - \boldsymbol{U}_j^*\boldsymbol{O}_j^T\|_{2\to\infty}.$$

Let $O_{j,1,i}$ denote the entry of $\boldsymbol{O}_j$ in the first row and i-th column and $\boldsymbol{U}_{j,-1}^*$ denote the matrix $\boldsymbol{U}_j^*$ with the first column $\boldsymbol{u}_{n-1}^*$ removed. Then

$$\begin{aligned}\sqrt{n}\|\boldsymbol{U}_j - \boldsymbol{U}_j^*\boldsymbol{O}_j^T\|_{2\to\infty} &\geq \sqrt{n}\left\|\boldsymbol{u}_{n-1} - \sum_{i=1}^{j} O_{j,1,i}\boldsymbol{u}_{n-i}^*\right\|_\infty \\ &\geq \sqrt{n}\|\boldsymbol{u}_{n-1} - O_{j,1,1}\boldsymbol{u}_{n-1}^*\|_\infty - \sqrt{n}\sqrt{\sum_{i\neq 1} O_{j,1,i}^2}\|\boldsymbol{U}_{j,-1}^*\|_{2\to\infty}.\end{aligned}$$

Furthermore, we know that

$$\begin{aligned}\sqrt{n}\|\boldsymbol{u}_{n-1} - O_{j,1,1}\boldsymbol{u}^*_{n-1}\|_\infty &= \sqrt{n}\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1} + \boldsymbol{u}^*_{n-1} - O_{j,1,1}\boldsymbol{u}^*_{n-1}\|_\infty \\ &\geq \sqrt{n}\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_\infty - \sqrt{n}|1 - O_{j,1,1}|\|\boldsymbol{u}^*_{n-1}\|_\infty \\ &= \sqrt{n}\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_\infty - |1 - O_{j,1,1}| \max\left\{\sqrt{\frac{n_1}{n_0}}, \sqrt{\frac{n_0}{n_1}}\right\} \\ &\geq \sqrt{n}\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_\infty - |1 - O_{j,1,1}|\sqrt{\xi}.\end{aligned}$$

The second last equality invokes Theorem 2.2.1. Since $\boldsymbol{O}_j$ is orthogonal,

$$\sum_{i=1}^{j} O^2_{j,1,i} = 1 \Longrightarrow \sum_{i\neq 1} O^2_{j,1,i} = 1 - O^2_{j,1,1} \leq 2(1 - O_{j,1,1}).$$

Also notice that $\|\boldsymbol{U}^*_{j,-1}\|_{2\to\infty} \leq \|\boldsymbol{U}^*_j\|_{2\to\infty}$. As a result,

$$\begin{aligned}&\sqrt{n}\|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_\infty \\ \leq&\sqrt{n}\|\boldsymbol{U}_j\boldsymbol{O}_j - \boldsymbol{U}^*_j\|_{2\to\infty} + \sqrt{2(1 - O_{j,1,1})}\left(\sqrt{n}\|\boldsymbol{U}^*_j\|_{2\to\infty}\right) + |1 - O_{j,1,1}|\sqrt{\xi}.\end{aligned} \tag{2.16}$$

To further simplify the second and the third terms, let $\boldsymbol{H}_j = \boldsymbol{U}_j^T\boldsymbol{U}^*_j$ with singular value decomposition $\boldsymbol{H}_j = \bar{\boldsymbol{U}}_j(\cos\Theta(\boldsymbol{U}_j, \boldsymbol{U}^*_j))\bar{\boldsymbol{V}}_j^T$, where $\cos\Theta(\boldsymbol{U}_j, \boldsymbol{U}^*_j) = \mathrm{diag}(\cos\theta_{j1}, \ldots, \cos\theta_{jj})$ and θ_{ji}'s are the principal angles between $\boldsymbol{u}_{n-i}$ and $\boldsymbol{u}^*_{n-i}$. By definition, $\boldsymbol{O}_j = \bar{\boldsymbol{U}}_j\bar{\boldsymbol{V}}_j^T$. As a result,

$$\|\boldsymbol{H}_j - \boldsymbol{O}_j\| = \|\bar{\boldsymbol{U}}_j(\boldsymbol{I} - \cos\Theta_j)\bar{\boldsymbol{V}}_j^T\| \leq \|\boldsymbol{I} - \cos\Theta(\boldsymbol{U}_j, \boldsymbol{U}^*_j)\|.$$

For any $\theta \leq \pi/2$,

$$1 - \cos\theta \leq 1 - \cos^2\theta = \sin^2\theta.$$

Therefore,

$$\|\boldsymbol{H}_j - \boldsymbol{O}_j\| \leq \|\sin\Theta(\boldsymbol{U}_j, \boldsymbol{U}^*_j))\|^2$$

On the other hand,

$$|1 - H_{j,1,1}| = |1 - \boldsymbol{u}^T_{n-1}\boldsymbol{u}^*_{n-1}| = |(\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1})^T\boldsymbol{u}^*_{n-1}| \leq \|\boldsymbol{u}_{n-1} - \boldsymbol{u}^*_{n-1}\|_2.$$

As a consequence,

$$\begin{aligned}&|1-O_{j,1,1}| \le |1-H_{j,1,1}| + |H_{j,1,1}-O_{j,1,1}|\\ \le&\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2 + \|\boldsymbol{H}_j-\boldsymbol{O}_j\| \le \|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2 + \|\sin\Theta(\boldsymbol{U}_j,\boldsymbol{U}^*_j)\|^2.\end{aligned}$$

This together with (2.16) imply

$$\begin{aligned}&\sqrt{n}\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_\infty\\ \le&\sqrt{n}\|\boldsymbol{U}_j\boldsymbol{O}_j-\boldsymbol{U}^*_j\|_{2\to\infty} + \left(\sqrt{2\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2}+\sqrt{2}\|\sin\Theta(\boldsymbol{U}_j,\boldsymbol{U}^*_j)\|\right)\left(\sqrt{n}\|\boldsymbol{U}^*_j\|_{2\to\infty}\right)\\ &+\sqrt{\xi}\left(\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2+\|\sin\Theta(\boldsymbol{U}_j,\boldsymbol{U}^*_j)\|^2\right).\end{aligned} \tag{2.17}$$

Let $\tilde{K}$ denote the number of eigenvalues that are strictly smaller than $n\underline{p}_*$. By Theorem 2.2.1, $\tilde{K}\le K$. Since $\xi=O(1)$, $K=O(1)$. Then for any $j\le\tilde{K}$, by part (4) of Theorem 2.2.1,

$$\|\boldsymbol{U}^*_j\|_{2\to\infty} \lesssim \frac{1}{\sqrt{n}}.$$

Thus, to prove $\sqrt{n}\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_\infty < \min\{\sqrt{n_0/n_1},\sqrt{n_1/n_0}\}$, it remains to prove

$$\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2\le c,\quad \|\sin\Theta(\boldsymbol{U}_j,\boldsymbol{U}^*_j)\|\le c,\quad \sqrt{n}\|\boldsymbol{U}_j\boldsymbol{O}_j-\boldsymbol{U}^*_j\|_{2\to\infty}\le c, \tag{2.18}$$

for a sufficiently small constant c that only depends on ξ for some $2\le j\le\tilde{K}$ with high probability. The first bound $\|\boldsymbol{u}_{n-1}-\boldsymbol{u}^*_{n-1}\|_2\le c$ has been proved in Theorem 2.3.2 if $C_{\ell_\infty}\ge C_{\ell_2}$. We will show the other two bounds in the following subsections.

2.5.5.1. *Choice of j via the pigeonhole principle.* By definition of $\tilde{K}$, $np_\emptyset=\lambda^*_{n-1}\le\cdots\lambda^*_{n-\tilde{K}+1}< n\underline{p}_*\le\lambda^*_{n-\tilde{K}}$. Let

$$\delta^*_j=\min\{n\underline{p}_*,\lambda^*_{n-j-1}\}-\lambda^*_{n-j},\quad \tilde{j}=\mathrm{argmax}_{j\le\tilde{K}-1}\delta^*_j. \tag{2.19}$$

By Theorem 2.2.1, $\tilde{K}\le K$ and thus

$$\delta^*_{\tilde{j}}\ge\frac{1}{\tilde{K}-1}\sum_{j=1}^{\tilde{K}-1}\delta^*_j=\frac{n(\underline{p}_*-p_\emptyset)}{\tilde{K}}. \tag{2.20}$$

Throughout the rest of the proof we will fix $j = \tilde{j}$ and depress the subscript j when no confusion can arise. This option guarantees a sufficiently large eigengap so that the off-the-shelf technical tools can be applied directly to obtain meaningful perturbation bounds.

For notational convenience, denote by $\mathbf{\Lambda}$ the diagonal matrix of the $\tilde{K}$ smallest eigenvalues, i.e.,

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_{n-1}, \ldots, \lambda_{n-\tilde{j}}), \quad \mathbf{\Lambda}^* = \mathrm{diag}(\lambda^*_{n-1}, \ldots, \lambda^*_{n-\tilde{j}}).$$

We write $\boldsymbol{U}, \boldsymbol{O}$ and $\boldsymbol{U}^*$ for $\boldsymbol{U}_{\tilde{j}}, \boldsymbol{O}_{\tilde{j}}$ and $\boldsymbol{U}^*_{\tilde{j}}$ throughout the rest of the section. Similar to the proof of Theorem 2.3.2, let

$$\boldsymbol{L}_\nu = \boldsymbol{L} + \nu \boldsymbol{J}, \quad \boldsymbol{L}^*_\nu = \boldsymbol{L}^* + \nu \boldsymbol{J}, \quad \text{where } \boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T. \tag{2.21}$$

Since $\mathbf{1}_n^T \boldsymbol{U} = \mathbf{1}_n^T \boldsymbol{U}^* = 0$,

$$\boldsymbol{L}_\nu \boldsymbol{U} = \boldsymbol{U}(\mathbf{\Lambda} + \nu \boldsymbol{I}), \quad \boldsymbol{L}^*_\nu \boldsymbol{U}^* = \boldsymbol{U}^*(\mathbf{\Lambda}^* + \nu \boldsymbol{I}).$$

Throughout we take

$$\nu = n\bar{p}^*. \tag{2.22}$$

2.5.5.2. *Bounding* $\|\sin\Theta(U, U^*)\|$. Applying Lemma 2.5.1, we have

$$\|\boldsymbol{L} - \boldsymbol{L}^*\| \le C(r)\left(\sqrt{n\bar{p}^* \log n} + \log n\right) \quad \text{with probability } 1 - n^{-r} \tag{2.23}$$

Since $\delta^*_{\tilde{j}}$, the eigengap defined in (2.20), is invariant to ν, by Davis-Kahan Theorem (Lemma 2.5.3) and (2.22),

$$\|\sin\Theta(\boldsymbol{U}, \boldsymbol{U}^*)\| \le \frac{2\|\boldsymbol{L}_\nu - \boldsymbol{L}^*_\nu\|}{\min\left\{\delta^*_{\tilde{j}}, np_\emptyset + \nu\right\}} = \frac{2\|\boldsymbol{L} - \boldsymbol{L}^*\|}{\delta^*_{\tilde{j}}} \le 2KC(r)\frac{\left(\sqrt{n\bar{p}^* \log n} + \log n\right)}{n(\underline{p}_* - p_\emptyset)}$$

By the condition (2.6),

$$(n\bar{p}^*)^4 \ge (n(\underline{p}_* - p_\emptyset))^4 \ge C_{\ell_\infty}(n\bar{p}^*)^3 \log n \Longrightarrow n\bar{p}^* \ge C_{\ell_\infty} \log n. \tag{2.24}$$

Thus, with probability $1 - n^{-r}$,

$$\|\sin\Theta(\boldsymbol{U}, \boldsymbol{U}^*)\| \leq 2KC(r)\frac{\left(\sqrt{n\bar{p}^* \log n} + \log n\right)}{C_{\ell_\infty}^{1/4}(n\bar{p}^*)^{3/4}(\log n)^{1/4}}$$
$$\leq 2KC(r)\frac{1 + C_{\ell_\infty}^{-1/2}}{C_{\ell_\infty^{1/2}}} = 2KC(r)\left(C_{\ell_\infty}^{-1/2} + C_{\ell_\infty}^{-1}\right).$$

Therefore, when $C_{\ell_\infty} \geq \max\{1, 4KC(r)/c\}$,

$$\|\sin\Theta(\boldsymbol{U}, \boldsymbol{U}^*)\| \leq c,$$

with probability $1 - n^{-r}$.

2.5.5.3. *Bounding* $\sqrt{n}\|UO - U^*\|_{2\to\infty}$. We will apply Theorem 2.6 of Lei (2019) on $\boldsymbol{L}_\nu$ and $\boldsymbol{L}_\nu^*$, defined in (2.21) with ν defined in (2.22). To be self-contained, we state the theorem in Appendix A.3 together with all necessary definitions. We can prove the following result.

Lemma 2.5.5. *Let* $\boldsymbol{U}, \boldsymbol{O}, \boldsymbol{U}^*$ *be defined as in Section 2.5.5.1. Fix any constant* $r, c > 0$. *Under the same setting as in Theorem 2.3.1, if* C_{ℓ_∞} *is a sufficiently large constant that only depends on* r *and* c, *with probability at least* $1 - (10K + 1)n^{-r}$,

$$\sqrt{n}\|\boldsymbol{UO} - \boldsymbol{U}^*\|_{2\to\infty} \leq c.$$

The proof is lengthy and hence relegated to Appendix A.3.

2.6. Conclusion and Discussion

In this chapter, we present a novel analysis of an unnormalized graph Laplacian-based recursive spectral clustering algorithm for sparse networks. Under a broad class of hierarchical network models, we show that the proposed algorithm is effective in both community detection and hierarchy estimation. Both weak and strong consistencies for mega-communities are established based on novel ℓ_2 and ℓ_∞ perturbation bounds of the Fiedler vector. Compared to earlier works on hierarchical and non-hierarchical community detection, our result substantially relaxes the constraints on connection probabilities, degree heterogeneity, and the hierarchical structure to handle sparse and multi-scale networks with an unbalanced hierarchy.

One limitation of our model is that the hierarchy is restricted to be a binary tree. Theoretically, a binary tree cannot encode, for example, a 3-block SBM with equal between-community connection probabilities due to the indistinguishability of the three primitive communities. However, the strict homogeneity as above is arguably a rare corner case in practice. When the between-community connection probabilities are mutually distinct, there exists a meaningful binary hierarchy in this case. Indeed, if $B_{12} = \max\{B_{12}, B_{13}, B_{23}\}$, the two mega-communities on the first level of the binary tree can be defined as $\{1, 2\}$ and $\{3\}$. We can verify that all between-mega-community connection probabilities are smaller than all within-mega-community connection probabilities in this case. Although it is beyond a general BTSBM, some of our results can be possibly extended to this case; see Remark 5 for instance. It would be interesting to investigate what hierarchical structures can be equivalently formulated as a binary one.

Our main results can also be interpreted as the adaptivity of graph Laplacian based spectral clustering to inhomogeneous connection probabilities. The general BTSBM studied in this work is a special case of inhomogeneous block models in which within-community connection probabilities are greater than p while between-community connection probabilities are smaller than $q < p$. Note that it is known that convex optimization approach would be adaptive to such inhomogeneous model (Moitra et al., 2016), which is usually computationally expensive and sometimes involves sensitive tuning parameters. It would be interesting to see whether certain spectral method also has such adaptivity. In particular, we are interested in studying such adaptivity under either multi-scale regime or the very sparse regime where the connection probabilities are on the order of $\omega(1/n)$.

Another important question is how degree heterogeneity degrades the performance of a clustering algorithm. Degree heterogeneity is ubiquitous in real-world networks while most theoretical works restrict the degrees to be on the same order. This work takes one step in mitigating the gap for a specific spectral clustering algorithm. As discussed in Remark 3, the degree variation condition (2.6) in Theorem 2.3.1 appears to be a mathematical artifact, but to relax it requires novel techniques beyond our proof strategies summarized in Section 2.3.4. We leave this as an open problem and look for affirmation or negation of our conjecture.

In terms of the technical proofs, as alluded to in Section 2.4, the ℓ_∞ perturbation bound is only sufficient yet not necessary for the strong consistency of mega-communities recovery. Previous

works (e.g. Abbe et al., 2017; Deng et al., 2020; Lei, 2019) suggest that the sign consistency is strictly weaker than ℓ_∞ consistency for eigenvectors under certain special SBMs or BTSBMs. It is mathematically intriguing to explore how those advanced techniques can be adapted to the more heterogeneous BTSBMs.

CHAPTER 3

Learning Linear Polytree Structural Equation Models

3.1. Introduction

Over the past three decades, the problem of learning directed graphical models from i.i.d. observations of a multivariate distribution has received enormous amount of attention since they provide a compact and flexible way to represent the joint distribution of the data, especially when the associated graph is a directed acyclic graph (DAG), which is a directed graph with no directed cycles. DAG models are popular in practice with applications in biology, genetics, machine learning and causal inference (Koller and Friedman, 2009; Sachs et al., 2005; Spirtes et al., 2000; Zhang et al., 2013). There exists an extensive literature on learning the graph structure from i.i.d. observations under DAG models. For a summary, see the survey papers Drton and Maathuis (2017); Heinze-Deml et al. (2018). Existing approaches generally fall into two categories, constraint-based methods (Pearl, 2009; Spirtes et al., 2000) and score-based methods (Chickering, 2002b). Constraint-based methods utilize conditional independence test to determine whether there exists an edge between two nodes and then orient the edges in the graph, such that the resulting graph is compatible with the conditional independencies determined in the data. Score-based methods formulate the structure learning task as optimizing a score function based on the unknown graph and the data.

A polytree is a connected DAG which contains no cycles even if the directions of all edges are ignored. It has been popularly used in practice due to tractability in both structure learning and inference. To the best of our knowledge, structure learning of polytree models was originally studied in Rebane and Pearl (1987), in which the skeleton of the polytree is estimated by applying the Chow-Liu algorithm (Chow and Liu, 1968) to pairwise mutual information quantities, a method that has been widely used in the literature of Markov random field to fit undirected tree models. Polytree graphical models have received a significant amount of research interests both empirically and theoretically ever since, see, e.g., Dasgupta (1999); Huete and de Campos (1993); Ouerd et al.

(2004). Skeleton recovery via Chow-Liu algorithm has also been used as an initial step for fitting more general sparse DAGs; see, e.g., Cheng et al. (2002).

This chapter aims to study sample size conditions of the method essentially proposed in Rebane and Pearl (1987) for the recovery of polytree structures, but we apply the Chow-Liu algorithm to pairwise sample correlations rather than estimated mutual information quantities. In particular, by restricting our study to the case of Gaussian linear structure equation models (SEM), we will establish sufficient conditions on the sample sizes for consistent recovery of both the skeleton and equivalence class for the underlying polytree structure. On the other hand, we will also establish necessary conditions on the sample sizes for these two tasks through information-theoretic lower bounds. Our sufficient and necessary conditions match in order in a broad regime of model parameters, and thereby characterize the difficulty of these two tasks in polytree learning. In addition, we extend the results to the sub-Gaussian case, and establish an upper bound for the estimation error of the inverse correlation matrix under the same models.

An important line of research that inspires our study is structure learning for tree-structured undirected graphical models, including both discrete cases (Anandkumar et al., 2012a,b; Bresler and Karzand, 2020; Heinemann and Globerson, 2014; Netrapalli et al., 2010) and Gaussian cases (Katiyar et al., 2019; Nikolakakis et al., 2019; Tan et al., 2010; Tavassolipour et al., 2018). In particular, conditions on the sample size for undirected tree structure learning via the Chow-Liu algorithm have been studied for both Ising and Gaussian models (Bresler and Karzand, 2020; Nikolakakis et al., 2019; Tavassolipour et al., 2018), and the analyses usually rely crucially on the so-called "correlation decay" property over the true undirected tree. The correlation decay properties can usually be explicitly quantified by the pairwise population correlations corresponding to the edges of the underlying true tree. Based on this result and some perturbation results of pairwise sample correlations to their population counterparts, sufficient conditions on the sample size for undirected tree recovery with the Chow-Liu algorithm can be straightforwardly obtained.

In order to apply the above technical framework to study the sample size conditions for polytree learning, a natural question is whether we have similar correlation decay phenomenon for the polytree models. In fact, this is suggested in the seminal paper Rebane and Pearl (1987). To be concrete, under some non-degeneracy assumptions, it has been shown in Rebane and Pearl (1987) that

there holds the "mutual information decay" over the skeleton of the underlying polytree. Roughly speaking, the mutual information decay is a direct implication of the well-known "data processing inequality" in information theory (Thomas and Joy, 2006). Restricted to the very special case of Gaussian linear SEM, the mutual information decay is indeed equivalent to the property of population correlation decay.

However, to obtain some meaningful sample complexity result, we need to quantify such correlation decay explicitly as what has been done in the study of Chow-Liu algorithm for undirected tree models (Bresler and Karzand, 2020; Nikolakakis et al., 2019; Tavassolipour et al., 2018). The mutual information decay given in Rebane and Pearl (1987) holds for general polytree models, but one can expect to further quantify such decay under more specific models. In fact, if we restrict the polytree model to linear SEM, by applying the well-known Wright's formula (Foygel et al., 2012; Nowzohour et al., 2017; Wright, 1960), the population correlation decay property can be quantified by the pairwise correlations corresponding to the tree edges. Note that such quantification of correlation decay holds even for non-Gaussian linear polytree SEM. This is interesting since in general mutual information decay does not directly imply population correlation coefficients decay for non-Gaussian models. With such quantification of correlation decay over the underlying polytree skeleton, we can apply the ideas from undirected tree structure learning to establish the sufficient conditions on sample size for polytree skeleton recovery via the Chow-Liu algorithm. Roughly speaking, if the maximum absolute correlation coefficient over the polytree skeleton is strictly bounded below 1, Chow-Liu algorithm recovers the skeleton exactly with probability at least $1-\delta$ when the number of samples satisfies $n > O(\frac{1}{\rho_{\min}^2} \log \frac{p}{\sqrt{\delta}})$, where p is the number of variables and $\rho_{\min}$ is the minimum absolute population correlation coefficient over the skeleton.

To determine the directions of the polytree over the skeleton, the concept of CPDAG (Verma and Pearl, 1991) captures the equivalence class of polytrees. We then consider the CPDAG recovery procedure introduced in Verma and Pearl (1992) and Meek (1995), which is a polynomial time algorithm based on identifying all the v-structures (Verma and Pearl, 1991). Therefore, conditional on the exact recovery of the skeleton, recovering the CPDAG is equivalent to recovering all v-structures. In a non-degenerate polytree model, a pair of adjacent edges form a v-structure if and only if the two non-adjacent node variables in this triplet are independent, so we consider a

natural v-structure identification procedure by thresholding the pairwise sample correlations over all adjacent pairs of edges with some appropriate threshold. In analogy to the result of skeleton recovery, we show that the CPDAG of the polytree can be exactly recovered with probability at least $1-\delta$ if the sample size satisfies $n > O(\frac{1}{\rho_{\min}^4} \log \frac{p}{\sqrt{\delta}})$.

Our sufficient condition on sample size for skeleton recovery is proportional to $1/\rho_{\min}^2$, whereas that for CPDAG recovery is proportional to $1/\rho_{\min}^4$. One may ask whether this discrepancy correctly captures the difficulties for the two tasks, or it is just a mathematical artifact. By using the Fano's method, we show that $n > O(\frac{\log p}{\rho_{\min}^2})$ is necessary for skeleton recovery, while $n > O(\frac{\log p}{\rho_{\min}^4})$ is necessary for CPDAG recovery. This means that we have sharply characterized the difficulties for the two tasks.

This chapter is organized as follows: In Section 3.2, we will review the concept of linear polytree SEM, the concepts of Markov equivalence and CPDAG, and the polytree learning method introduced in Rebane and Pearl (1987) as well as the CPDAG recovery method introduced in Verma and Pearl (1992) and Meek (1995). In Section 3.3, we will first explain the phenomenon of correlation decay under linear polytree models, and then focus on the Gaussian polytree model and establish sufficient conditions on sample sizes for both skeleton and CPDAG recovery. In addition, we will also establish information-theoretic bounds as necessary conditions on sample sizes for these two tasks. In Section 3.4, we extend the sample size conditions to sub-Gaussian linear polytree models, and then give an upper bound in the entry-wise ℓ_1 norm for the estimation of the inverse correlation matrix. All proofs are deferred to Section 3.6. Our theoretical findings as well as empirical robustness of polytree learning will be illustrated by numerical experiments in Section 3.5. A brief summary of our work and some potential future research will be discussed in Section 3.7.

3.2. Linear Polytree Models and Learning

The aim of this section is to give an overview of the concepts of DAG models, linear polytree models, equivalence classes characterized by CPDAG, and Chow-Liu algorithm for polytree learning. Most materials are not new, but we give a self-contained introduction of these important concepts and methods so that our main results introduced in the subsequent sections will be more accessible to a wider audience.

3.2.1. Linear Polytree Models. Let $G = (V, E)$ be a directed graph with vertex set $V = \{1, 2, ..., p\}$ and edge set E. We use $i \to j \in E$ to denote that there is a directed edge from node i to node j in G. A directed graph with no directed cycles is referred to as a directed acyclic graph (DAG). The parent set of node j in G is denoted as $Pa(j) \coloneqq \{i \in V : i \to j \in E\}$. Correspondingly, denote by $Ch(j) \coloneqq \{k : j \to k \in E\}$ the children set of j.

Let $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ be a random vector where each random variable X_j corresponds to a node $j \in V$. The edge set E usually encodes the causal relationships among the variables. The random vector $\boldsymbol{X}$ is said to be Markov on a DAG G if its joint density function (or mass function) $p(\boldsymbol{x})$ can be factorized according to G as $p(\boldsymbol{x}) = \prod_{j=1}^{p} p(x_j | x_{Pa(j)})$, where $p(x_j | x_{Pa(j)})$ is the conditional density/probability of X_j given its parents $X_{Pa(j)} \coloneqq \{X_i : i \in Pa(j)\}$. We usually refer to $(G, p(\boldsymbol{x}))$ as a DAG model.

For any DAG, if we ignore the directions of all its directed edges, the resulting undirected graph is referred to as the *skeleton* of the DAG. A polytree is a connected DAG whose skeleton does not possess any undirected cycles. A polytree model is a multivariate probability distribution $p(\boldsymbol{x})$ that is Markov to a polytree $T = (V, E)$. As mentioned in Introduction, polytree models are an important and tractable class of directed graphical models, largely because they permit fast exact inference.

Throughout this work, we restrict our discussion to an important sub-class of DAG models: linear structure equation models (SEM), in which the dependence of each X_j on its parents is linear with an additive noise. The parameterization of the linear SEM with directed graph $G = (V, E)$ would be $X_j = \sum_{i=1}^{p} \beta_{ij} X_i + \epsilon_j = \sum_{i \in Pa(j)} \beta_{ij} X_i + \epsilon_j$ for $j = 1, \ldots, p$, where $\beta_{ij} \neq 0$ *if and only if* $i \to j \in E$, and all ϵ_j's are independent with mean zero, usually with different variances. Let $\boldsymbol{B} = \left[\beta_{ij}\right] \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_p]^\top$. Then the SEM can be represented as

$$\boldsymbol{X} = \boldsymbol{B}^\top \boldsymbol{X} + \boldsymbol{\epsilon}. \tag{3.1}$$

Denote $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \left[\sigma_{ij}\right] \in \mathbb{R}^{p \times p}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Omega} = \mathrm{Diag}(\omega_{11}, \ldots, \omega_{pp})$. Here $\boldsymbol{\Omega}$ is diagonal since all additional noises are assumed to be mutually independent. Note that when we say that a linear SEM is associated to a polytree $T = (V, E)$, this is in general stronger than Markov property, since we can determine the directed edges in T from the non-zero patterns of $\boldsymbol{B}$. In addition,

if the noises ϵ_i's are Gaussian, then the linear polytree model is referred to as a *Gaussian linear polytree model*. Similarly, if ϵ_j's are sub-Gaussian, then the linear polytree model is referred to as a *sub-Gaussian linear polytree model*.

3.2.2. Markov Equivalence and CPDAG. Let's briefly review the concept of Markov equivalence of DAGs. In fact, there are several equivalent definitions for this concept. The most intuitive definition is perhaps the following: if any multivariate distribution $p(\boldsymbol{x})$ that is Markov to G_1 is Markov to G_2, and vice versa, then we say DAGs G_1 and G_2 are Markov equivalent. Characterization of Markov equivalence between DAGs through multivariate Gaussian distributions is given in Ghassami et al. (2020). Another intuitive definition is from the concept of conditional independence. Note that each DAG G entails a list of statements of conditional independence, which are satisfied by any joint distribution Markov to G. Then two DAGs are equivalent if they entail the same list of conditional independencies. In the present paper, the recovery of equivalence class of DAG hinges on the following famous and neat result given in Verma and Pearl (1991): Two DAGs are Markov equivalent if and only if they have the same skeleton and sets of v-structures, where a v-structure is a node triplet $i \to k \leftarrow j$ where i and j are non-adjacent.

An important concept to intuitively capture equivalence classes of DAGs is the completed partially DAG (CPDAG): a graph K with both directed and undirected edges representing the Markov equivalence class of a DAG G if: (1) K and G have the same skeleton; (2) K contains a directed edge $i \to j$ if and only if any DAG G' that is Markov equivalent to G contains the same directed edge $i \to j$. The CPDAG of G is denoted as $K = C_G$. It has been shown in Chickering (2002a) that two DAGs have the same CPDAG if and only if they belong to the same Markov equivalence class.

It would be interesting to have some intuitions on what the CPDAG of a polytree looks like. To this end, we introduce the following result, the proof of which can be found in Section 3.6.2.

THEOREM 3.2.1. The undirected sub-graph containing undirected edges of the CPDAG of a polytree forms a forest. All equivalent DAGs can be obtained by orienting each undirected tree of the forest into a rooted tree, that is, by selecting any node as the root and setting all edges going away from it.

3.2.3. Polytree Learning. A major purpose of this chapter is to study the problem of polytree learning, i.e., the recovery of the CPDAG of the polytree $T = (V, E)$ under the linear SEM (3.1) from a finite sample of observations. To be concrete, suppose that we have observed i.i.d. samples $\boldsymbol{X}^{(1:n)} = [\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}]^\top \in \mathbb{R}^{n \times p}$ generated from the linear SEM (3.1) that is associated to a polytree $T = (V, E)$. We aim to consistently recover the CPDAG of T, namely C_T, from $\boldsymbol{X}^{(1:n)}$.

The procedure of polytree learning we are considering in this chapter has been in principle introduced in Rebane and Pearl (1987), but we use the sample correlation coefficients rather than estimated mutual information quantities. For multivariate Gaussian distributions, of course, the Chow-Liu algorithm applying to empirical mutual information quantities is the same as the one applying to pairwise sample correlations. The key idea is to first recover the skeleton of the polytree by applying the Chow-Liu algorithm (Chow and Liu, 1968) to the pairwise sample correlations of the data matrix. After the skeleton is recovered, the set of all v-structures can be correctly identified via a simple thresholding approach to pairwise sample correlations. Finally, the CPDAG can be found by applying Rule 1 introduced in Verma and Pearl (1992), as guaranteed theoretically in Meek (1995).

3.2.3.1. *Chow-Liu Algorithm for Skeleton Recovery.* To the best of our knowledge, it was first proposed in Rebane and Pearl (1987) to recover the skeleton of a polytree by applying the Chow-Liu algorithm introduced in Chow and Liu (1968) that was originally intended for undirected tree graphical models. Notice that given we are interested in linear polytree models, we directly apply the Chow-Liu algorithm to the sample correlations.

The Chow-Liu tree associated to pairwise correlations, which is the estimated skeleton of the underlying polytree, is defined as below.

Definition 3.2.2 (Chow-Liu tree associated to pairwise sample correlations)**.** Consider the linear polytree model (3.1) associated to a polytree $T = (V, E)$, whose skeleton is denoted as $\mathcal{T} = (V, \mathcal{E})$. Let $\mathbb{T}_p$ denote the set of undirected trees over p nodes. Given n i.i.d. samples $\boldsymbol{X}^{(1:n)} = [\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}]^\top \in \mathbb{R}^{n \times p}$, we obtain the sample correlation $\hat{\rho}_{ij}$ between X_i and X_j for all $1 \leq i < j \leq p$. The Chow-Liu tree associated to the pairwise sample correlations is defined as the maximum-weight spanning tree over the p nodes where the weights are absolute values of sample

correlations:

$$\mathcal{T}^{CL} = \text{argmax}_{\mathcal{T}=(V,\mathcal{E})\in\mathbb{T}_p} \sum_{i-j\in\mathcal{E}} |\hat{\rho}_{ij}|. \tag{3.2}$$

For tree-structured undirected graphical models, it has been established in Chow and Liu (1968) that the maximum likelihood estimation of the underlying tree structure is the Chow-Liu tree associated to the empirical mutual information quantities (which are used to find the maximum-weight spanning tree). The rationale of applying Chow-Liu algorithm to polytree learning has been carefully explained in Rebane and Pearl (1987), to which interested readers are referred. The step of skeleton recovery can be summarized in Algorithm 1.

Algorithm 1 Chow-Liu algorithm

Input: n i.i.d. samples $\boldsymbol{X}^{(1:n)}$
Output: Estimated skeleton $\widehat{\mathcal{T}}$.

1: Compute the pairwise sample correlations $\hat{\rho}_{ij}$ for all $1 \le i < j \le p$;
2: Construct a maximum-weight spanning tree using $|\hat{\rho}_{ij}|$ as the edge weights, i.e., $\widehat{\mathcal{T}} = \mathcal{T}^{CL}$ defined in (3.2).

It is noteworthy that Algorithm 1 can be implemented efficiently by applying the Kruskal's algorithm (Kruskal, 1956) to pairwise sample correlations $|\hat{\rho}_{ij}|$ for the construction of maximum weight spanning tree. The computational complexity for Kruskal's algorithm is known to be $O(p^2 \log p)$, which is generally no larger than that for computing the sample correlations, which is $O(p^2 n)$.

3.2.3.2. *CPDAG Recovery.* In the second part of the procedure of polytree learning, we aim to extend the estimated skeleton $\widehat{\mathcal{T}} = \mathcal{T}^{CL}$ to an estimated CPDAG of the underlying polytree T. Intuitively speaking, this amounts to figuring out all the edges whose orientations can be determined.

The first step of this part is to identify all the v-structures. Recall that in the linear polytree model (3.1), we assume that $\beta_{ij} \neq 0$ if and only if $i \to j \in E$, which implies the non-degeneracy of the polytree. In this case, it has also been observed in Rebane and Pearl (1987) that, for any pair of non-adjacent nodes i and j with common neighbor k, they form a v-structure $i \to k \leftarrow j$ if and only X_i and X_j are mutually independent. Interestingly, the criterion of mutual independence can be replaced with zero correlation under the linear polytree model, even under non-Gaussian models. Then, we can determine the existence of a v-structure $i \to k \leftarrow j$ when the sample correlation

$|\hat{\rho}_{ij}| < \rho_{crit}$. The discussion on the practical choice of ρ_{crit} is deferred to Section 3.5. Theoretical discussions on the threshold will be elaborated in Sections 3.3 and 3.4.

After recovering all the v-structures, as aforementioned, it is guaranteed in Meek (1995) that the CPDAG of the polytree can be recovered by iteratively applying the four rules originally introduced in Verma and Pearl (1992). However, given our discussion is restricted to the polytree models, Rules 2, 3, and 4 in Verma and Pearl (1992) and Meek (1995) do not apply. We only need to apply Rule 1 repeatedly. This rule can be stated as follows: Orient any undirected edge $j - k$ into $j \to k$ whenever there is a directed edge $i \to j$ coming from a third node i.

These two steps in the second part of polytree structure learning are summarized as Algorithm 2.

Algorithm 2 Extending the skeleton to a CPDAG

Input: Estimated skeleton $\hat{S}$, sample correlations $\hat{\rho}_{ij}$'s, critical value for correlation ρ_{crit}.

Output: Estimated CPDAG $\hat{C}$.

1: **for** Each pair of non-adjacent variables i, j with common neighbor k **do**
2: **if** $|\hat{\rho}_{ij}| < \rho_{crit}$ **then**
3: replace $i - k - j$ in $\hat{S}$ by $i \to k \leftarrow j$
4: In the resulting graph, orient as many undirected edges as possible by repeatedly applying the rule: orient an undirected edge $j - k$ into $j \to k$ whenever there is a directed edge $i \to j$ for some i.

3.3. Main Results for Gaussian Polytree Models

In this section, we discuss sample size conditions for the recovery of skeleton and CPDAG under a Gaussian linear polytree model $T = (V, E)$, and the sub-Gaussian case will be discussed in the next section. We first establish a correlation decay property on the polytree skeleton by applying the famous Wright's formula.

3.3.1. Wright's Formula and Correlation Decay on Polytree Skeleton. First, the polytree learning method introduced in Section 3.2.3 depends solely on the marginal correlation coefficients, and is thereby invariant to scaling. Therefore, without loss of generality, we can assume that X_j's have unit variance for all $j \in V$, i.e. $\mathbf{\Sigma}$ is the correlation matrix. It is obvious that the standardized version of a linear SEM is still a linear SEM. In fact, let $\boldsymbol{D}$ be the diagonal matrix

with the j-th diagonal being the standard deviation of X_j. Then the standardized random variables are $\boldsymbol{D}^{-1}\boldsymbol{X}$, which satisfies

$$\boldsymbol{D}^{-1}\boldsymbol{X} = (\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^{-1})^\top \boldsymbol{D}^{-1}\boldsymbol{X} + \boldsymbol{D}^{-1}\boldsymbol{\epsilon}. \tag{3.3}$$

In other words, the standardized random variables admit a linear SEM with the coefficient matrix $\tilde{\boldsymbol{B}} = \boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^{-1}$ and the diagonal noise variance matrix $\tilde{\boldsymbol{\Omega}} = \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}^{-1} = \boldsymbol{D}^{-2}\boldsymbol{\Omega}$. One should note that $\tilde{\boldsymbol{B}}$ and $\boldsymbol{B}$ correspond to the same DAG. When the variables are all variance-one, denoting the pairwise correlations as $\rho_{ij} \coloneqq \mathrm{corr}(X_i, X_j)$, we have $\sigma_{ij} = \rho_{ij}$ for all $1 \le i, j \le p$.

Under the linear SEM (3.1), we know that $\boldsymbol{B}$ is permutationally similar to an upper triangular matrix, which implies that all eigenvalues of $\boldsymbol{I} - \boldsymbol{B}$ are 1's, and further implies that $\boldsymbol{I} - \boldsymbol{B}$ is invertible. Then, $(\boldsymbol{I} - \boldsymbol{B})^\top \boldsymbol{X} = \boldsymbol{\epsilon}$ implies $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\epsilon}$, and further implies that $\boldsymbol{X}$ is mean-zero, and has covariance $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\Omega}(\boldsymbol{I} - \boldsymbol{B})^{-1}$. This suggests that we can represent the entries of $\boldsymbol{\Sigma}$ by (β_{ij}) and (ω_{ii}). In fact, this can be conveniently achieved by using the Wright's path tracing formula (Wright, 1960). We first introduce some necessary definitions in order to obtain such expression. A *trek* connecting nodes i and j in a directed graph $G = (V, E)$ is a sequence of non-colliding consecutive edges connecting i and j of the form

$$i = v_l^L \leftarrow v_{l-1}^L \leftarrow \cdots \leftarrow v_1^L \leftarrow v_0 \rightarrow v_1^R \rightarrow \cdots \rightarrow v_{r-1}^R \rightarrow v_r^R = j.$$

We define the left-hand side of τ as $Left(\tau) = v_l^L \leftarrow \cdots \leftarrow v_0$, the right-hand side of τ as $Right(\tau) = v_0 \rightarrow \cdots \rightarrow v_r^R$, and the head of τ as $H_\tau = v_0$. A trek τ is said to be a *simple trek* if $Left(\tau)$ and $Right(\tau)$ do not have common edges.

Denoting the sets of simple treks $\mathcal{S}^{ij} = \{\tau : \tau \text{ is a simple trek connecting } i \text{ and } j \text{ in } G\}$, the following rules (Foygel et al., 2012; Nowzohour et al., 2017; Wright, 1960) express the off-diagonal entries of the covariance matrix $\boldsymbol{\Sigma}$ as a summation over treks:

$$\sigma_{ij} = \sum_{\tau \in \mathcal{S}^{ij}} \sigma_{H_\tau H_\tau} \prod_{s \to t \in \tau} \beta_{st} \text{ for } i \neq j.$$

Now let us simplify the above trek rules under the linear polytree models with standardized variables. In a polytree model, any two nodes i and j are connected by a unique path, which is

either a simple trek or a path with collision. As a direct consequence of the trek rules introduced above, we have the following result.

Lemma 3.3.1. Consider the linear polytree model (3.1) with the associated polytree $T = (V, E)$ over p nodes, i.e., $\beta_{ij} \neq 0$ if and only if $i \to j \in E$. Also assume that X_j has unit variance for all $j \in V$. Then, given $\{\beta_{ij}\}_{1 \leq i,j \leq p}$, we have the following results regarding the entries of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$:

(1) For any $i \neq j$,

$$\rho_{ij} = \sigma_{ij} = \begin{cases} \prod_{s \to t \in \tau_{ij}} \beta_{st} & \text{the path connecting } i \text{ and } j \text{ is a simple trek;} \\ 0 & \text{otherwise,} \end{cases} \tag{3.4}$$

where τ_{ij} is the simple trek connecting i and j when there is one.

(2) The diagonal entries of $\mathbf{\Omega}$ are given by

$$\omega_{jj} = 1 - \sum_{i \in Pa(j)} \beta_{ij}^2, \quad j = 1, \ldots, p. \tag{3.5}$$

REMARK 1. Here Eq. (3.5) can be derived by the following simple argument: Since T is a polytree, all variables in $Pa(j)$ are independent and are also independent with ϵ_j. Evaluating the variance on both sides of $X_j = \sum_{i \in Pa(j)} \beta_{ij} X_i + \epsilon_j$ leads to Eq. (3.5).

Note that $i \to j \in E$ (or $j \to i \in E$) is a simple trek itself. Then Lemma 3.3.1 implies that $\rho_{ij} = \sigma_{ij} = \beta_{ij}$(or β_{ji}). Then Lemma 3.3.1 implies the following corollary.

Corollary 3.3.2. Consider the linear polytree model (3.1) with the associated polytree $T = (V, E)$ over p nodes. The pairwise population correlation coefficients satisfy that

$$\rho_{ij} = \begin{cases} \prod_{s \to t \in \tau_{ij}} \rho_{st} & \text{the path connecting } i \text{ and } j \text{ is a simple trek} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

REMARK 2. We need to emphasize that in this corollary the assumption that the variables $X_1, \ldots, X_p$ have unit variances is unnecessary. One can simply use (3.3) to standardize the linear polytree and then obtain (3.6), which still holds for the original linear SEM since correlation coefficients are invariant under standardization.

3.3.2. Connection to Correlation Decay on Undirected Gaussian Tree Models. A noteworthy fact is that undirected tree models (Markov random fields) can be viewed as special cases of polytree DAGs. Suppose $\mathcal{T} = (V, \mathcal{E})$ is an undirected tree. An undirected tree model is a multivariate distribution satisfies

$$p(\boldsymbol{x}) = \prod_{i=1}^{p} p_i(x_i) \prod_{(i,j)\in\mathcal{E}} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)}.$$

If we choose any node in a tree as the root, then we can transform the undirected tree into a unique *rooted tree*, i.e., a directed tree in which each non-root node has a unique parent node. Without loss of generality, let's choose node 1 as the root, and let $T = (V, E)$ be the resulting rooted tree, which implies that $\mathcal{E}$ is the skeleton of E. Then we can rewrite the joint pdf/pmf as

$$p(\boldsymbol{x}) = p_1(x_1) \prod_{i=2}^{p} p_i(x_i) \frac{p_{iPa(i)}(x_i, x_{Pa(i)})}{p_i(x_i)p_{Pa(i)}(x_{Pa(i)})} = p_1(x_1) \prod_{i=2}^{p} \frac{p_{iPa(i)}(x_i, x_{Pa(i)})}{p_{Pa(i)}(x_{Pa(i)})} = \prod_{j=1}^{p} p(x_j|x_{Pa(j)}),$$

which is the polytree model according to the rooted tree $T = (V, E)$. Similarly, one can show that a undirected Gaussian tree model can be viewed as a Gaussian linear polytree model.

Since an undirected Gaussian tree model can be represented as a Gaussian linear polytree model according to a rooted tree, any two nodes are connected by a simple trek as there are no colliding edges. Then Eq. (3.6) becomes $\rho_{ij} = \prod_{s\to t\in\tau_{ij}} \rho_{st}$ for any $i \neq j$. This is exactly the correlation decay property used in the literature to study the sample complexity for undirected tree structure learning, see, e.g. Bresler and Karzand (2020); Nikolakakis et al. (2019); Tavassolipour et al. (2018). We aim to apply the similar techniques employed in these works to derive sufficient conditions on the sample size for polytree learning.

3.3.3. Skeleton and CPDAG Recovery for Gaussian Models. For the convenience of quantifying the correlation decay rates, we need the following definitions.

Definition 3.3.3. In a standardized linear polytree model (3.1), let $\rho_{\min}$ and $\rho_{\max}$ be the minimum and maximum absolute correlation over the tree skeleton, that is

$$\rho_{\min} \coloneqq \min_{i\to j\in E} |\rho_{ij}|, \quad \rho_{\max} \coloneqq \max_{i\to j\in E} |\rho_{ij}|.$$

It is noteworthy that in general we cannot assume that $\rho_{\min}$ is independent of n or p. In fact, the second part of Lemma 3.3.1 gives rise to the following relationship between the noise variance and the correlation coefficients with parents for each node: $\sum_{i \in Pa(j)} \rho_{ij}^2 < 1$, which further implies the following corollary.

Corollary 3.3.4. Let d_* represent the highest in-degree for a polytree. Then $\rho_{\min} < \frac{1}{\sqrt{d_*}}$.

In contrast, it is reasonable to assume $\rho_{\min}$ to be a positive constant independent of p under the undirected tree-structured Gaussian graphical model, since after transforming it to a rooted tree as in Section 3.2.1, the highest in-degree satisfies $d_* = 1$.

3.3.3.1. *Skeleton Recovery.* We now introduce a sufficient condition on the sample size for skeleton recovery under the Gaussian linear polytree model, in which the independent noise variables satisfy $\epsilon_j \sim \mathcal{N}(0, \omega_{jj})$ for $j = 1, \ldots, p$. Then by $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})^{-\top} \boldsymbol{\epsilon}$, we know that $\boldsymbol{X}$ is also multivariate Gaussian. This fact will help quantify the discrepancy between population and sample pairwise correlations as characterized in Lemma 3.6.1.

THEOREM 3.3.5. Consider a Gaussian linear SEM (3.1) associated to a polytree $T = (V, E)$, in which all variables have variance one. Also, assume $\rho_{\min} > 0$ and $\rho_{\max} < 1$. Denote by $\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)})$ the estimated skeleton from the Chow-Liu algorithm (3.2) and by $\mathcal{T}$ the true skeleton from the polytree T. For any $\delta \in (0, 1)$, we have $\mathbb{P}(\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) \neq \mathcal{T}) < \delta$, provided

$$n > \left(\frac{8}{\rho_{\min}^2 (1 - \rho_{\max}^2)} + \frac{1}{2} \right) \left(\log \frac{3p^2}{2\delta (1 - \rho_{\max})^{7/2}} + \log n \right) + 4. \tag{3.7}$$

Moreover, if we further assume $n < p^{10}$ and $\rho_{\max} < 0.95$, then a sufficient condition for the exact skeleton recovery with probability at least $1 - \delta$ is $n > C_0 \log(p/\sqrt{\delta})/\rho_{\min}^2$ for some absolute constant C_0.

REMARK 3. The assumption that all variables have variance one can be removed since both the algorithm and polytree structure are scaling invariant.

REMARK 4. With the assumptions $\rho_{\min} > 0$ and $\rho_{\max} < 1$, Eq. (3.6) implies strict correlation decay over the tree skeleton, i.e., the population correlation coefficient between any two non-adjacent variables X_i and X_j is strictly smaller than that between any two consecutive variables on the

unique path between X_i and X_j in terms of absolute value. Putting this property and Lemma 3.6.1 together, we can derive the above sufficient condition on sample size for skeleton recovery by standard techniques in the literature of undirected tree learning, e.g., Bresler and Karzand (2020); Nikolakakis et al. (2019); Tavassolipour et al. (2018). We will give a self-contained proof in this chapter. The crux for the proof is Lemma 3.6.2 that guarantees the exact recovery of tree skeleton by Chow-Liu algorithm provided the sample correlation decay over the tree.

REMARK 5. The above condition implies some dependence of the sample size on the maximum in-degree d_*. In fact, together with Corollary 3.3.4, the sample size condition is essentially $n \geq O(d_* \log(p/\sqrt{\delta}))$ if $\rho_{\min} \asymp 1/\sqrt{d_*}$.

REMARK 6. As discussed in Section 3.3.2, Gaussian undirected tree models are equivalent to Gaussian linear rooted polytree models, so it is permissible to assume that $\rho_{\min} = \Omega(1)$ and $1 - \rho_{\max} = \Omega(1)$. In this case, the sufficient condition on sample size for skeleton recovery in Theorem 3.3.5 is $n = O(\log(p/\sqrt{\delta}))$. Note that Gaussian undirected tree structure learning based on the Chow-Liu algorithm has been studied in a recent work (Nikolakakis et al., 2019), in which the sufficient condition on the sample size is $n = O(\log^4(p/\delta))$.

3.3.3.2. *CPDAG Recovery.* As described in Section 3.2.3.2, after obtaining the estimated skeleton, the next step is to identify all v-structures by comparing ρ_{ij} for all node triplets $i - k - j$ in the skeleton with a threshold ρ_{crit}. Then the orientation propagation rule described in Algorithm 2 can be applied iteratively to orient as many undirected edges as possible. If both the skeleton and v-structures are correctly identified, the orientation rule will be able to recover the true CPDAG, i.e. equivalence class (Meek, 1995).

THEOREM 3.3.6. Consider a Gaussian linear SEM (3.1) associated to a polytree $T = (V, E)$, in which all variables have variance one. Also, assume $\rho_{\min} > 0$ and $\rho_{\max} < 1$. Denote by $\hat{C}(\boldsymbol{X}^{(1:n)})$ the estimated CPDAG from the entire algorithm in Sections 3.2.3.1 and 3.2.3.2 with threshold ρ_{crit}, and by C_T the true CPDAG from the polytree T. Denote $\gamma = \min\left\{\frac{\rho_{\min}}{3}, \frac{1-\rho_{\max}}{2}\right\}\rho_{\min}$. For any $\delta \in (0, 1)$, on an event with probability at least $1 - \delta$, we have $|\hat{\rho}_{ij} - \rho_{ij}| < \gamma$ for any $i < j$ and

$\hat{C}(\boldsymbol{X}^{(1:n)}) = C_T$, provided $\gamma < \rho_{crit} < \rho_{\min}^2 - \gamma$ and

$$n > \left(\frac{2}{\gamma^2} + \frac{1}{2}\right)\left(\log \frac{3p^2}{2\delta(1-\rho_{\max})^{7/2}} + \log n\right) + 4. \tag{3.8}$$

Moreover, if we further assume $n < p^{10}$ and $\rho_{\max} < 0.95$, then a sufficient condition for the exact CPDAG recovery with probability at least $1-\delta$ is $n > C_0 \log(p/\sqrt{\delta})/\rho_{\min}^4$ for some absolute constant C_0.

REMARK 7. Again, the assumption that all variables have unit variance can be removed due to the scaling invariance of the algorithm and the polytree structure.

REMARK 8. It is noteworthy to observe the difference between the sample size conditions in Theorems 3.3.5 and 3.3.6. In particular, if $\rho_{\min} \asymp 1/\sqrt{d_*}$, the above sufficient condition on sample size for CPDAG recovery is essentially $n \geq O(d_*^2 \log(p/\sqrt{\delta}))$, while recall that the sample size condition for skeleton recovery is $n \geq O(d_* \log(p/\sqrt{\delta}))$.

REMARK 9. In spite of the above implicit dependence on the maximum in-degree, the sufficient condition on the sample size does not rely on the maximum out-degree except for $n \gtrsim \log p$. Note that most existing theory on general sparse DAG recovery usually requires the sample size to be greater than the maximum neighborhood size, e.g., Theorem 2 of Kalisch and Bühlman (2007). Our result demonstrates the benefit by exploiting the polytree structures.

3.3.4. Information-theoretic Lower Bounds on the Sample Size. In this subsection, we will establish necessary conditions on the sample size for both skeleton and CPDAG recovery under Gaussian linear polytree models. In particular, we will use Fano's method to derive information-theoretic bounds.

THEOREM 3.3.7. Let $\mathcal{C}(\rho_{\min})$ be a collection of Gaussian linear polytree models, such that $\rho_{\min} := \min_{i\to j\in E} |\rho_{ij}|$ is fixed and satisfies $0 < \rho_{\min} < 1/\sqrt{p}$. In each model out of this class, assume that $\rho_{\max} := \max_{i\to j\in E} |\rho_{ij}| < 1/2$. Assume $p \geq 10$. Suppose that $\mathcal{C}(\rho_{\min})$ is indexed by θ, with corresponding polytree T_θ, covariance matrix $\boldsymbol{\Sigma}_\theta$, tree skeleton $\mathcal{T}_\theta$, and CPDAG C_{T_θ}. Then

for any skeleton estimator $\widehat{\mathcal{T}}$, there holds

$$\sup_{\theta \in \mathcal{C}(\rho_{\min})} \mathbb{P}_{\boldsymbol{\Sigma}_\theta}(\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) \neq \mathcal{T}_\theta) \geq 1/2$$

provided $n < (\log(p-2)-2)/\rho_{\min}^2$. Moreover, for any CPDAG estimator $\widehat{C}$, there holds

$$\sup_{\theta \in \mathcal{C}(\rho_{\min})} \mathbb{P}_{\boldsymbol{\Sigma}_\theta}(\widehat{C}(\boldsymbol{X}^{(1:n)}) \neq C_{T_\theta}) \geq 1/2$$

provided $n < \left(\log \frac{(p-1)(p-2)}{2} - 2\right)/(5\rho_{\min}^4)$.

Compare Theorem 3.3.7 with Theorems 3.3.5 and 3.3.6, we can conclude that our derived sufficient conditions on the sample sizes for the recovery of both skeleton and CPDAG are sharp.

3.4. Extension to Sub-Gaussian Models and Inverse Correlation Matrix Estimation

3.4.1. Sub-Gaussian Case. We now move to the sub-Gaussian case for the linear SEM (3.1) associated to a polytree $T = (V, E)$, in which we replace the Gaussian assumption with the following sub-Gaussian assumption imposed on the independent noise variables.

Assumption 3.4.1. $\epsilon_1, \ldots, \epsilon_p$ are independent mean-zero sub-Gaussian random variables satisfying $\mathbb{E}[e^{\lambda \epsilon_i}] \leq e^{\frac{1}{2}\lambda^2 \kappa \omega_{ii}}$ for all $\lambda \in \mathbb{R}$, where κ is some constant and $\omega_{ii} = \mathrm{var}(\epsilon_i)$. In other words, the squared sub-Gaussian parameter of ϵ_i is upper bounded by $\kappa \omega_{ii}$.

Remark 10. If ϵ_i is a mean-zero Gaussian random variable, then $\kappa = 1$.

Remark 11. Assumption 3.4.1 actually implies the linear invariance for the parameter κ. In fact, any linear combination $X \coloneqq a_1\epsilon_1 + a_2\epsilon_2 + \ldots + a_p\epsilon_p$ satisfies

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &= \mathbb{E}[e^{\lambda(a_1\epsilon_1 + \ldots + a_p\epsilon_p)}] = \mathbb{E}[e^{\lambda(a_1\epsilon_1)}] \cdots \mathbb{E}[e^{\lambda(a_p\epsilon_p)}] \\ &\leq \exp\left[\frac{\lambda^2 \kappa}{2}\left(a_1^2\omega_{11} + \ldots + a_p^2\omega_{pp}\right)\right] = e^{\frac{\lambda^2 \kappa}{2}\mathrm{var}(X)}, \quad \forall \lambda \in \mathbb{R},\end{aligned}$$

which implies that X is a sub-Gaussian random variable whose sub-Gaussian parameter is controlled by $\kappa \mathrm{var}(X)$. This fact can be applied to the components of the feature variables $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\epsilon}$ and their standardized counterparts $\tilde{\boldsymbol{X}} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\epsilon}$. As before, $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are the standard deviations of $X_1, \ldots, X_p$.

Recall that in the Gaussian case the discrepancy between population and sample pairwise correlations are quantified in Lemma 3.6.1. For the sub-Gaussian case, such discrepancy is quantified in Lemma 3.6.4. Combining it with the correlation decay property in Corollary 3.3.2, we can establish the following CPDAG exact recovery result under the sub-Gaussian case. Given the proof is exactly the same as that of Theorem 3.3.6, we skip the detailed argument and directly give the statement.

THEOREM 3.4.2. Consider the sub-Gaussian linear SEM (3.1) associated to a polytree $T = (V, E)$, in which Assumption 3.4.1 is assumed to be true. Also, assume $\rho_{\min} > 0$ and $\rho_{\max} < 1$. Denote by $\hat{C}(\boldsymbol{X}^{(1:n)})$ the estimated CPDAG from the entire algorithm discussed in Sections 3.2.3.1 and 3.2.3.2 with threshold ρ_{crit}, and by C_T the true CPDAG of the polytree T. Denote $\gamma = \min\left\{\frac{\rho_{\min}}{3}, \frac{1-\rho_{\max}}{2}\right\}\rho_{\min}$. For any $\delta \in (0,1)$, on an event with probability at least $1-\delta$, we have $|\hat{\rho}_{ij} - \rho_{ij}| < \gamma$ for any $i < j$ and $\hat{C}(\boldsymbol{X}^{(1:n)}) = C_T$, provided $\gamma < \rho_{crit} < \rho_{\min}^2 - \gamma$ and

$$n > 2\max\left\{\frac{128\kappa^2(2+\gamma)^2}{\gamma^2}, \frac{8\kappa(2+\gamma)}{\gamma}\right\}\log\frac{4p^2}{\delta}.$$

3.4.2. Inverse Correlation Matrix Estimation. In this section, let's consider the linear polytree SEM and also assume the event described in either Theorem 3.3.6 or Theorem 3.4.2, that is, the true CPDAG is exactly recovered, and $|\hat{\rho}_{ij} - \rho_{ij}| < \gamma$ for any $i < j$. Under this situation, we are interested in recovering the inverse correlation matrix of the polytree model, which can be used to estimate the partial correlations, and is thereby useful in constructing undirected graphical models empirically with some tuning threshold. To estimate the inverse correlation matrix, due to scaling invariance of population and sample correlations, without loss of generality, we assume that all X_i's have unit variances. Then the inverse correlation matrix is $\boldsymbol{\Theta} \coloneqq \boldsymbol{\Sigma}^{-1} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{B}^\top)$. The major goal of this subsection is to study how well we can estimate $\boldsymbol{\Theta}$.

At first, let's choose one realization from the equivalence class represented by this CPDAG, and still refer to it as T with no confusion. By $\boldsymbol{\Theta} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{B}^\top)$, the elements in the inverse

correlation matrix $\mathbf{\Theta}$ are given by

$$\theta_{ij} = \begin{cases} -\beta_{ij}/\omega_{jj} & \text{if } i \to j \in T \\ -\beta_{ji}/\omega_{ii} & \text{if } j \to i \in T \\ \beta_{ik}\beta_{jk}/\omega_{kk} & \text{if } i \to k \leftarrow j \in T \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i \neq j$$

$$\theta_{jj} = \frac{1}{\omega_{jj}} + \sum_{k \in Ch(j)} \frac{\beta_{jk}^2}{\omega_{kk}}, \text{ for } j = 1, \dots p.$$

Notice that the k in $i \to k \leftarrow j \in T$ must be unique in a polytree.

Since all variables have unit variance, it has been explained in Section 3.3.1 that for each $i \to j \in T$, β_{ij} is actually the correlation coefficient ρ_{ij} between X_i and X_j, so we can represent the entries of the inverse correlation matrix by the correlation coefficients over the polytree as

$$\theta_{ij} = \begin{cases} -\rho_{ij}/\omega_{jj} & \text{if } i \to j \in T \\ -\rho_{ji}/\omega_{ii} & \text{if } j \to i \in T \\ \rho_{ik}\rho_{jk}/\omega_{kk} & \text{if } i \to k \leftarrow j \in T \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i \neq j \tag{3.9}$$

$$\theta_{jj} = \frac{1}{\omega_{jj}} + \sum_{k \in Ch(j)} \frac{\rho_{jk}^2}{\omega_{kk}}, \text{ for } j = 1, \dots p. \tag{3.10}$$

where $\omega_{jj} = 1 - \sum_{i \in Pa(j)} \rho_{ij}^2$ for $j = 1, \dots, p$.

A natural question is whether we can represent the inverse correlation matrix only through the CPDAG C_T. This question is important given we can only hope to recover C_T by the algorithms introduced in Sections 3.2.3.1 and 3.2.3.2. We first give a useful lemma, which explains for what kind of node j, the noise variance $\omega_{jj} = 1 - \sum_{i \in Pa(j)} \rho_{ij}^2$ is well-defined on the CPDAG C_T, i.e., invariant to any particular polytree chosen from the equivalence class.

Lemma 3.4.3. Denote by C_T the true CPDAG of the polytree T. We denote by V_m the collection of nodes j such that there is at least one undirected edge $i - j$ in C_T. On the other hand, we denote V_d the collection of nodes j such that all its neighbors are connected to it with a directed

edge in C_T. This means that V_m and V_d form a partition of all nodes. Then, we have the following properties:

(1) For each $j \in V_m$, there is no i satisfying $i \to j \in C_T$.

(2) For each $j \in V_m$ and any polytree T' within the equivalence class C_T, j has at most one parent in T'.

(3) For each $j \in V_d$, since the set of parents of j is determined by the CPDAG C_T, the corresponding noise variance $\omega_{jj} = 1 - \sum_{i \in Pa(j)} \rho_{ij}^2$ is well-defined.

(4) Combining the third property and the contrapositive of the first property, we know for each $i \to j \in C_T$, we have $j \in V_d$, and the corresponding noise variance ω_{jj} is thereby well-defined.

We omit the proof since this result can be directly implied by the fact that v-structures are kept unchanged in all polytrees within the equivalence class determined by C_T. Then the following result shows that the inverse correlation matrix can be represented by the pairwise correlations on the skeleton as well as the CPDAG.

Lemma 3.4.4. Let V_m and V_d be the partition of all nodes defined in Lemma 3.4.3. Then, the inverse correlation matrix can be represented as

$$\theta_{ij} = \begin{cases} -\rho_{ij}/\omega_{jj} & \text{if } i \to j \in C_T \\ -\rho_{ji}/\omega_{ii} & \text{if } j \to i \in C_T \\ -\rho_{ij}/(1-\rho_{ij}^2) & \text{if } i - j \in C_T \\ \rho_{ik}\rho_{jk}/\omega_{kk} & \text{if } i \to k \leftarrow j \in C_T \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i \neq j$$

and

$$\theta_{jj} = \begin{cases} \frac{1}{\omega_{jj}} + \sum_{j \to k \in C_T} \frac{\rho_{jk}^2}{\omega_{kk}}, & j \in V_d, \\ 1 + \sum_{j - k \in C_T} \frac{\rho_{jk}^2}{1-\rho_{jk}^2} + \sum_{j \to k \in C_T} \frac{\rho_{jk}^2}{\omega_{kk}}, & j \in V_m. \end{cases}$$

Here $\omega_{jj} = 1 - \sum_{i \in Pa(j)} \rho_{ij}^2$ is well-defined in all of the above formulas, since $Pa(j)$ is well-defined for any $j \in V_d$.

The result can be obtained relatively straightforward by the facts listed in Lemma 3.4.3. How to obtain the formula of θ_{jj} for $j \in V_m$ from (3.10) may not be too obvious, since for polytree corresponding to C_T, j may have one or zero parent. It turns out that these two cases lead to the same formula given in the lemma. We omit the detailed argument for the proof.

Based the above lemma, we can give an estimate of the inverse correlation matrix through pairwise sample correlations over the estimated tree skeleton in combination with the estimated CPDAG $\widehat{C}_T$:

$$\hat{\theta}_{ij} = \begin{cases} -\hat{\rho}_{ij}/\hat{\omega}_{jj} & \text{if } i \to j \in \widehat{C}_T \\ -\hat{\rho}_{ji}/\hat{\omega}_{ii} & \text{if } j \to i \in \widehat{C}_T \\ -\hat{\rho}_{ij}/(1-\hat{\rho}_{ij}^2) & \text{if } i - j \in \widehat{C}_T \\ \hat{\rho}_{ik}\hat{\rho}_{jk}/\hat{\omega}_{kk} & \text{if } i \to k \leftarrow j \in \widehat{C}_T \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i \neq j \tag{3.11}$$

and

$$\hat{\theta}_{jj} = \begin{cases} \frac{1}{\hat{\omega}_{jj}} + \sum\limits_{j \to k \in \widehat{C}_T} \frac{\hat{\rho}_{jk}^2}{\hat{\omega}_{kk}}, & j \in \widehat{V}_d, \\ 1 + \sum\limits_{j - k \in \widehat{C}_T} \frac{\hat{\rho}_{jk}^2}{1-\hat{\rho}_{jk}^2} + \sum\limits_{j \to k \in \widehat{C}_T} \frac{\hat{\rho}_{jk}^2}{\hat{\omega}_{kk}}, & j \in \widehat{V}_m. \end{cases} \tag{3.12}$$

Here $\widehat{V}_d$ and $\widehat{V}_m$ are similarly defined as in Lemma 3.4.3. Also, for any $j \in \widehat{V}_d$, we have $\hat{\omega}_{jj} = 1 - \sum\limits_{i \in \widehat{Pa}(j)} \hat{\rho}_{ij}^2$, where $\widehat{Pa}(j)$ is the estimated parent set determined in the estimated CPDAG $\widehat{C}_T$.

Finally, we introduce our result regarding the estimation error bounds of inverse correlation matrix estimation defined above.

THEOREM 3.4.5. Consider the linear polytree SEM (3.1) where X_j has unit variance for each $j \in V$. Denote by $\hat{C}(\boldsymbol{X}^{(1:n)})$ the estimated CPDAG and by C_T the true CPDAG from the polytree T. For $\varepsilon > 0$, consider the events $E_\rho(\varepsilon) = \{|\hat{\rho}_{ij} - \rho_{ij}| \leq \varepsilon, \forall i \to j \in T\}$ and $E_{C_T} = \{\hat{C}(\boldsymbol{X}^{(1:n)}) = C_T\}$. Then on the event $E_\rho(\varepsilon) \cap E_{C_T}$, the estimated inverse correlation matrix defined in (3.11) and (3.12) satisfies

$$\sum_{j=1}^{p} |\hat{\theta}_{jj} - \theta_{jj}| \leq C_0 \left(\frac{d_* p}{\omega_{\min}^2} \right) \epsilon,$$

and

$$\sum_{i \neq j} |\hat{\theta}_{ij} - \theta_{ij}| \leq C_0 \left(\frac{d_*^2 p}{\omega_{\min}^2} \right) \epsilon,$$

for some absolute constant C_0 provided $\varepsilon < \omega_{\min}/(4d_*)$. Here $d_* = \max\{d_j^{in} : j \in V_d\} \vee 1$ and $\omega_{\min} = \min\{\omega_{jj} : j \in V_d\} \wedge \min\{1 - \rho_{ij}^2 : i - j \in C_T\}$, both of which only depend on C_T.

REMARK 12. Under Gaussian or sub-Gaussian cases, the event $E_\rho(\varepsilon) \cap E_{C_T}$ occurs with high probability when the sample size is enough. Taking Theorem 3.3.6 for instance, when the sufficient sample size condition is satisfied, $E_\rho(\varepsilon) \cap E_{C_T}$ occurs with high probability with $\varepsilon = \min\left\{ \frac{\rho_{\min}}{3}, \frac{1-\rho_{\max}}{2} \right\} \rho_{\min}$.

REMARK 13. Here d_* and $\omega_{\min}$ have clear interpretations for any polytree T' chosen from the equivalence class determined by C_T. In fact, for each $j \in V_m$, there holds $d_j^{in} \leq 1$ in T', so d_* is the maximum in-degree of T'. Again, for each $j \in V_m$, it is easy to verify that $\omega_{jj} \geq \min\{1 - \rho_{ij}^2 : i - j \in C_T\}$, so we know $\omega_{jj} \geq \omega_{\min}$ for $j = 1, \ldots, p$, i.e., $\omega_{\min}$ is a uniform lower bound of the noise variances of $\epsilon_1, \ldots, \epsilon_p$ induced by any polytree within the equivalence class.

3.5. Numerical Experiments

To illustrate the feasibility and quantitative performance of the polytree learning method, we implement Algorithms 1 and 2 in Python and test on simulated data (Section 3.5.1). We further test on commonly used benchmark datasets (Section 3.5.2) to assess the robustness and applicability to real-world data. In all experiments, we set the threshold ρ_{crit} (Algorithm 2) for rejecting a pair of nodes being independent based on the testing zero correlation for Gaussian distributions. Specifically, $\rho_{crit} = \sqrt{1 - \frac{1}{1+t_{\alpha/2}^2/(n-2)}}$, where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a t-distribution with $df = n - 2$, and we use $\alpha = 0.1$. As comparisons, we run these same data using two basic and representative structural learning methods: the score-based hill-climbing (Gámez et al., 2011) and the constraint-based PC algorithm (Spirtes et al., 2000). We use R implementations of these two algorithms from `bnlearn` and `pcalg` packages, respectively, along with all the default options and parameters. An $\alpha = 0.01$ is used for the PC algorithm as recommended in Kalisch and Bühlman (2007). All codes are available at `https://github.com/huyu00/linear-polytree-SEM`.

We assess the results by comparing the true and inferred CPDAGs G and $\hat{G}$. On the skeleton level, there can be edges in G that are *missing* in $\hat{G}$, and vice versa $\hat{G}$ can have *extra* edges. For the CPDAG, we consider a directed edge to be *correct* if it occurs with the same direction in both CPDAGs. For an undirected edge, it needs to be undirected in both CPDAGs to be considered correct. Any other edges that occur in both CPDAGs are considered having *wrong directions*. With these notions, we can calculate the False Discovery Rate (FDR) for the skeleton as $\frac{|\text{extra}|}{|\hat{G}|}$, and for the CPDAG as $\frac{|\text{extra}|+|\text{wrong direction}|}{|\hat{G}|}$. Here $|\text{extra}|$ is the number of extra edges, $|\hat{G}|$ is the number of edges in $\hat{G}$ and so on. To quantify the overall similarity and taken into account the true positives, we calculate the Jaccard index (JI), which is $\frac{|\text{correct}|+|\text{wrong direction}|}{|G \cup \hat{G}|} = \frac{|\text{correct}|+|\text{wrong direction}|}{|\text{missing}|+|\hat{G}|}$ for the skeleton, and $\frac{|\text{correct}|}{|G|+|\hat{G}|-|\text{correct}|}$ for the CPDAG.

3.5.1. Testing on Simulated Polytree Data. Here we briefly describe how we generate linear polytree SEMs. Additional implementation details can be found in Section 3.5.3. First, we generate a polytree by randomly assigning directions to a random undirected tree. Next, the standardized SEM parameters β_{ij}'s (as in Lemma 3.3.1) are randomly chosen within a range, which in turn determine ω_{ii} (Eq. (3.5)). Motivated by the theoretical results (Theorems 3.3.5 to 3.3.7), we make sure that in the above procedures, the generated SEM satisfies $\rho_{\min} \leq |\beta_{ij}| \leq \rho_{\max}$, the maximum in-degree is $d^{\text{in}}_{\max}$, and $\omega_{\min} \leq \omega_{ii}$ for all $i \in V$. Here $\rho_{\min}, \rho_{\max}, d^{\text{in}}_{\max}, \omega_{\min}$ are pre-specified constants (values used are listed in Fig. 3.1 caption).

Figures 3.1 and 3.2 show the performance for $p = 100$ and n ranging from 50 to 1000. We see that the polytree learning performs much better than the hill-climbing, and overall has an accuracy similar to or better than that of the PC algorithm. For small sample size less than 400, the PC algorithm has a smaller FDR for skeleton recovery than the polytree learning, but this is likely at the expense of the true positive rate, as reflected by the similar or lower JI of the PC algorithm comparing to the polytree learning (Panels BD of Figs. 3.1 and 3.2). As $\rho_{\min}$ becomes smaller or as $d^{\text{in}}_{\max}$ increases, the accuracy of the polytree learning decreases, which is consistent with the theory (Theorems 3.3.5 to 3.3.7). For the hill-climbing and the PC, the accuracy is less affected by $\rho_{\min}$ or $d^{\text{in}}_{\max}$ (Fig. 3.1 vs Fig. 3.2). Interestingly, the running time of the PC algorithm is significantly affected by $d^{\text{in}}_{\max}$: the running time increases 40 folds when $d^{\text{in}}_{\max}$ changes from 10 to 20 (Table 3.3), and the code may even fail to stop (running for more than 8 hours) when $d^{\text{in}}_{\max} = 40$ (data not

shown). This phenomenon can be explained by the relationship between the maximal number of neighbors and the maximal number of iterations in the PC algorithm; see Proposition 1 of Kalisch and Bühlman (2007). On the other hand, the polytree learning is significantly more favorable in terms of running time. It is up to 80 times faster than the slowest alternative algorithm (Table 3.3) and, importantly, has a running time that is constant across the SEM parameters (this is also true for all other experiments described later).

3.5.2. Testing on DAG Benchmark Data. The ALARM dataset (Beinlich et al., 1989) is a widely used benchmark data. The true DAG (Fig. 3.3) has 37 nodes and 46 edges, hence there has to be at least 10 edges missing in the inferred polytree. In fact, a three-phase algorithm initialized by polytree learning has been demonstrated to be effective on this data (Cheng et al., 2002). We simply conducted the polytree learning algorithm introduced in Sections 3.2.3.1 and 3.2.3.2, and found that it still performs better than hill-climbing and PC algorithm in terms of the metrics (Table 3.1) as well as intuitively by the resulting graph (Fig. 3.3). At $n = 5000$, it even achieves the best possible accuracy for skeleton recovery as polytree learning can achieve (10 missing edges and 0 extra).

$n = 500$	Correct	Wrong d.	Missing	Extra	FDR sk.	JI sk.	FDR CPDAG	JI CPDAG
Polytree	**28.0**	**4.0**	14.0	**4.0**	**0.11**	**0.64**	**0.22**	**0.52**
Hill-climbing	24.0	17.0	**5.0**	60.0	0.59	0.39	0.76	0.2
PC	14.0	17.0	15.0	13.0	0.3	0.53	0.68	0.18
$n = 5000$	Correct	Wrong d.	Missing	Extra	FDR sk.	JI sk.	FDR CPDAG	JI CPDAG
Polytree	25.0	**11.0**	10.0	**0.0**	**0.0**	**0.78**	**0.31**	**0.44**
Hill-climbing	**27.0**	18.0	**1.0**	62.0	0.58	0.42	0.75	0.21
PC	24.0	17.0	5.0	12.0	0.23	0.71	0.55	0.32

TABLE 3.1. Performance on ALARM data. See text for the details of the accuracy measures: the number of correct, missing, extra and wrong direction edges, FDR and Jaccard index for skeleton and CPDAG. The best results across the three algorithms are in bold.

Another benchmark we test is the ASIA dataset (Lauritzen and Spiegelhalter, 1988), which is a simulated DAG dataset with eight nodes. Note that the ground truth is sparse but not exactly a polytree. At $n = 500$ samples, the performance of polytree learning is comparable to that of hill-climbing and PC algorithm, while the hill-climbing gives the best result at $n = 5000$ (Table 3.2). We illustrate the comparison intuitively by plotting the most likely inference outcomes of each algorithm across the bootstrap trials in Fig. 3.4 (where we resample n observations from the original

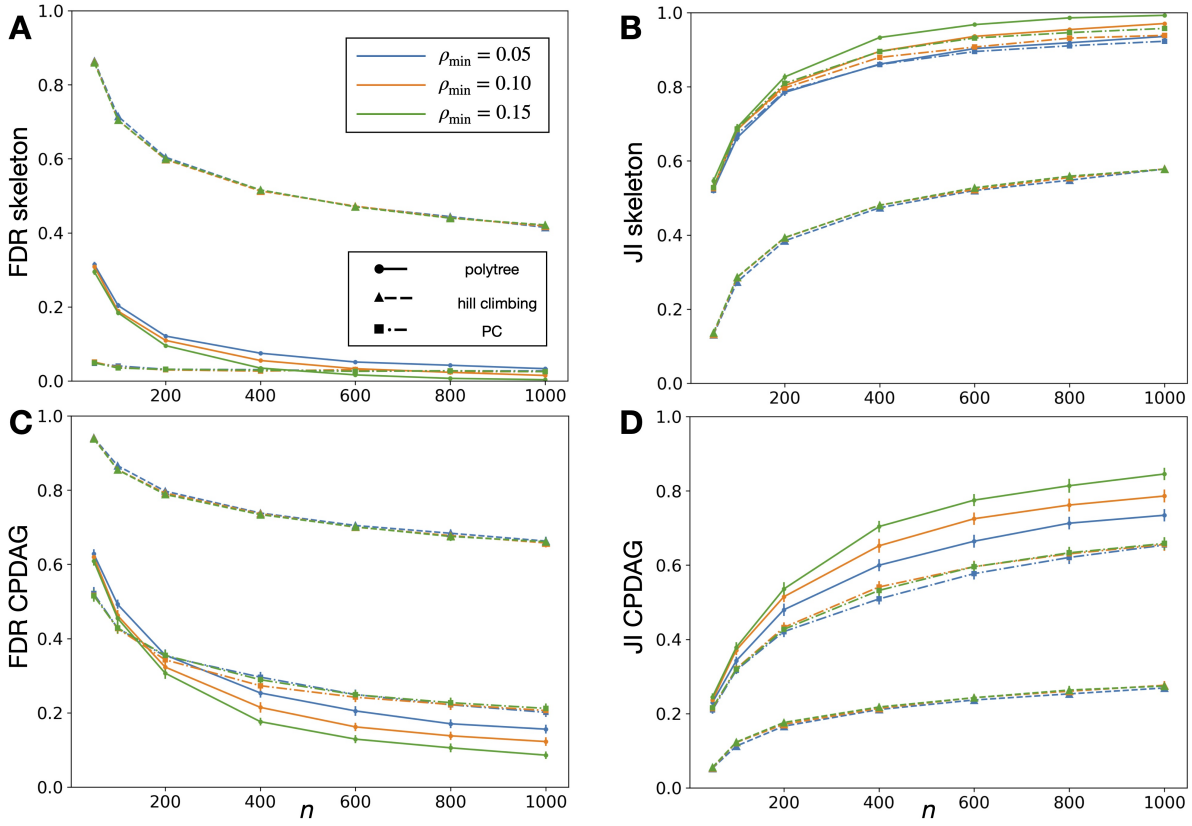


FIGURE 3.1. Performance on the polytree simulated data at $p = 100$ and the maximum in-degree $d^{\text{in}}_{\max} = 10$. The results from the algorithms are represented by solid lines and dot markers (polytree), dash lines and triangle markers (hill-climbing), and dash-dot lines and square markers (PC). Colors correspond to three different values of $\rho_{\min}$. The rest of the SEM parameters are: $\rho_{\max} = 0.8$, and $\omega_{\min} = 0.1$). Panels A,C show the FDR (the smaller the better) for skeleton and CPDAG recovery. Panels B,D show the Jaccard Index (the larger the better). For each combination of the SEM parameters, we randomly generate a polytree, the detailed generation of the β_{ij}'s and ω_{ii}'s are described in Section 3.5.3. Then we draw iid samples from the SEM of different sizes (the x-axis, $n = 50, 100, 200, 400, 600, 800, 1000$) and repeat 100 times. Each point on the curves is averaged over the 100 repeats and the errorbars are 1.96 times the standard error of the mean (many are smaller than the marker).

5000 samples). Note the polytree learning graph (occurs at 23%) is the best possible result it can achieve. This is because at least one edge must be missing by polytree learning, and the v-structure involving B, E, D can no longer be identified once missing the edge ED, leading to BD being undirected.

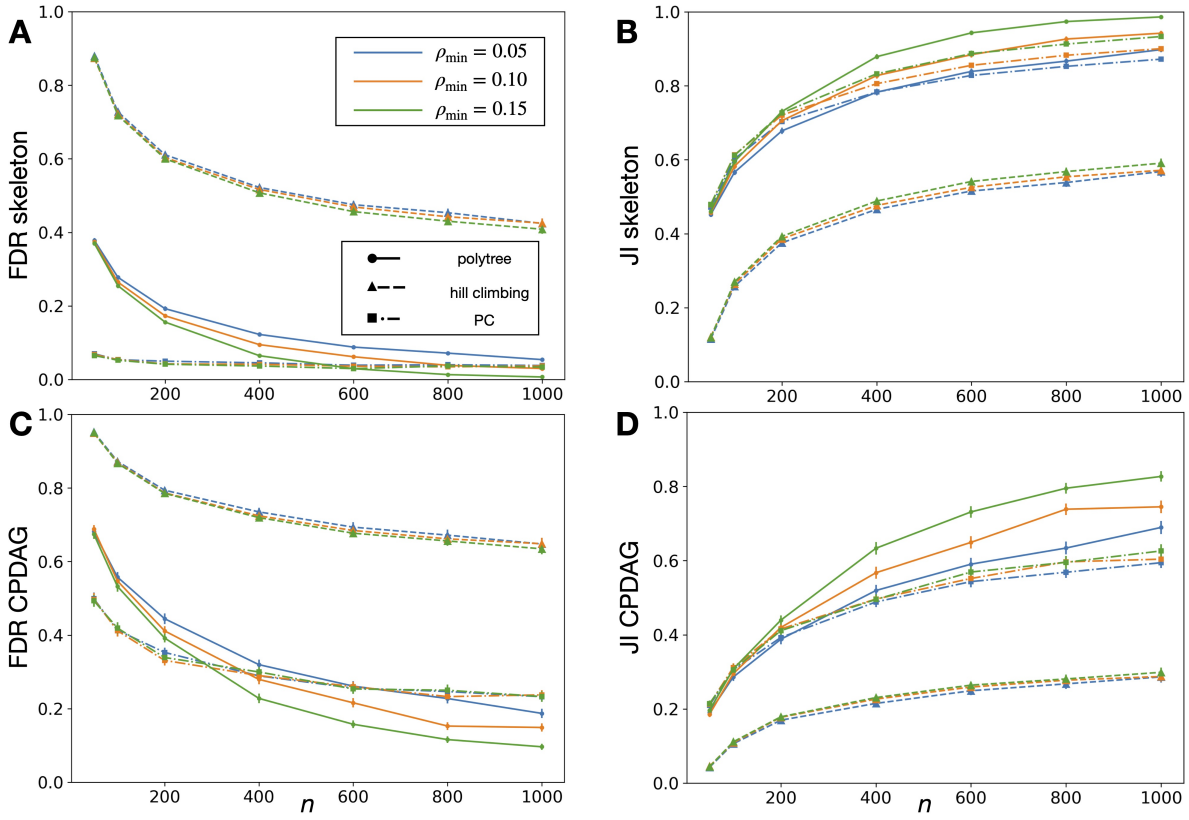


FIGURE 3.2. Same as Fig. 3.1 but for a maximum in-degree of $d^{\text{in}}_{\max} = 20$.

3.5.3. Details on polytree data generation. In simulated polytree data, we draw i.i.d. samples from a Gaussian linear SEM with a polytree structure. First, we generate an undirected tree with p nodes from a random Prufer sequence. The Prufer sequence which has a one-to-one correspondence to all the trees with p nodes is obtained by sampling $p-2$ numbers with replacement from $\{1, 2, \ldots, p\}$. Next, a polytree is obtained by randomly orienting the edges of the undirected tree. We also ensure that one of the nodes has a specified large in-degree $d^{\text{in}}_{\max}$. This is done by making a node i occur at least $d^{\text{in}}_{\max} - 1$ times in the Prufer sequence, so the node will have degree at least $d^{\text{in}}_{\max}$ in the undirected tree. We then orient all edges connected to i by selecting $d^{\text{in}}_{\max}$ of them to be incoming edges. The rest of the edges in the tree are oriented randomly as before.

In the next step, we choose the value of the standardized β_{ij} corresponding to the correlation matrix (as in Lemma 3.3.1). Note that once β_{ij}'s are given, ω_{ii} are determined by Eq. (3.5). Motivated by the theoretical conditions on n, p such as those in Theorems 3.3.5 and 3.3.6, we

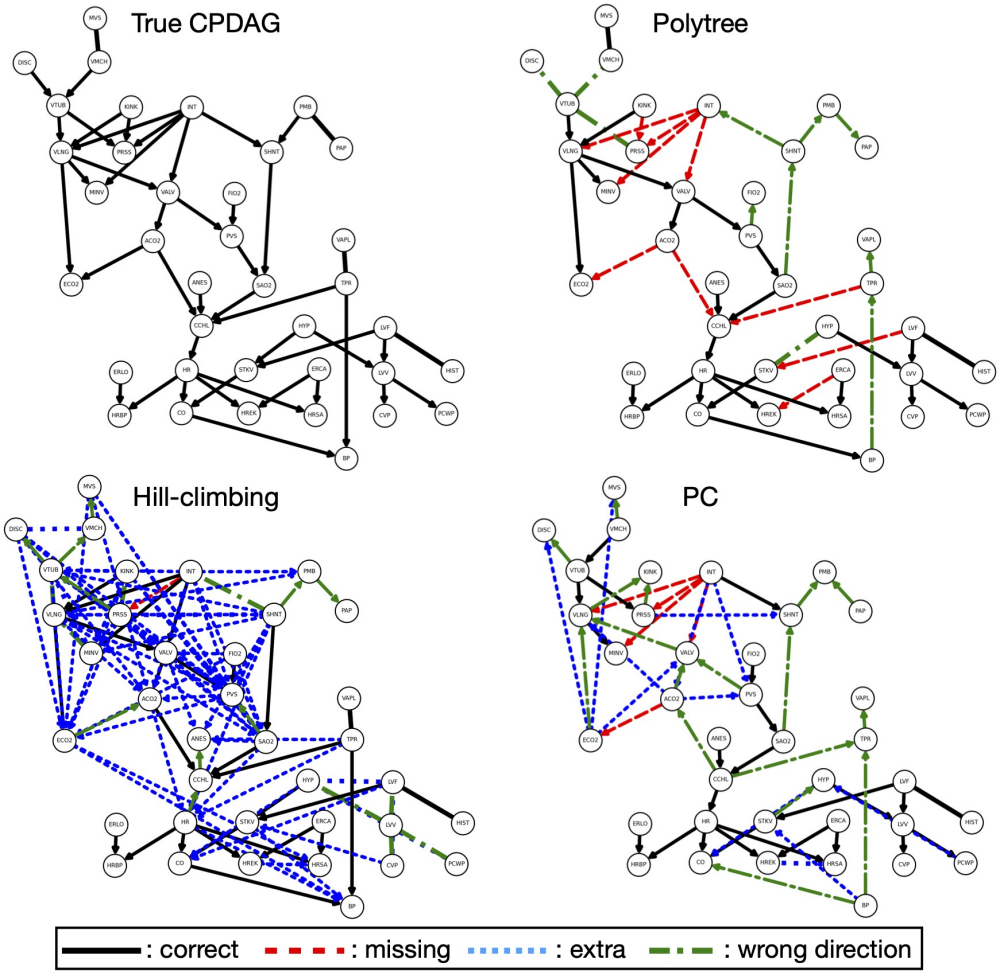


FIGURE 3.3. Comparing the true CPDAG of the ALARM data and the inferred one from the three algorithms at $n = 5000$. There are 37 nodes and 46 edges in the true CPDAG.

choose β_{ij} according to some pre-specified values $\rho_{\min}$ and $\rho_{\max}$, and study the effects of these parameters on the recovery accuracy. To avoid ill-conditioned cases, we require that $\omega_{ii} \geq \omega_{\min}$, where $\omega_{\min}$ is another parameter. This adds constraints on β_{ij}, $\sum_{j=1}^{p} \beta_{ij}^2 \leq 1 - \omega_{\min}$, in addition to $\rho_{\min} \leq |\beta_{ij}| \leq \rho_{\max}$. We sample β_{ij}^2 uniformly among the set of non-negative values satisfying the above inequality constraints. This sampling is implemented by drawing β_{ij}^2, (corresponding to all the edges in the polytree) sequentially in a random order as $\min(\rho_{\max}^2, \rho_{\min}^2 + v_j x)$, where x is drawn from the beta distribution $B(1, \tilde{d}_j^{\text{in}})$. Here $\tilde{d}_j^{\text{in}}$ is the number of incoming edges to node j whose β_{ij}^2 has not yet been chosen, and $v_j = 1 - \omega_{\min} - d_j^{\text{in}} \rho_{\min}^2 - \sum_k \beta_{kj}^2$, where the sum is over all edges

$n = 500$	Correct	Wrong d.	Missing	Extra
Polytree	3.56(1.47)	2.24(1.42)	2.2(0.85)	1.2(0.85)
Hill-climbing	**4.11(1.62)**	2.31(1.29)	**1.58(0.82)**	1.29(0.97)
PC	3.67(0.86)	**1.54(0.9)**	2.78(0.65)	**0.14(0.42)**
$n = 5000$	Correct	Wrong d.	Missing	Extra
Polytree	4.31(1.42)	2.17(1.36)	1.52(0.51)	0.52(0.51)
Hill-climbing	**6.96(0.92)**	**0.33(0.68)**	**0.7(0.57)**	0.88(0.97)
PC	4.56(1.03)	1.81(1.02)	1.63(0.53)	**0.15(0.37)**

(continue)

$n = 500$	FDR sk.	JI sk.	FDR CPDAG	JI CPDAG
Polytree	0.17(0.12)	0.64(0.13)	0.49(0.21)	0.33(0.17)
Hill-climbing	0.16(0.11)	**0.7(0.13)**	0.46(0.21)	0.38(0.21)
PC	**0.02(0.06)**	0.64(0.09)	**0.31(0.17)**	**0.39(0.11)**
$n = 5000$	FDR sk.	JI sk.	FDR CPDAG	JI CPDAG
Polytree	0.07(0.07)	0.77(0.11)	0.38(0.2)	0.43(0.18)
Hill-climbing	0.1(0.1)	**0.83(0.12)**	**0.13(0.15)**	**0.78(0.18)**
PC	**0.02(0.05)**	0.78(0.07)	0.29(0.17)	0.47(0.14)

TABLE 3.2. Performance on ASIA data. The accuracy measures (the number of correct, missing, extra and wrong direction edges, FDR and Jaccard index for skeleton and CPDAG; see text) are averaged over 1000 bootstraps (resampling n observations) and the standard deviations are in the parentheses. The best results across the three algorithms are in bold.

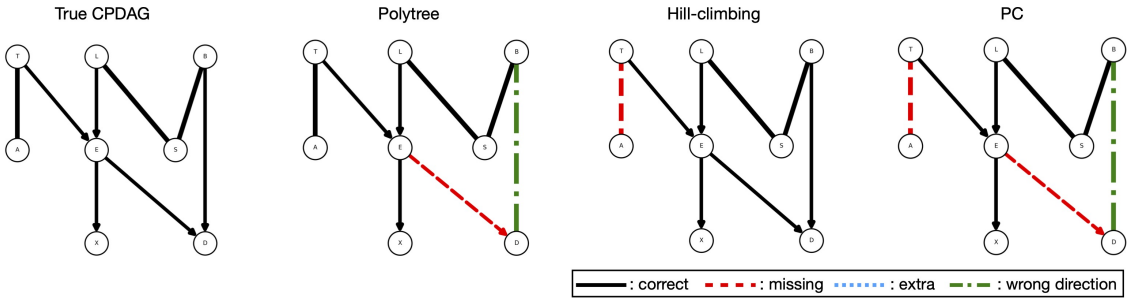


FIGURE 3.4. The true CPDAG and the typical inferred CPDAG with $n = 5000$ samples. We plot the most likely inferred graph across 1000 bootstraps for each algorithm, which occurs at 23% (polytree), 44% (hill-climbing), 42% (PC), respectively.

$k \to j$ whose β_{kj}^2 have already been chosen, d_j^{in} is the total number of incoming edges to j. The use of beta distribution here is based on the fact of the order statistics of independent uniformly distributed random variables. As an exception, we first set two $|\beta_{ij}|$ values to attain equality in the constraints by $\rho_{\min}$ and $\rho_{\max}$ before choosing the rest of β_{ij}'s according to the above sampling procedure. For $\rho_{\max}$, we randomly choose a node i that satisfies $\rho_{\min}^2(d_i^{\text{in}} - 1) + \rho_{\max}^2 \le 1 - \omega_{\min}$, $d_i^{\text{in}} > 0$ (always exists if $\rho_{\max}^2 + \omega_{\min} \le 1$ and the minimum nonzero in-degree is 1), and set one of its incoming edges to have $|\beta_{ji}| = \rho_{\max}$. For $\rho_{\min}$, we choose a node among the rest of nodes with

(Unit: sec)	Polytree $p = 100$, $d^{in}_{\max} = 10$	Polytree $p = 100$, $d^{in}_{\max} = 20$	ASIA $p = 8$	ALARM $p = 37$
Polytree	**0.01**	**0.01**	**0.01**	**0.01**
Hill-climbing	0.87	1.00	0.01	1.48
PC	0.07	2.86	0.03	0.53

TABLE 3.3. Running time comparison. The columns correspond to the SEM data in Figs. 3.1 and 3.2 (polytree), Table 3.1 (ALARM) and Table 3.2 (ASIA). $n = 5000$ for the AISA and ALARM. The running time is for one inference (averaged across trials/bootstraps when applicable). All computation is done on a 2019 i7 quad-core CPU desktop computer.

$d^{\text{in}}_k > 0$ and set $|\beta_{lk}| = \rho_{\min}$ for one of its incoming edges. Lastly, a positive or negative sign is given to each β_{ij} with equal probability. After the β_{ij}'s (i.e., matrix $\boldsymbol{B}$) are chosen (and hence $\boldsymbol{\Omega}$), the zero mean Gaussian samples $\boldsymbol{X}$ are drawn according to $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\epsilon}$.

3.6. Proofs

3.6.1. Supporting Lemmata. Here we introduce some important results that would be helpful for proving our main results.

Lemma 3.6.1 (Kalisch and Bühlman (2007), Lemma 1)**.** Consider the Gaussian linear polytree SEM (3.1) with $\rho_{\max} < 1$. For any $0 < \gamma \leq 2$, there holds

$$\sup_{i \neq j} \mathbb{P}(|\hat{\rho}_{ij} - \rho_{ij}| \geq \gamma) \leq C_1(n-2)\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right),$$

where $C_1 = \frac{(1-\rho^2_{\min})^{3/2}(3-\rho_{\max})}{(1-\rho_{\max})^{7/2}}$. This further implies that

$$\mathbb{P}\left(\bigcap_{1 \leq i < j \leq p} \{|\hat{\rho}_{ij} - \rho_{ij}| < \gamma\}\right) \geq 1 - C_1\binom{p}{2}(n-2)\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right).$$

Note that $\rho_{\min}$ is only defined over the skeleton.

Lemma 3.6.2 (e.g. Bresler and Karzand (2020), Lemma 6.1 and Lemma 8.8)**.** Let $\mathcal{T}$ be the skeleton of true polytree $T = (V, E)$ and $\widehat{\mathcal{T}}$ be the estimated tree through Chow-Liu algorithm (3.2). If an edge $(w, \tilde{w}) \in \mathcal{T}$ and $(w, \tilde{w}) \notin \widehat{\mathcal{T}}$, i.e. this edge is incorrectly missed, then there exists an edge $(v, \tilde{v}) \in \widehat{\mathcal{T}}$ and $(v, \tilde{v}) \notin \mathcal{T}$ such that $(w, \tilde{w}) \in path_{\mathcal{T}}(v, \tilde{v})$ and $(v, \tilde{v}) \in path_{\widehat{\mathcal{T}}}(w, \tilde{w})$. On such an error event, we have $|\hat{\rho}_{v\tilde{v}}| \geq |\hat{\rho}_{w\tilde{w}}|$.

Lemma 3.6.3 (Harris and Drton (2013), Lemma 7)**.** Let (X, Y) be a bivariate random vector with mean zero and covariance matrix $\mathbf{\Sigma}$. Denote the empirical covariance matrix with $\hat{\mathbf{\Sigma}}_n$ from an i.i.d. sample of size n. If $\mathbf{\Sigma}$ is positive definite with $\Sigma_{11}, \Sigma_{22} \geq 1$ and $\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_{\max} < t < 1$, where $\|\cdot\|_{\max}$ represents the elementwise maximum absolute value of a matrix, then for the population and sample correlation between X and Y, we have

$$|\hat{\rho}_{XY} - \rho_{XY}| < \frac{2t}{1-t}.$$

Lemma 3.6.4. Assume X_i and X_j are jointly distributed mean-zero sub-Gaussian random variables whose sub-Gaussian parameters are controlled by $\kappa \mathrm{Var}(X_i)$ and $\kappa \mathrm{Var}(X_j)$, respectively. Assume we have i.i.d. samples from their joint distribution as $(X_i^{(1)}, X_j^{(1)})$, ..., $(X_i^{(n)}, X_j^{(n)})$. Then, for any $0 < \gamma \leq 2$, their population and sample correlation coefficients satisfy

$$\mathbb{P}(|\hat{\rho}_{ij} - \rho_{ij}| > \gamma) \leq 8 \exp\left\{-\frac{n}{2} \min\left\{\frac{\gamma^2}{128\kappa^2(2+\gamma)^2}, \frac{\gamma}{8\kappa(2+\gamma)}\right\}\right\}.$$

Proof. Notice that both population and sample correlation coefficients are scaling invariant. Therefore, WLOG, we can assume that $\mathrm{Var}(X_i) = \mathrm{Var}(X_j) = 1$. Remark 11 implies that their sub-Gaussian parameters are both controlled by κ.

It is known that X_i^2 is sub-Exponential with parameters $(32\kappa^2, 4\kappa)$ (Honorio and Jaakkola, 2014). In other words, it holds that

$$\mathbb{E}[e^{\lambda(X_i^2 - \mathbb{E}[X_i^2])}] \leq e^{16\kappa^2}, \quad \forall |\lambda| \leq \frac{1}{4\kappa}.$$

Note that our assumption gives $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) = 1$. By sub-Exponential tail bound, for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{l=1}^{n}(X_i^{(l)})^2 - \mathrm{Var}(X_i)\right| > t\right) \leq 2\exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{32\kappa^2}, \frac{t}{4\kappa}\right\}\right\}.$$

Similarly, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{l=1}^{n}(X_j^{(l)})^2 - \mathrm{Var}(X_j)\right| > t\right) \leq 2\exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{32\kappa^2}, \frac{t}{4\kappa}\right\}\right\}.$$

For the covariance term between X_i and X_j, note that $X_i X_j = \frac{(X_i+X_j)^2 - (X_i - X_j)^2}{4}$. Since X_i and X_j are sub-Gaussian variables, we have that $X_i \pm X_j$ are both sub-Gaussian with parameter 4κ.

Then, the sub-Exponential tail bound can be applied to $(X_i \pm X_j)^2$ to obtain the following result.

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{l=1}^{n} X_i^{(l)} X_j^{(l)} - \mathrm{Cov}(X_i, X_j)\right| > t\right) \\
\leq &\mathbb{P}\left(\left|\frac{1}{4n}\sum_{l=1}^{n}(X_i^{(l)} + X_j^{(l)})^2 - \frac{1}{4}\mathbb{E}(X_i^{(1)} + X_j^{(1)})^2\right| > \frac{t}{2}\right) \\
&+ \mathbb{P}\left(\left|\frac{1}{4n}\sum_{l=1}^{n}(X_i^{(l)} - X_j^{(l)})^2 - \frac{1}{4}\mathbb{E}(X_i^{(1)} - X_j^{(1)})^2\right| > \frac{t}{2}\right) \\
\leq &4\exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{128\kappa^2}, \frac{t}{8\kappa}\right\}\right\}.
\end{aligned}$$

Denote by $\mathbf{\Sigma}^{ij}$ and $\widehat{\mathbf{\Sigma}}_n^{ij}$ the population and sample covariance matrices of X_i and X_j, respectively. A union bound gives

$$\mathbb{P}(\|\widehat{\mathbf{\Sigma}}_n^{ij} - \mathbf{\Sigma}^{ij}\|_{\max} > t) \leq 8\exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{128\kappa^2}, \frac{t}{8\kappa}\right\}\right\}.$$

Applying Lemma 3.6.3 and setting $t = \frac{\gamma}{2+\gamma}$ for some $0 < \gamma \leq 2$, we have

$$\begin{aligned}
\mathbb{P}(|\hat{\rho}_{ij} - \rho_{ij}| > \gamma) &\leq \mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_n^{ij} - \mathbf{\Sigma}^{ij}\|_{\max} > \frac{\gamma}{2+\gamma}\right) \\
&\leq 8\exp\left\{-\frac{n}{2}\min\left\{\frac{\gamma^2}{128\kappa^2(2+\gamma)^2}, \frac{\gamma}{8\kappa(2+\gamma)}\right\}\right\}.
\end{aligned}$$

□

3.6.2. Proof of Theorem 3.2.1.

Proof. Each connected component of the undirected edges is a sub-graph of the polytree G's skeleton, thus is a tree. If a node of the tree also has directed edges, they must be outgoing according to Line 6 of Algorithm 2 (Rule 1 in Meek (1995)). This means that when we convert each undirected tree into a rooted tree, it does not create any additional v-structures in the resulting DAG G'. So the original CPDAG is also the CPDAG of G', i.e., G' is equivalent to G. On the other hand, if G' is an equivalent DAG, for each undirected tree T in the CPDAG, let i be a source node of T according to G'. Then T in G' must be a rooted tree with i being the root to avoid having v-structures within T (and hence contradicting with G' shares the same CPDAG). This shows that

all equivalent class members can be obtained by orienting undirected trees into rooted trees and completes the proof. □

3.6.3. Proof of Theorem 3.3.5.

PROOF. Denote $\gamma = \rho_{\min}(1-\rho_{\max})/2$. Consider the event

$$E = \bigcap_{1\le i<j\le p} \{|\hat{\rho}_{ij} - \rho_{ij}| < \gamma\}.$$

By Lemma 3.6.1, we have

$$\mathbb{P}(E) \ge 1 - C_1\binom{p}{2}(n-2)\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right),$$

where $C_1 = \frac{(1-\rho_{\min}^2)^{3/2}(3-\rho_{\max})}{(1-\rho_{\max})^{7/2}}$.

Consider any undirected edge $(w,\tilde{w}) \in \mathcal{T}$ and any non-adjacent pair $(v,\tilde{v})$ such that $(w,\tilde{w}) \in path_{\mathcal{T}}(v,\tilde{v})$. According to Corollary 3.3.2, there are two possible cases. If $path_{S_T}(v,\tilde{v})$ corresponds to a simple trek in the polytree T, then $\rho_{v\tilde{v}}$ consists of the product among several correlation coefficients containing $\rho_{w\tilde{w}}$. Hence $|\rho_{v\tilde{v}}| \le |\rho_{w\tilde{w}}|\rho_{\max}$. On the contrary, if $path_{\mathcal{T}}(v,\tilde{v})$ is not a simple trek in T, then we have $\rho_{v\tilde{v}} = 0$. Overall, we can obtain an upper bound for $|\rho_{v\tilde{v}}| - |\rho_{w\tilde{w}}|$.

$$|\rho_{v\tilde{v}}| - |\rho_{w\tilde{w}}| \le |\rho_{w\tilde{w}}|(\rho_{\max}-1) \le \rho_{\min}(\rho_{\max}-1). \tag{3.13}$$

Under the event E, the triangular inequality gives that

$$\begin{aligned} |\hat{\rho}_{v\tilde{v}}| - |\hat{\rho}_{w\tilde{w}}| &= |\hat{\rho}_{v\tilde{v}}| - |\rho_{v\tilde{v}}| + |\rho_{v\tilde{v}}| - |\rho_{w\tilde{w}}| - (|\hat{\rho}_{w\tilde{w}}| - |\rho_{w\tilde{w}}|) \\ &\le |\hat{\rho}_{v\tilde{v}} - \rho_{v\tilde{v}}| + |\hat{\rho}_{w\tilde{w}} - \rho_{w\tilde{w}}| + |\rho_{v\tilde{v}}| - |\rho_{w\tilde{w}}| \\ &< 2\gamma + \rho_{\min}(\rho_{\max}-1) = 0. \end{aligned} \tag{3.14}$$

Notice that this holds uniformly for any undirected edge $(w,\tilde{w}) \in \mathcal{T}$ and any non-adjacent pair $(v,\tilde{v})$ such that $(w,\tilde{w}) \in path_{\mathcal{T}}(v,\tilde{v})$.

Under the event $\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) \ne \mathcal{T}$, by Lemma 3.6.2, we know there is an edge $(w,\tilde{w}) \in \mathcal{T}$ and a non-adjacent pair $(v,\tilde{v})$, such that $(w,\tilde{w}) \in path_{\mathcal{T}}(v,\tilde{v})$ while $|\hat{\rho}_{v\tilde{v}}| \ge |\hat{\rho}_{w\tilde{w}}|$. Then we have

$$E \subset \{\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) = \mathcal{T}\}.$$

It suffices to study the tail probability of E, which satisfies

$$C_1(n-2)\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right)\binom{p}{2} \le C_1\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right)\frac{np^2}{2}.$$

Then we have $\mathbb{P}(E) < \delta$ by requiring the sample size to satisfy

$$n > \log\left(\frac{(1-\rho_{\min}^2)^{3/2}(3-\rho_{\max})np^2}{2(1-\rho_{\max})^{7/2}\delta}\right)/\log\left(\frac{4+\gamma^2}{4-\gamma^2}\right)+4.$$

This condition can be implied by the sample complexity condition (3.7) by the equality $\log(1+x) \ge \frac{x}{1+x}$ for any positive x. □

3.6.4. Proof of Theorem 3.3.6.

PROOF. With $\gamma = \min\left\{\frac{\rho_{\min}}{3}, \frac{1-\rho_{\max}}{2}\right\}\rho_{\min}$, consider the event

$$E' = \bigcap_{1\le i<j\le p}\{|\hat{\rho}_{ij} - \rho_{ij}| < \gamma\}.$$

By Lemma 3.6.1, we have

$$\mathbb{P}(E') \ge 1 - C_1\binom{p}{2}(n-2)\exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right),$$

where $C_1 = \frac{(1-\rho_{\min}^2)^{3/2}(3-\rho_{\max})}{(1-\rho_{\max})^{7/2}}$. Similar to the argument in Theorem 3.3.5, under this event the the Chow–Liu algorithm recovers the true skeleton of the polytree exactly, i.e.,

$$E' \subset \{\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) = \mathcal{T}\}.$$

It suffices to show that by choosing ρ_{crit} that satisfies $\gamma < \rho_{crit} < \rho_{\min}^2 - \gamma$ in Algorithm 2, all v-structures are correctly identified on the event E'. Let's consider all node triplets $i-k-j$ in $\mathcal{T}$. If the ground truth is $i \to k \leftarrow j$, we know that $\rho_{ij} = 0$ and then on E' we have $|\hat{\rho}_{ij}| \le \gamma \le \rho_{crit}$. This means the v-structure is identified by Algorithm 2. In contrast, if the ground truth is $i \leftarrow k \leftarrow j$ or $i \leftarrow k \to j$ or $i \to k \to j$, Corollary 3.3.2 implies that $|\rho_{ij}| = |\rho_{ik}||\rho_{kj}| \ge \rho_{\min}^2$, and then on E' there holds $|\hat{\rho}_{ij}| \ge |\rho_{ij}| - \gamma \ge \rho_{\min}^2 - \gamma > \rho_{crit}$. This means this triplet is correctly identified as a non-v-structure. In sum, we know that on the event E', we identify all the v-structures exactly. Then the CPDAG of T can be exactly recovered by Algorithm 2 as guaranteed in Meek (1995).

Finally, we have

$$\mathbb{P}(\hat{C}(\boldsymbol{X}^{(1:n)}) \neq C_T) \leq \mathbb{P}((E')^c) \leq C_1 \binom{p}{2} (n-2) \exp\left((n-4)\log\left(\frac{4-\gamma^2}{4+\gamma'^2}\right)\right).$$

Then, we know that $\mathbb{P}(\hat{C}(\boldsymbol{X}^{(1:n)}) \neq C_T) \leq \delta$ under the sample size condition (3.8) by the same argument in Theorem 3.3.5.

□

3.6.5. Proof of Theorem 3.3.7.

Proof. The key idea is to apply Fano's method to appropriate sub-classes of $\mathcal{C}(\rho_{\min})$ to establish the intended information-theoretic lower bounds for both skeleton and CPDAG recovery. Generally speaking, let $\mathcal{C}_M = \{T_1, \ldots, T_M\}$ be a sub-class of polytree models $\mathcal{C}(\rho_{\min})$ whose respective covariance matrices are denoted as $\boldsymbol{\Sigma}(T_1), \ldots, \boldsymbol{\Sigma}(T_M)$. Let model index θ be chosen uniformly at random from $\{1, \ldots, M\}$. Given the observations $\boldsymbol{X}^{(1:n)} \in \mathbb{R}^{n \times p}$, the decoder ψ estimates the underlying polytree structure with maximal probability of decoding error defined as

$$p_{\text{err}}(\psi) = \max_{1 \leq j \leq M} \mathbb{P}_{\boldsymbol{\Sigma}(T_j)}\left(\psi(\boldsymbol{X}^{(1:n)}) \neq T_j\right).$$

By Fano's inequality (Thomas and Joy, 2006), the maximal probability of error over $\mathcal{C}_M$ can be lower bounded as

$$\inf_{\psi} p_{\text{err}}(\psi) \geq 1 - \frac{I(\theta; \boldsymbol{X}^{(1:n)}) + 1}{\log M}.$$

Given all involved distributions are multivariate Gaussian, we will apply the following entropy-based bound of the mutual information that can be found in Wang et al. (2010):

$$I(\theta; \boldsymbol{X}^{(1:n)}) \leq \frac{n}{2} F(\mathcal{C}), \quad \text{where}$$

$$F(\mathcal{C}) \coloneqq \operatorname{logdet}(\overline{\boldsymbol{\Sigma}}) - \frac{1}{M}\sum_{j=1}^{M} \operatorname{logdet}(\boldsymbol{\Sigma}(T_j)) \tag{3.15}$$

and the averaged covariance matrix $\overline{\boldsymbol{\Sigma}} \coloneqq \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{\Sigma}(T_j)$.

Lower Bound for Skeleton Recovery. In the following we consider a class of polytree models $\mathcal{C}_M = \{T_1, \ldots, T_M\}$ where $M = p-2$. These polytrees share $p-2$ common directed edges $1 \to (p-1)$,

$2 \to (p-1)$, ..., $(p-2) \to (p-1)$. For the $(p-1)$-th directed edge, we let $p \to 1$ in T_1, $p \to 2$ in T_2, ..., $p \to (p-2)$ in T_{p-2}. Also, we assume that all variables have variance one, and the correlation coefficients on the skeleton are all ρ that satisfies $0 < \rho < \frac{1}{\sqrt{p}}$. Here we write $\rho = \rho_{\min}$ for simplicity. Note that the polytrees in this sub-class of $\mathcal{C}(\rho)$ (defined in the statement of Theorem 3.3.7) have distinct skeletons, so

$$\inf_{\widehat{\mathcal{T}}} \sup_{\theta \in \mathcal{C}(\rho_{\min})} \mathbb{P}_{\boldsymbol{\Sigma}_\theta}(\widehat{\mathcal{T}}(\boldsymbol{X}^{(1:n)}) \neq \mathcal{T}_\theta) \geq \inf_{\psi} \max_{1 \leq j \leq M} \mathbb{P}_{\boldsymbol{\Sigma}(T_j)} \left(\psi(\boldsymbol{X}^{(1:n)}) \neq T_j \right).$$

We can easily obtain the formula for each covariance $\boldsymbol{\Sigma}(T_j)$ for $j = 1, \ldots, M$ by using Corollary 3.3.2. For example, for T_1, we have

$$\boldsymbol{\Sigma}(T_1) = \left[\begin{array}{cccc:cc} 1 & 0 & \ldots & 0 & \rho & \rho \\ 0 & 1 & \ldots & 0 & \rho & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & \rho & 0 \\ \hdashline \rho & \rho & \ldots & \rho & 1 & \rho^2 \\ \rho & 0 & \ldots & 0 & \rho^2 & 1 \end{array}\right] \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{D} \end{bmatrix}$$

The Schur complement of $\boldsymbol{A} = \boldsymbol{I}$ is thereby

$$\boldsymbol{D} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B} = \begin{bmatrix} 1 - (p-2)\rho^2 & 0 \\ 0 & 1 - \rho^2 \end{bmatrix}.$$

Then $\det(\boldsymbol{\Sigma}(T_1)) = \det(\boldsymbol{A}) \det(\boldsymbol{D} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B}) = (1-\rho^2)(1-(p-2)\rho^2)$. Similarly, for all $j = 1, \ldots, p-2$, there holds $\det(\boldsymbol{\Sigma}(T_j)) = (1-\rho^2)(1-(p-2)\rho^2)$.

On the other hand, the average covariance is

$$\overline{\boldsymbol{\Sigma}} = \frac{1}{p-2} \sum_{j=1}^{p-2} \boldsymbol{\Sigma}(T_j) = \left[\begin{array}{cccc:cc} 1 & 0 & \ldots & 0 & \rho & \rho/(p-2) \\ 0 & 1 & \ldots & 0 & \rho & \rho/(p-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & \rho & \rho/(p-2) \\ \hdashline \rho & \rho & \ldots & \rho & 1 & \rho^2 \\ \rho/(p-2) & \rho/(p-2) & \ldots & \rho/(p-2) & \rho^2 & 1 \end{array}\right].$$

As with above, we can use Schur complement to obtain $\det(\overline{\boldsymbol{\Sigma}}) = (1-\rho^2/(p-2))(1-(p-2)\rho^2)$. Plug these results into (3.15), we have

$$F(\mathcal{C}) = \log\left(1 + \frac{(p-3)\rho^2}{(p-2)(1-\rho^2)}\right) \leq \frac{(p-3)\rho^2}{(p-2)(1-\rho^2)} \leq \frac{(p-3)\rho^2}{(p-2)(1-1/p)} \leq \rho^2.$$

Then $p_{\text{err}} \geq 1 - (\frac{n}{2}\rho^2 + 1)/\log(p-2)$. To ensure $p_{\text{err}} > 1/2$, we only need to require $1 - (\frac{n}{2}\rho^2 + 1)/\log(p-2) > 1/2$, which is equivalent to $n < (\log(p-2) - 2)/\rho^2$.

Lower Bound for CPDAG Recovery. Let's now consider another class of polytree models $\mathcal{C}_M = \{T_1, \ldots, T_M\}$ where $M = \binom{p-1}{2}$. All polytrees in this class are stars with hub node p, and p is directed to all but two nodes in $\{1, \ldots, p-1\}$. In T_1, the directed edges are $1 \to p$, $2 \to p$, $p \to 3$, $p \to 4$, ..., $p \to (p-1)$. In T_2, the directed edges are $1 \to p$, $p \to 2$, $3 \to p$, $p \to 4$, ..., $p \to (p-1)$. And so on until in T_M, the directed edges are $p \to 1$, $p \to 2$, ..., $p \to (p-3)$, $(p-2) \to p$, $(p-1) \to p$. Also, assume that all variables have variance one, and the correlation coefficients on the skeleton are all ρ that satisfies $0 < \rho < \frac{1}{2}$. Again, we write $\rho = \rho_{\min}$ for simplicity. Although the polytrees in this sub-class of $\mathcal{C}(\rho)$ have the same skeletons, but they have distinct CPDAGs since they have distinct sets of v-structures. Therefore,

$$\inf_{\widehat{C}} \sup_{\theta \in \mathcal{C}(\rho_{\min})} \mathbb{P}_{\boldsymbol{\Sigma}_\theta}(\widehat{C}(\boldsymbol{X}^{(1:n)}) \neq C_{T_\theta}) \geq \inf_{\psi} \max_{1 \leq j \leq M} \mathbb{P}_{\boldsymbol{\Sigma}(T_j)}\left(\psi(\boldsymbol{X}^{(1:n)}) \neq T_j\right).$$

Again, we have the formula for each covariance $\boldsymbol{\Sigma}(T_j)$ for $j = 1, \ldots, M$ by using Corollary 3.3.2. For example, for T_1, we have

$$\boldsymbol{\Sigma}(T_1) = \left[\begin{array}{cc:ccc:c} 1 & 0 & \rho^2 & \ldots & \rho^2 & \rho \\ 0 & 1 & \rho^2 & \ldots & \rho^2 & \rho \\ \hdashline \rho^2 & \rho^2 & 1 & \ldots & \rho^2 & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^2 & \rho^2 & \rho^2 & \ldots & 1 & \rho \\ \hdashline \rho & \rho & \rho & \ldots & \rho & 1 \end{array}\right]$$

Recall that in a linear polytree model there holds $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{\Omega}(\boldsymbol{I} - \boldsymbol{B})$. Since $\boldsymbol{B}$ can be transformed to a strict upper triangular matrix by permuting the p nodes, we know that $\det(\boldsymbol{I} - \boldsymbol{B}) =$

1. Then

$$\det(\mathbf{\Sigma}) = \det(\mathbf{\Omega}) = \prod_{j=1}^{p} \omega_{jj} = \prod_{j=1}^{p} \left(1 - \sum_{i \in Pa(j)} \rho_{ij}^2\right).$$

Then for $j = 1, \ldots, M$, there holds $\det(\mathbf{\Sigma}(T_j)) = (1-\rho^2)^{p-3}(1-2\rho^2)$, which implies that

$$\text{logdet}(\mathbf{\Sigma}(T_j)) = (p-3)\log(1-\rho^2) + \log(1-2\rho^2).$$

On the other hand, we have

$$\overline{\mathbf{\Sigma}} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{\Sigma}(T_j) = \left[\begin{array}{cccc:c} 1 & \frac{M-1}{M}\rho^2 & \ldots & \frac{M-1}{M}\rho^2 & \rho \\ \frac{M-1}{M}\rho^2 & 1 & \ldots & \frac{M-1}{M}\rho^2 & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{M-1}{M}\rho^2 & \frac{M-1}{M}\rho^2 & \ldots & 1 & \rho \\ \hdashline \rho & \rho & \ldots & \rho & 1 \end{array}\right] \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{D} \end{bmatrix}.$$

The Schur complement of $\boldsymbol{D} = 1$ is

$$\overline{\mathbf{\Sigma}}/\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\top == \begin{bmatrix} 1-\rho^2 & -\rho^2/M & \ldots & -\rho^2/M \\ -\rho^2/M & 1-\rho^2 & \ldots & -\rho^2/M \\ \vdots & \vdots & \ddots & \vdots \\ -\rho^2/M & -\rho^2/M & \ldots & 1-\rho^2 \end{bmatrix}.$$

It's easy to obtain all the eigenvalues of $\overline{\boldsymbol{\Sigma}}/\boldsymbol{D}$: $1-\frac{p+1}{p-1}\rho^2$ with multiplicity 1 and $1-\frac{p(p-3)}{(p-1)(p-2)}\rho^2$ with multiplicity $p-2$. Plug these results into (3.15), we have

$$\begin{aligned}
F(\mathcal{C}) &= \log\left(1-\frac{p+1}{p-1}\rho^2\right)+(p-2)\log\left(1-\frac{p(p-3)}{(p-1)(p-2)}\rho^2\right) \\
&\quad -\log(1-2\rho^2)-(p-3)\log(1-\rho^2) \\
&= \log\left(1-\frac{p+1}{p-1}\rho^2\right)+(p-2)\log\left(1+\frac{2}{(p-1)(p-2)}\frac{\rho^2}{1-\rho^2}\right)+\log\left(1+\frac{\rho^2}{1-2\rho^2}\right) \\
&\leq -\frac{p+1}{p-1}\rho^2+\frac{2}{p-1}\frac{\rho^2}{1-\rho^2}+\frac{\rho^2}{1-2\rho^2} \\
&= -\frac{p+1}{p-1}\rho^2+\frac{2}{p-1}\left(\rho^2+\frac{\rho^4}{1-\rho^2}\right)+\rho^2+\frac{2\rho^4}{1-2\rho^2} \\
&= \frac{2}{p-1}\frac{\rho^4}{1-\rho^2}+\frac{2\rho^4}{1-2\rho^2}<5\rho^4,
\end{aligned}$$

where the first inequality is due to $\log(1+x)\leq x$, and the second inequality is due to the assumption that $\rho^2<1/4$ and $p\geq 10$. As with the case of skeleton recovery, we know that $p_{\text{err}}>1/2$ as long as we require that

$$n<\frac{1}{5\rho^4}\left(\log\frac{(p-1)(p-2)}{2}-2\right).$$

□

3.6.6. Proof of Theorem 3.4.5.

Proof. In the CPDAG C_T, we denote $coPa(j)\coloneqq\{i:\exists k,\ s.t.\ i\to k\leftarrow j\in C_T\}$ and refer to it as the co-parent set of node j. Let's first study the off-diagonal entries of the inverse correlation matrix. We can represent the estimation error as

$$\sum_{i\neq j}|\hat{\theta}_{ij}-\theta_{ij}|=\sum_{j=1}^{p}\left(\sum_{i\to j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|+\sum_{j\to i\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|+\sum_{i-j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|+\sum_{i\in coPa(j)}|\hat{\theta}_{ij}-\theta_{ij}|\right).$$

Due to the symmetry of the inverse correlation matrix, we have

$$\sum_{j=1}^{p}\sum_{j\to i\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|=\sum_{j=1}^{p}\sum_{j\to i\in C_T}|\hat{\theta}_{ji}-\theta_{ji}|=\sum_{i=1}^{p}\sum_{i\to j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|=\sum_{j\in V_d}\sum_{i\to j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}|.$$

Then,

$$\sum_{i\neq j}|\hat{\theta}_{ij}-\theta_{ij}| = 2\sum_{j\in V_d}\sum_{i\to j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}| + \sum_{j\in V_m}\sum_{i-j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}| + \sum_{j=1}^{p}\sum_{i\in coPa(j)}|\hat{\theta}_{ij}-\theta_{ij}|.$$

For each node $j \in V_d$, recall that d_j^{in} denotes its in-degree. Lemma 3.4.4 implies that

$$|\hat{\omega}_{jj}-\omega_{jj}| = |\sum_{i\in Pa(j)}(\hat{\rho}_{ij}^2-\rho_{ij}^2)| \leq \sum_{i\in Pa(j)}|\hat{\rho}_{ij}^2-\rho_{ij}^2| \leq 2\sum_{i\in Pa(j)}|\hat{\rho}_{ij}-\rho_{ij}| \leq 2d_j^{in}\varepsilon.$$

Then

$$\begin{aligned}\sum_{i\to j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}| &= \sum_{i\to j\in C_T}\left|\frac{\hat{\rho}_{ij}}{\hat{\omega}_{jj}}-\frac{\rho_{ij}}{\omega_{jj}}\right| \\ &\leq \sum_{i\to j\in C_T}\frac{1}{\hat{\omega}_{jj}}|\hat{\rho}_{ij}-\rho_{ij}| + |\rho_{ij}|\left|\frac{1}{\hat{\omega}_{jj}}-\frac{1}{\omega_{jj}}\right| \\ &\leq d_j^{in}\frac{\varepsilon(\omega_{jj}+2\rho_{\max}d_j^{in})}{(\omega_{jj}-2d_j^{in}\varepsilon)\omega_{jj}},\end{aligned}$$

$$\begin{aligned}\sum_{i-j\in C_T}|\hat{\theta}_{ij}-\theta_{ij}| &= \sum_{i-j\in C_T}\left(\frac{\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}^2}-\frac{\rho_{ij}}{1-\rho_{ij}^2}\right) \\ &\leq \sum_{i-j\in C_T}\left[\frac{|\hat{\rho}_{ij}-\rho_{ij}|}{1-\hat{\rho}_{ij}^2}+|\rho_{ij}|\left|\frac{1}{1-\hat{\rho}_{ij}^2}-\frac{1}{1-\rho_{ij}^2}\right|\right] \\ &\leq \sum_{i-j\in C_T}\frac{3\varepsilon}{(1-\rho_{ij}^2)(1-\rho_{ij}^2-2\varepsilon)},\end{aligned}$$

and

$$\begin{aligned}\sum_{i\in coPa(j)}|\hat{\theta}_{ij}-\theta_{ij}| &= \sum_{i\in coPa(j)}\left|\frac{\hat{\rho}_{ik_i}\hat{\rho}_{jk_i}}{\hat{\omega}_{k_ik_i}}-\frac{\rho_{lk_i}\rho_{jk_i}}{\omega_{k_ik_i}}\right| \\ &\leq \sum_{i\in coPa(j)}\frac{1}{\hat{\omega}_{k_ik_i}}|\hat{\rho}_{ik_i}\hat{\rho}_{jk_i}-\rho_{ik_i}\rho_{jk_i}| + |\rho_{ik_i}\rho_{jk_i}|\left|\frac{1}{\hat{\omega}_{k_ik_i}}-\frac{1}{\omega_{k_ik_i}}\right| \\ &\leq \sum_{i\in coPa(j)}\frac{(2\rho_{\max}+\varepsilon)\varepsilon\omega_{k_ik_i}+2\rho_{\max}^2d_{k_i}^{in}\varepsilon}{(\omega_{k_ik_i}-2d_{k_i}^{in}\varepsilon)\omega_{k_ik_i}},\end{aligned}$$

where k_i denotes the v-node such that $j \to k_i \leftarrow i$ is a v-structure in C_T.

$$\sum_{i\neq j}|\hat{\theta}_{ij}-\theta_{ij}| \leq 2\sum_{j\in V_d}\frac{(\omega_{jj}+2\rho_{\max}d_j^{in})d_j^{in}\varepsilon}{(\omega_{jj}-2d_j^{in}\varepsilon)\omega_{jj}}+\sum_{j\in V_m}\sum_{i-j\in C_T}\frac{3\varepsilon}{(1-\rho_{ij}^2)(1-\rho_{ij}^2-2\varepsilon)}$$
$$+\sum_{j=1}^{p}\sum_{i\in coPa(j)}\frac{(2\rho_{\max}+\varepsilon)\varepsilon\omega_{k_ik_i}+2\rho_{\max}^2 d_{k_i}^{in}\varepsilon}{(\omega_{k_ik_i}-2d_{k_i}^{in}\varepsilon)\omega_{k_ik_i}}$$
$$\leq \frac{\varepsilon\left[\sum_{j\in V_d}(2d_j^{in}+4(d_j^{in})^2)+\sum_{j\in V_m}\sum_{i-j\in C_T}3+\sum_{j=1}^{p}\sum_{i\in coPa(j)}(2+\varepsilon+2d_{k_i}^{in})\right]}{(\omega_{\min}-2d_*\varepsilon)\omega_{\min}}.$$

The facts $\sum_{j\in V_d} d_j^{in}\leq p-1$ and $0\leq d_j^{in}\leq d_*$ imply that $\sum_{j\in V_d}(d_j^{in})^2\leq pd_*$. It is also obvious that $2\sum_{j\in V_d} d_j^{in}+\sum_{j\in V_m}\sum_{i-j\in C_T}1=2(p-1)$. Moreover, by counting the number of v-structures, there holds

$$\sum_{j=1}^{p}\sum_{i\in coPa(j)}d_{k_i}^{in}=2\sum_{k\in V_d:d_k^{in}\geq 2}d_k^{in}\binom{d_k^{in}}{2}\leq pd_*^2$$

and

$$\sum_{j=1}^{p}\sum_{i\in coPa(j)}1=2\sum_{k\in V_d:d_k^{in}\geq 2}\binom{d_k^{in}}{2}\leq pd_*.$$

Putting the above together, we have

$$\sum_{i\neq j}|\hat{\theta}_{ij}-\theta_{ij}|\leq\frac{(7+6d_*+2d_*^2)p\varepsilon}{(\omega_{\min}-2d_*\varepsilon)\omega_{\min}}.$$

Let's move on to the diagonal entries of the inverse correlation matrix. For each $j\in V_d$,

$$|\hat{\theta}_{jj}-\theta_{jj}|=\left|\frac{1}{\hat{\omega}_{jj}}-\frac{1}{\omega_{jj}}+\sum_{j\to k\in C_T}\left(\frac{\hat{\rho}_{jk}^2}{\hat{\omega}_{kk}}-\frac{\rho_{jk}^2}{\omega_{kk}}\right)\right|$$
$$\leq\left|\frac{1}{\hat{\omega}_{jj}}-\frac{1}{\omega_{jj}}\right|+\sum_{j\to k\in C_T}\left[\frac{|\hat{\rho}_{jk}^2-\rho_{jk}^2|}{\hat{\omega}_{kk}}+\rho_{jk}^2\left|\frac{1}{\hat{\omega}_{kk}}-\frac{1}{\omega_{kk}}\right|\right]$$
$$\leq\frac{2d_j^{in}\varepsilon}{(\omega_{jj}-2d_j^{in}\varepsilon)\omega_{jj}}+\sum_{j\to k\in C_T}\left[\frac{2\varepsilon}{\omega_{kk}-2d_k^{in}\varepsilon}+\rho_{\max}^2\frac{2d_k^{in}\varepsilon}{(\omega_{kk}-2d_k^{in}\varepsilon)\omega_{kk}}\right]$$
$$\leq\frac{2d_j^{in}\varepsilon}{(\omega_{jj}-2d_j^{in}\varepsilon)\omega_{jj}}+\sum_{j\to k\in C_T}\frac{2\varepsilon(\omega_{kk}+\rho_{\max}^2 d_k^{in})}{(\omega_{kk}-2d_k^{in}\varepsilon)\omega_{kk}}.$$

Similarly, for each $j \in V_m$,

$$
\begin{aligned}
|\hat{\theta}_{jj} - \theta_{jj}| &= \left| \sum_{i-j \in C_T} \left(\frac{\hat{\rho}_{ij}^2}{1-\hat{\rho}_{ij}^2} - \frac{\rho_{ij}^2}{1-\rho_{ij}^2} \right) + \sum_{j \to k \in C_T} \left(\frac{\hat{\rho}_{jk}^2}{\hat{\omega}_{kk}} - \frac{\rho_{jk}^2}{\omega_{kk}} \right) \right| \\
&\leq \sum_{i-j \in C_T} \left[\frac{|\hat{\rho}_{ij}^2 - \rho_{ij}^2|}{1-\hat{\rho}_{ij}^2} + \rho_{ij}^2 \left| \frac{1}{1-\hat{\rho}_{ij}^2} - \frac{1}{1-\rho_{ij}^2} \right| \right] \\
&\quad + \sum_{j \to k \in C_T} \left[\frac{|\hat{\rho}_{jk}^2 - \rho_{jk}^2|}{\hat{\omega}_{kk}} + \rho_{jk}^2 \left| \frac{1}{\hat{\omega}_{kk}} - \frac{1}{\omega_{kk}} \right| \right] \\
&\leq \sum_{i-j \in C_T} \frac{2\varepsilon}{(1-\rho_{ij}^2)(1-\rho_{ij}^2-2\varepsilon)} + \sum_{j \to k \in C_T} \frac{2\varepsilon(\omega_{kk} + \rho_{\max}^2 d_k^{in})}{(\omega_{kk} - 2d_k^{in}\varepsilon)\omega_{kk}}.
\end{aligned}
$$

Combine the above together,

$$
\begin{aligned}
\sum_{j=1}^{p} |\hat{\theta}_{jj} - \theta_{jj}| &\leq \sum_{j \in V_d} \frac{2d_j^{in}\varepsilon}{(\omega_{jj} - 2d_j^{in}\varepsilon)\omega_{jj}} + \sum_{j \in V_m} \sum_{i-j \in C_T} \frac{2\varepsilon}{(1-\rho_{ij}^2)(1-\rho_{ij}^2-2\varepsilon)} \\
&\quad + \sum_{j=1}^{p} \sum_{j \to k \in C_T} \frac{2\varepsilon(\omega_{kk} + \rho_{\max}^2 d_k^{in})}{(\omega_{kk} - 2d_k^{in}\varepsilon)\omega_{kk}} \\
&\leq \frac{2\varepsilon}{(\omega_{\min} - 2d_*\varepsilon)\omega_{\min}} \left(2\sum_{j \in V_d} d_j^{in} + \sum_{j \in V_m} \sum_{i-j \in C_T} 1 + \sum_{j=1}^{p} \sum_{j \to k \in C_T} d_k^{in} \right).
\end{aligned}
$$

Similar to the case of off-diagonal entries, we have

$$
\sum_{j=1}^{p} |\hat{\theta}_{jj} - \theta_{jj}| \leq \frac{2(2+d_*)p\varepsilon}{(\omega_{\min} - 2d_*\varepsilon)\omega_{\min}}.
$$

□

3.7. Discussions

This chapter studies the problem of polytree learning, a special case of DAG learning where the skeleton of the directed graph is a tree. This model has been widely used in the literature for both prediction and structure learning. We consider the linear polytree model, and consider the Chow-Liu algorithm (Chow and Liu, 1968) that has been proposed in Rebane and Pearl (1987) for polytree learning. Our major contribution in this theoretical work is to study the sample size conditions under which the polytree learning algorithm recovers the skeleton and the CPDAG

exactly. Under certain mild assumptions on the correlation coefficients over the polytree skeleton, we show that the skeleton can be exactly recovered with probability at least $1-\delta$ if the sample size satisfies $n > O(\frac{1}{\rho_{\min}^2}\log\frac{p}{\sqrt{\delta}})$, and the CPDAG of the polytree can be exactly recovered with probability at least $1-\delta$ if the sample size satisfies $n > O(\frac{1}{\rho_{\min}^4}\log\frac{p}{\sqrt{\delta}})$. We also establish necessary conditions on sample size for both skeleton and CPDAG recovery, which are consistent with the sufficient conditions and thereby give sharp characterization of the difficulties for these two tasks. In addition, under the event of exact recovery of CPDAG, we also establish the accuracy of inverse correlation matrix estimation. Under the component-wise ℓ_1 metric, we give an estimation error bound that relies on the estimation error of pairwise correlations, the minimum noise variance $\omega_{\min}$, and the maximum in-degree d_*.

There are a number of remaining questions to study in future. It would be interesting to study how to relax the polytree assumption. In fact, the benchmark data analysis (Section 3.5.2) is very insightful, since it shows that the considered Chow-Liu based CPDAG recovery algorithm, which seemingly relies heavily on the polytree assumption, could lead to reasonable and accurate structure learning result when the ground truth deviates from a polytree to some degree. This inspires us to consider the robustness of the proposed approach against such structural assumptions. For example, if the ground truth can only be approximated by a polytree, can the structure learning method described in Sections 3.2.3.1 and 3.2.3.2 lead to an approximate recovery of the ground truth CPDAG with theoretical guarantees?

When the ground truth DAG is a polytree, our result Theorem 3.4.5 regarding inverse correlation matrix estimation relies on the assumption that the ground truth of the CPDAG must be exactly recovered. Naturally, we wonder whether this is necessary. In other words, if the sample size is not large enough and the CPDAG is thereby unable to be recovered exactly, can we still obtain an accurate estimate of the inverse correlation matrix? Which method should be used for such estimate?

As aforementioned, polytree modeling is usually used in practice only as initialization, and post-processing could give better structural recovery result. A well-known method of this type is given in Cheng et al. (2002) without theoretical guarantees. An interesting future research direction is to

include such post processing steps into our theoretical analysis, such that our structural learning results (e.g., Theorems 3.3.6) hold for more general sparse DAGs.

APPENDIX A

Appendix for Chapter 2

A.1. ℓ_2 Perturbation Theory for Fiedler Vector Under Unrestricted Heterogeneity Within Mega-Communities

THEOREM A.1.1. *Let $(\mathcal{G}_0, \mathcal{G}_1)$ denote a partition of $\{1, \ldots, n\}$, with $n_0 = |\mathcal{G}_0|$ and $n_1 = |\mathcal{G}_1|$, and $\boldsymbol{A} \in \{0,1\}^{n\times n}$ the adjacency matrix such that*

$$\mathbb{E}[\boldsymbol{A}_{ij}] \begin{cases} = p_\emptyset & (i \in \mathcal{G}_0, j \in \mathcal{G}_1) \\ \geq p_0 & (i, j \in \mathcal{G}_0) \\ \geq p_1 & (i, j \in \mathcal{G}_1) \end{cases},$$

where $p_\emptyset < \min(p_0, p_1)$. Further let $\boldsymbol{u}_{n-1}$ and $\boldsymbol{u}^*_{n-1}$ denote the Fiedler vector of $\boldsymbol{A}$ and $\mathbb{E}[\boldsymbol{A}]$, respectively. Then, for any fixed $r > 0$, there exists a constant $B_{\ell_2}(r)$ that only depends on r such that, with probability at least $1 - 2n^{-r}$,

$$\|\boldsymbol{u}_{n-1}\text{sign}(\boldsymbol{u}_{n-1}^T\boldsymbol{u}^*_{n-1}) - \boldsymbol{u}^*_{n-1}\|_2 \leq B_{\ell_2}(r)\frac{\sqrt{(np_\emptyset + \log n)\log n}}{\min\{n_0(p_0 - p_\emptyset), n_1(p_1 - p_\emptyset)\}}.$$

PROOF. Note that the proofs for Lemma 2.3.3 and Theorem 2.3.2 do not rely on the modelling assumptions within the mega-communities, except that $\boldsymbol{P}_{ij} \geq p_0$ if $i, j \in \mathcal{G}_0$ and $\boldsymbol{P}_{ij} \geq p_1$ if $i, j \in \mathcal{G}_1$. Therefore, the proofs carry over to Theorem A.1.1. □

A.2. Eigen-structure for Two-layer Hierarchical SBMs

In Theorem 2.2.1, we give a full description of the eigenstructure for the population unnormalized graph Laplacian under general BTSBMs. The demonstration of strong consistency, which is shown by Theorem 2.3.1, heavily depends on this population eigenstructure. In this section, we reveal that the results in Theorem 2.2.1 hold under a more general SBM without hierarchical structure within mega-communities, so as Theorem 2.3.1.

Recall that in the inhomogeneous stochastic block model introduced in Remark 5, there are two mega-communities $\mathcal{G}_0$ and $\mathcal{G}_1$. In this section, we assume that each mega-community is a SBM itself. More specifically, let's assume that the subgraph $\mathcal{G}_0$ follows a SBM with K_0 communities, and the subgraph $\mathcal{G}_1$ follows a SBM with K_1 communities. These $K \coloneqq K_0 + K_1$ communities are referred to as the primitive communities, whose sizes are denoted as $l_1, \ldots, l_K$, respectively. As a result, we have $n_0 = l_1 + \cdots + l_{K_0}$ and $n_1 = l_{K_0+1} + \cdots + l_K$. To combine a standard SBM with the aforementioned inhomogeneous block model together, the $K \times K$ connection probability matrix is assumed to satisfy

$$B_{st} \begin{cases} = q & (1 \le s \le K_0,\ K_0 + 1 \le t \le K) \\ \ge p_0 & (1 \le s,\ t \le K_0) \\ \ge p_1 & (K_0 + 1 \le s,\ t \le K) \end{cases} . \tag{A.1}$$

We refer to such a hierarchical SBM as *Two-layer Hierarchical SBM*, although it includes a general class of hierarchical models, with the general BTSBM as a special case.

As with before, here we also need the following notations in establishing relevant theory.

$$p^* = \max_{1 \le i,j \le n} P_{ij}, \quad \bar{p}^* = \max_{1 \le i \le n} \frac{1}{n} \sum_{j=1}^{n} P_{ij}, \quad \underline{p}_* = \min_{1 \le i \le n} \frac{1}{n} \sum_{j=1}^{n} P_{ij}.$$

Noting $\boldsymbol{P} = \mathbb{E}[\boldsymbol{A}]$, p^* is the largest connection probability across the whole network, and $n\bar{p}^*$ ($n\underline{p}_*$) is the largest (smallest) expected degree. Obviously, $p^* \ge \bar{p}^* \ge \underline{p}_*$.

As an analogy to Theorem 2.2.1, we have the following results under the more general two-layer hierarchical SBM. Also note that the first two items have also been proven in Balakrishnan et al. (2011).

THEOREM A.2.1. *Under the K-group two-layer hierarchical SBM (A.1) with parameters q, p_0, p_1 and community sizes $l_1, \ldots, l_K$, the eigenvalues and eigenvectors of the population graph Laplacian $\boldsymbol{L}^*$ satisfy*

(1) $\lambda^*_{n-1} = nq$ with multiplicity 1 and the entries of the corresponding eigenvector $\boldsymbol{u}^*_{n-1}$ obeys

$$\boldsymbol{u}^*_{n-1,i} = \pm \begin{cases} \sqrt{n_1/(n_0 n)} & (i \in \mathcal{G}_0) \\ -\sqrt{n_0/(n_1 n)} & (i \in \mathcal{G}_1) \end{cases} ;$$

(2) $\lambda^*_{n-2} \geq \min\{n_1p_1 + n_0q, n_0p_0 + n_1q\}$;

(3) $\lambda^*_{n-K} \geq n\underline{p}_*$;

(4) For any j *with* $\lambda^*_j < n\underline{p}_*$, $\|\boldsymbol{u}^*_j\|_\infty \leq \max_{1\leq s\leq K}\sqrt{1/l_s}$.

PROOF OF THEOREM A.2.1. (1) It is easy to verify that $\boldsymbol{u}^*_{n-1}$ defined above is an eigenvector of $\boldsymbol{L}^*$ with eigenvalue nq. We will show that nq is the second smallest eigenvalue with multiplicity 1.

Suppose $\boldsymbol{P}^{'}$ is an $n \times n$ matrix such that

$$P^{'}_{ij} = \begin{cases} p_0 & (i \in \mathcal{G}_0, j \in \mathcal{G}_0) \\ p_1 & (i \in \mathcal{G}_1, j \in \mathcal{G}_1) \\ q & (\text{otherwise}) \end{cases}.$$

Let $\boldsymbol{L}^{'}$ denote the corresponding unnormalized graph Laplacian. It can be verified through simple algebra that $\lambda_{n-1}(\boldsymbol{L}^{'}) = nq$ and $\lambda_{n-2}(\boldsymbol{L}^{'}) = \min\{n_1p_1 + n_0q, n_0p_0 + n_1q\}$. Note that $\boldsymbol{L}^* - \boldsymbol{L}^{'}$ is also a Laplacian matrix, which means $\lambda_n(\boldsymbol{L}^* - \boldsymbol{L}^{'}) = 0$. By Wely's Inequality, $\lambda^*_{n-1} \geq nq$ and $\lambda^*_{n-2} \geq \min\{n_1p_1 + n_0q, n_0p_0 + n_1q\} > nq$. Therefore, nq is the second smallest eigenvalue of $\boldsymbol{L}^*$ with multiplicity 1.

(2) See the proof of (1).

(3) Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ represent node sets for the K primitive communities. Without loss of generality, we assume the nodes have been properly ordered such that $\mathcal{C}_1 = \{1, \ldots, l_1\}$, $\mathcal{C}_2 = \{l_1 + 1, \ldots, l_1 + l_2\}$ and so on. For any $i \in \mathcal{C}_s, s = 1, \ldots, K$, the expected degree of i is $d^*_s = \sum_{t=1}^K l_t B_{st}$. Considering the sub-matrix of $\boldsymbol{L}^*$ at $\mathcal{C}_s$ rows and $\mathcal{C}_s$ columns, denoted by $\boldsymbol{L}^*_{\mathcal{C}_s\times\mathcal{C}_s}$, it is easy to verify that $\mathbf{1}_{l_s}$ is an eigenvector of $\boldsymbol{L}^*_{\mathcal{C}_s\times\mathcal{C}_s}$, and the other $l_s - 1$ eigenvectors which are orthogonal to $\mathbf{1}_{l_s}$ have the same eigenvalue d^*_s. For any $\boldsymbol{v} \in \mathbb{R}^{l_s}$ in the eigenspace of d^*_s, extending $\boldsymbol{v}$ into $\mathbb{R}^n$ by filling all additional coordinates with 0, $\boldsymbol{v}$ becomes an eigenvector of $\boldsymbol{L}^*$ with eigenvalue d^*_s. Hence d^*_s is an eigenvalue of $\boldsymbol{L}^*$ with multiplicity at least $l_s - 1$, $s = 1, \ldots, K$. So λ^*_{n-K} is at least $\min_s\{d^*_s\} = n\underline{p}_*$.

(4) Continue with the proof of (3). We have specified $n-K$ eigenvectors of $\boldsymbol{L}^*$. Now let us examine the other K eigenvectors in a constructive way.

First, if we write the population Laplacian as $\boldsymbol{L}^* = \boldsymbol{D}^* - \boldsymbol{P}$, then the diagonal entries of $\boldsymbol{D}^*$ are d_s^* for $s = 1, \ldots, K$ with multiplicities $l_1, \ldots, l_K$. Suppose $\boldsymbol{v} = [w_1 \mathbf{1}_{l_1}^\top, \ldots, w_K \mathbf{1}_{l_K}^\top]^\top$ be an eigenvector of $\boldsymbol{L}^*$, i.e. $\boldsymbol{L}^* \boldsymbol{v} = \lambda \boldsymbol{v}$ for some $\lambda \in \mathbb{R}$. Then for each $s = 1, \ldots, K$, there holds

$$d_s^* w_s - \sum_{t=1}^{K} l_t B_{st} w_t = \lambda w_s. \tag{A.2}$$

Define the diagonal matrix $\widetilde{\boldsymbol{D}} \in \mathbb{R}^{K \times K}$ whose diagonal entries are $\tilde{d}_s = d_s^*/n$ for $s = 1, \ldots, K$. Define another diagonal matrix $\widetilde{\boldsymbol{D}}^l \in \mathbb{R}^{K \times K}$ whose diagonal entries are $\tilde{d}_s^l = l_s/n$. Also define $\widetilde{\boldsymbol{L}} = \widetilde{\boldsymbol{D}} - \boldsymbol{B}\widetilde{\boldsymbol{D}}^l$. Then Equation (A.2) becomes $\widetilde{\boldsymbol{L}}\boldsymbol{w} = \frac{\lambda}{n}\boldsymbol{w}$. In other words, $\boldsymbol{w} = [w_1, \ldots, w_K]^\top$ is an eigenvector of $\widetilde{\boldsymbol{L}}$ with corresponding eigenvalue λ/n. Note that $\widetilde{\boldsymbol{L}}$ is diagonalizable since

$$(\widetilde{\boldsymbol{D}}^l)^{\frac{1}{2}} \widetilde{\boldsymbol{L}} (\widetilde{\boldsymbol{D}}^l)^{-\frac{1}{2}} = (\widetilde{\boldsymbol{D}}^l)^{\frac{1}{2}} \widetilde{\boldsymbol{D}} (\widetilde{\boldsymbol{D}}^l)^{-\frac{1}{2}} - (\widetilde{\boldsymbol{D}}^l)^{\frac{1}{2}} \boldsymbol{B} (\widetilde{\boldsymbol{D}}^l)^{\frac{1}{2}}$$

is symmetric and thereby diagonalizable, which means $\widetilde{\boldsymbol{L}}$ has K eigenvectors. Based on each of them, we can construct an eigenvector of $\boldsymbol{L}^*$ by repeating the s-th element l_s times, i.e. $\boldsymbol{v} = [w_1 \mathbf{1}_{l_1}^\top, \ldots, w_K \mathbf{1}_{l_K}^\top]^\top$. Obviously, eigenvectors in such form are orthogonal to the $n - K$ eigenvectors specified in the proof of (3) and span the orthogonal complement. Thus for any $\lambda_j^* < n\underline{p}_*$, the corresponding $\boldsymbol{u}_j^*$ must take such form and satisfy $\|\boldsymbol{u}_j^*\|_\infty \leq \max_s \sqrt{1/l_s}$.

□

With an argument similar to the proof of Theorem 2.3.1, we can obtain the following result:

THEOREM A.2.2 (ℓ_∞ perturbation). *Under the K-group two-layer hierarchical SBM (A.1) with parameters q, p_0, p_1 and community sizes $l_1, \ldots, l_K$, assume further that $\xi = \max_s \frac{n}{l_s}$ is a constant. Then, for any fixed constant $r > 0$, there exists a constant C_{ℓ_∞} that only depends on r and ξ, such that*

$$\sqrt{n}\|\boldsymbol{u}_{n-1}\mathrm{sign}(\boldsymbol{u}_{n-1}^T \boldsymbol{u}_{n-1}^*) - \boldsymbol{u}_{n-1}^*\|_\infty < \min\{\sqrt{n_0/n_1}, \sqrt{n_1/n_0}\}$$

with probability at least $1-(10K+4)n^{-r}$*, provided the following two conditions:*

$$\begin{aligned} &\textit{Density gap} && \min\{n_0(p_0-q), n_1(p_1-q)\} \geq C_{\ell_\infty}\sqrt{(n_0p_0+n_1p_1)\log n}, \\ &\textit{Degree variation} && (n(\underline{p}_*-q))^4 \geq C_{\ell_\infty}(n\bar{p}^*)^3\log n, \end{aligned}$$

where $\underline{p}_*$ *and* $\bar{p}^*$ *are defined in* (2.1).

A.3. $\ell_{2\to\infty}$ Perturbation Theory for Unnormalized Laplacians

A.3.0.1. *A Generic* $\ell_{2\to\infty}$ *Perturbation Bound.* In this subsection, we rephrase the weaker version Theorem 2.6 of Lei (2019), discussed in their Remark 2.3, by only keeping the parts that are relevant to our purpose. Throughout this section, we consider two generic symmetric real matrices $\boldsymbol{G}$ and $\boldsymbol{G}^*$ with

$$\boldsymbol{E} = \boldsymbol{G} - \boldsymbol{G}^*. \tag{A.3}$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_n^*$ be the eigenvalues of $\boldsymbol{G}$ and $\boldsymbol{G}^*$, respectively. Given positive integers s and r, let

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{s+1}, \lambda_{s+2}, \ldots, \lambda_{s+r}), \quad \boldsymbol{\Lambda}^* = \mathrm{diag}(\lambda_{s+1}^*, \lambda_{s+2}^*, \ldots, \lambda_{s+r}^*). \tag{A.4}$$

Let $\boldsymbol{U}, \boldsymbol{U}^* \in \mathbb{R}^{n\times r}$ be a matrix of eigenvectors such that

$$\boldsymbol{G}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}, \quad \boldsymbol{G}^*\boldsymbol{U}^* = \boldsymbol{U}^*\boldsymbol{\Lambda}^*. \tag{A.5}$$

To state the generic bound, we define the following quantities.

- *Modified perturbation matrix* $\tilde{\boldsymbol{E}}$:

$$\tilde{\boldsymbol{E}} = \boldsymbol{G} - \boldsymbol{\Sigma} - (\boldsymbol{G}^* - \boldsymbol{\Sigma}^*)$$

  where

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{G}), \quad \boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{G}^*).$$

- *Condition number* κ^*:

$$\lambda_{\max}^* = \lambda_{\max}(\boldsymbol{\Lambda}^*), \quad \lambda_{\min}^* = \lambda_{\min}(\boldsymbol{\Lambda}^*), \quad \kappa^* = \lambda_{\max}^*/\lambda_{\min}^*. \tag{A.6}$$

- *Effective eigengap* Δ^*:

$$\Delta^* \triangleq \min\{\text{sep}^*, \lambda^*_{\min}\}, \tag{A.7}$$

where $\text{sep}^* = \min\{\lambda^*_s - \lambda^*_{s+1}, \lambda^*_{s+r} - \lambda^*_{s+r+1}\}$ and $\lambda^*_0 = \infty, \lambda^*_{n+1} = -\infty$.

The assumptions for the generic bound are stated below.

A1 For any $\delta \in (0,1)$,

$$\frac{\min_{j\in[s+1,s+r]} |\Lambda^*_{jj}|}{\min_{j\in[s+1,s+r],k\in[n]} |\Lambda^*_{jj} - \Sigma_{kk}|} \le \Theta(\delta),$$

with probability at least $1-\delta$ for some deterministic function $\Theta(\delta) > 0$.

A2 For any $\delta \in (0,1)$, there exists a random matrix $\boldsymbol{G}^{(k)} \in \mathbb{R}^{n\times n}$ such that

$$d_{TV}\left(\mathbb{P}_{(\tilde{\boldsymbol{E}}_k, \boldsymbol{G}^{(k)})}, \mathbb{P}_{\tilde{\boldsymbol{E}}_k} \times \mathbb{P}_{\boldsymbol{G}^{(k)}}\right) \le \delta/n.$$

where d_{TV} denotes the total variation distance and it holds simultaneously for all k and all contiguous subsets $S \subset [r]$ that

$$\|\boldsymbol{G}^{(k)} - \boldsymbol{G}\|_{\text{op}} \le L_1(\delta), \quad \frac{\|(\boldsymbol{G}^{(k)} - \boldsymbol{G})\boldsymbol{U}\|_{\text{op}}}{\lambda^*_{\min}} \le (\kappa(\boldsymbol{\Lambda}^*)L_2(\delta) + L_3(\delta))\,\|\boldsymbol{U}\|_{2\to\infty},$$

with probability at least $1-\delta$ for some deterministic functions $L_1(\delta), L_2(\delta), L_3(\delta)$.

A3 There exists deterministic functions $\lambda_-(\delta), E_+(\delta), \tilde{E}_\infty(\delta)$, such that for any $\delta \in (0,1)$, the following event holds with probability at least $1-\delta$:

$$\|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^*\|_{\max} \le \lambda_-(\delta), \quad \|\boldsymbol{E}\boldsymbol{U}^*\|_{\text{op}} \le E_+(\delta), \quad \|\tilde{\boldsymbol{E}}\|_{2\to\infty} \le \tilde{E}_\infty(\delta).$$

A4 There exists deterministic functions $\tilde{b}_\infty(\delta), \tilde{b}_2(\delta) > 0$, such that for any $\delta \in (0,1)$, $k \in [n]$, and fixed matrix $\boldsymbol{W} \in \mathbb{R}^{n\times \tilde{j}}$,

$$\|\tilde{\boldsymbol{E}}_k^T \boldsymbol{W}\|_2 \le \tilde{b}_\infty(\delta)\|\boldsymbol{W}\|_{2\to\infty} + \tilde{b}_2(\delta)\|\boldsymbol{W}\|_{\text{op}}, \text{ with probability at least } 1-\delta/n.$$

A5 $\Delta^* \ge 4\left(\Theta(\delta)\tilde{\sigma}(\delta) + L_1(\delta) + \lambda_-(\delta) + E_+(\delta)\right)$ where

$$\tilde{\eta}(\delta) = \tilde{E}_\infty(\delta) + \tilde{b}_\infty(\delta) + \tilde{b}_2(\delta), \quad \tilde{\sigma}(\delta) = \{\kappa^* L_2(\delta) + L_3(\delta) + 1\}\tilde{\eta}(\delta) + E_+(\delta). \tag{A.8}$$

THEOREM A.3.1 (Theorem 2.6 of Lei (2019), Remark 2.3). *Under assumptions **A1** - **A5**,*

$$\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*)-\boldsymbol{U}^*\|_{2\to\infty}\le C\Big\{\frac{\Theta(\delta)}{\lambda^*_{\min}}\|\boldsymbol{E}\boldsymbol{U}^*\|_{2\to\infty}+\left(\frac{E_+(\delta)^2}{(\Delta^*)^2}+\frac{\Theta(\delta)\tilde{\sigma}(\delta)}{\Delta^*}\right)\|\boldsymbol{U}^*\|_{2\to\infty}$$
$$+\frac{\Theta(\delta)(\tilde{b}_2(\delta)+\|\boldsymbol{G}^*-\boldsymbol{\Sigma}^*\|_{2\to\infty})E_+(\delta)}{\lambda^*_{\min}\Delta^*}\Big\},$$

with probability at least $1-B(r)\delta$, where C is a universal constant (that can be chosen as 136) and

$$B(r)=10\min\{r,1+\log_2\kappa^*\}. \tag{A.9}$$

A.3.0.2. *Proof of Lemma 2.5.5.* Let $\boldsymbol{G}=\boldsymbol{L}+\nu\boldsymbol{J}$ and $\boldsymbol{G}^*=\boldsymbol{L}^*+\nu\boldsymbol{J}$. Now we verify each of assumptions **A**1 - **A**5. We add the subscript ν into all quantities defined in Supplement A.3.0.1 to highlight their dependence on ν, including $\Lambda_\nu,\Sigma_\nu,\Theta_\nu$ (or $\Lambda^*_\nu,\Sigma^*_\nu,\Theta^*_\nu$), and $\boldsymbol{E}_\nu$, $\tilde{\boldsymbol{E}}_\nu$, $L_{1,\nu}$, $L_{2,\nu}$, $L_{3,\nu}$, $\tilde{E}_{\infty,\nu}$, $E_{+,\nu}$, $\lambda_{-,\nu}$, $\tilde{b}_{\infty,\nu}$, $\tilde{b}_{2,\nu}$, κ^*_ν, Δ^*_ν, $\tilde{\eta}_\nu$, $\tilde{\sigma}_\nu$. We remove the subscript ν when $\nu=0$. Moreover, we let

$$M(\delta)=\sqrt{n\bar{p}^*\log(n/\delta)}+\log(n/\delta),\quad R(\delta)=\log(n/\delta)+\tilde{j}. \tag{A.10}$$

By definition (2.19), $\tilde{j}\le\tilde{K}\le K=O(1)$. In each of the following steps, δ is always set to be n^{-r}. Unless otherwise specified, $a\gtrsim b$ ($a\lesssim b$) iff $a\ge Cb$ ($a\le Cb$) for some constant C that only depends on r and ξ. To apply Theorem A.3.1 in Supplement A.3.0.1, we need to verify Assumptions **A**1 - **A**5.

Checking Assumption A1: We recall Lemma 3.12 of Lei (2019), rephrased for our purpose.

Lemma A.3.2. *Let $\Theta(\delta)$ be defined in assumption **A1** in Supplement A.3.0.1. Further let*

$$\Theta^*=\frac{\min_{j\in[n-\tilde{j},n-1]}|\Lambda^*_{jj}|}{\min_{j\in[n-\tilde{j},n-1],k\in[n]}|\Lambda^*_{jj}-\Sigma^*_{kk}|}.$$

Then $\Theta(\delta)\le 5\Theta^*$ if

$$\min_{j\in[n-\tilde{j},n-1],k\in[n]}|\Lambda^*_{jj}-\Sigma^*_{kk}|\ge 5M(\delta).$$

In this case, $\Lambda^*_{jj} < n\underline{p}_*$ for all $j \in [n-\tilde{j}, n-1]$ and $\Sigma^*_{kk} \geq n\underline{p}_*$ for all $k \in [n]$. Thus,

$$\min_{j\in[n-\tilde{j},n-1],k\in[n]} |\Lambda^*_{jj} - \Sigma^*_{kk}| = \min_{k\in[n]} \Sigma^*_{kk} - \max_{j\in[n-\tilde{j},n-1]} \Lambda^*_{jj} = n\underline{p}_* - \lambda^*_{n-\tilde{j}}.$$

By definition of $\tilde{j}$ and equation (2.20),

$$n\underline{p}_* - \lambda^*_{n-\tilde{j}} \geq \frac{n(\underline{p}_* - p_\emptyset)}{K}.$$

By definition,

$$\Lambda^*_{\nu,jj} = \Lambda^*_{jj} + \nu, \quad \Sigma^*_{\nu,kk} = \Sigma^*_{kk} + \frac{n-1}{n}\nu.$$

Thus,

$$\min_{j\in[n-\tilde{j},n-1],k\in[n]} |\Lambda^*_{\nu,jj} - L^*_{\nu,kk}| \geq \frac{n(\underline{p}_* - p_\emptyset)}{K} - \frac{\nu}{n}.$$

By the condition (2.6) and (2.24),

$$\frac{n(\underline{p}_* - p_\emptyset)}{K} \geq \frac{C_{\ell_\infty}^{1/4}(n\bar{p}^*)^{3/4}(\log n)^{1/4}}{K} \geq \frac{C_{\ell_\infty}}{K}\log n.$$

If $C_{\ell_\infty} \geq 3\xi \geq 3K$,

$$\frac{n(\underline{p}_* - p_\emptyset)}{K} \geq 2 \geq 2\bar{p}^*.$$

By the definition (2.22) of ν,

$$\frac{\nu}{n} = \bar{p}^* \leq \frac{n(\underline{p}_* - p_\emptyset)}{2K}.$$

As a result,

$$\min_{j\in[n-\tilde{j},n-1],k\in[n]} |\Lambda^*_{\nu,jj} - L^*_{\nu,kk}| \geq \frac{n(\underline{p}_* - p_\emptyset)}{2K}.$$

On the other hand, by (2.24),

$$M(n^{-r}) \leq (r+1)\left(1 + C_{\ell_\infty}^{-1}\right)\sqrt{n\bar{p}^*\log n}.$$

By the condition (2.6) and (2.24) again, if $C_{\ell_\infty}^{1/2} \geq 20\xi(r+1) \geq 20K(r+1)$,

$$\frac{n(\underline{p}_* - p_\emptyset)}{2K} \geq \frac{C_{\ell_\infty}^{1/4}(n\bar{p}^*)^{3/4}(\log n)^{1/4}}{2K} \geq \frac{C_{\ell_\infty}^{1/2}\sqrt{(n\bar{p}^*)\log n}}{2K} \geq 5M(n^{-r}).$$

Therefore,

$$\Theta_\nu(n^{-r}) \le 5\Theta^*_\nu \le \frac{5(np_\emptyset+\nu)}{n(\underline{p}_*-p_\emptyset)/2K} \lesssim \frac{n\bar{p}^*}{n(\underline{p}_*-p_\emptyset)}. \tag{A.11}$$

Checking Assumption A2: We recall Lemma 3.10 of Lei (2019).

Lemma A.3.3. *There exists $\boldsymbol{G}^{(1)},\dots,\boldsymbol{G}^{(n)}$ satisfying **A**2 for $\boldsymbol{G}=\boldsymbol{L}$ with*

$$L_1(\delta)\lesssim M(\delta),\quad L_2(\delta)=1,\quad L_3(\delta)\lesssim \frac{n\bar{p}^*+\log(n/\delta)}{\lambda_{\min}(\boldsymbol{\Lambda}^*)},$$

*where $\lesssim$ only hides absolute constants and $L_1(\delta),L_2(\delta),L_3(\delta)$ are defined in Assumption **A**2 in Supplement A.3.0.1.*

In this case, let

$$\boldsymbol{G}^{(k)}_\nu = \boldsymbol{G}^{(k)}+\nu\boldsymbol{J}.$$

Then it is easy to see that

$$\boldsymbol{G}^{(k)}_\nu - \boldsymbol{L}_\nu = \boldsymbol{G}^{(k)}-\boldsymbol{L}.$$

Therefore, Lemma A.3.3 holds for any $\nu>0$. Let $\delta=n^{-r}$, we have

$$L_{1,\nu}(n^{-r})\lesssim\sqrt{n\bar{p}^*\log n},\quad L_{2,\nu}(n^{-r})\lesssim 1,\quad L_{3,\nu}(n^{-r})\lesssim\frac{n\bar{p}^*+\log n}{np_\emptyset+\nu}\lesssim\frac{n\bar{p}^*+\log n}{n\bar{p}^*}\lesssim 1, \tag{A.12}$$

where the last inequality uses (2.24).

Checking Assumption A3: We recall Lemma 3.8 of Lei (2019).

Lemma A.3.4. *Assumption **A**3 is satisfied for $\boldsymbol{G}=\boldsymbol{L}$ with*

$$\tilde{E}_\infty(\delta)\lesssim\sqrt{n\bar{p}^*}+\sqrt{\log(n/\delta)},\quad E_+(\delta),\lambda_-(\delta)\lesssim M(\delta),$$

*where $\lesssim$ only hides absolute constants and $\tilde{E}_\infty(\delta),E_+(\delta),\lambda_-(\delta)$ are defined in Assumption **A**3 in Supplement A.3.0.1.*

Since $\mathbf{\Lambda}_\nu = \mathbf{\Lambda} + \nu \boldsymbol{I}$ and $\mathbf{\Lambda}^*_\nu = \mathbf{\Lambda}^* + \nu \boldsymbol{I}$, $\mathbf{\Lambda}_\nu - \mathbf{\Lambda}^*_\nu = \mathbf{\Lambda} - \mathbf{\Lambda}^*$ is invariant to ν. Similarly, $\boldsymbol{E}_\nu = \boldsymbol{E}$ and $\tilde{\boldsymbol{E}}_\nu = \tilde{\boldsymbol{E}}$. Thus Lemma A.3.4 holds for any $\nu > 0$. By (2.24),

$$\tilde{E}_{\infty,\nu}(n^{-r}) \lesssim \sqrt{n\bar{p}^*}, \quad E_{+,\nu}(n^{-r}), \lambda_{-,\nu}(n^{-r}) \lesssim \sqrt{n\bar{p}^* \log n}. \tag{A.13}$$

Checking Assumption A4: We recall Lemma 3.7 of Lei (2019).

Lemma A.3.5. *Assumption **A4** is satisfied for $\boldsymbol{G} = \boldsymbol{L}$ with*

$$\tilde{b}_\infty(\delta) \lesssim \frac{R(\delta)}{\alpha \log R(\delta)}, \quad \tilde{b}_2(\delta) \lesssim \frac{\sqrt{p^*} R(\delta)^{(1+\alpha)/2}}{\alpha \log R(\delta)},$$

*where $\lesssim$ only hides absolute constants and $\tilde{b}_\infty(\delta), \tilde{b}_2(\delta)$ are defined in Assumption **A4** in Supplement A.3.0.1.*

As with Assumption **A**3, $\tilde{E}$ is invariant to ν. Thus Lemma A.3.5 holds for any $\nu > 0$. Let $\alpha = 1/\log R(\delta)$. Since $\tilde{j} \le K = O(1)$,

$$\tilde{b}_{\infty,\nu}(n^{-r}) \lesssim R(n^{-r}) \lesssim \log n, \quad \tilde{b}_{2,\nu}(\delta) \lesssim \sqrt{R(\delta)p^*} \lesssim \sqrt{(\log n)p^*} \lesssim \sqrt{(\log n)\bar{p}^*}. \tag{A.14}$$

where the last inequality uses the fact that $\bar{p}^* \ge (\max n_s)p^*/n$.

Checking Assumption A5: We first refer the readers to Supplement A.3.0.1 for the definitions of $\kappa^*, \Delta^*, \tilde{\eta}(\delta)$ and $\tilde{\sigma}(\delta)$. By definition,

$$\kappa^*_\nu = \frac{\lambda_{\max}(\Lambda^*_\nu)}{\lambda_{\min}(\Lambda^*_\nu)} = \frac{\lambda^*_{n-\tilde{j}} + \nu}{np_\emptyset + \nu} \lesssim 1, \tag{A.15}$$

and

$$\Delta^*_\nu = \min\{\nu, np_* - \lambda_{n-\tilde{j}+1}\} \ge \min\left\{\nu, \frac{n(p_* - p_\emptyset)}{K}\right\} = \frac{n(p_* - p_\emptyset)}{K}. \tag{A.16}$$

By definition of $\tilde{\eta}$, (A.13) and (A.14),

$$\tilde{\eta}_\nu(n^{-r}) \lesssim \sqrt{n\bar{p}^*} + \log n.$$

By (A.12), (A.13) and (2.24),

$$\tilde{\sigma}_\nu(n^{-r}) \lesssim \tilde{\eta}(n^{-r}) + \sqrt{n\bar{p}^* \log n} \lesssim \sqrt{n\bar{p}^*} + \log n + \sqrt{n\bar{p}^* \log n} \lesssim \sqrt{n\bar{p}^* \log n}. \tag{A.17}$$

By (A.11), (A.12) and (A.13),

$$\begin{aligned}
&\Theta_\nu(n^{-r})\tilde{\sigma}_\nu(n^{-r}) + L_{1,\nu}(n^{-r}) + \lambda_{-,\nu}(n^{-r}) + E_{+,\nu}(n^{-r}) \\
\lesssim &\frac{n\bar{p}^*}{n(\underline{p}_* - p_\emptyset)}\sqrt{n\bar{p}^* \log n} + \sqrt{n\bar{p}^* \log n} \lesssim \frac{n\bar{p}^*}{n(\underline{p}_* - p_\emptyset)}\sqrt{n\bar{p}^* \log n}.
\end{aligned}$$

By (A.16) and the condition (2.6), if C_{ℓ_∞} is sufficiently large,

$$\Delta_\nu^* \geq 4\left(\Theta_\nu(n^{-r})\tilde{\sigma}_\nu(n^{-r}) + L_{1,\nu}(n^{-r}) + \lambda_{-,\nu}(n^{-r}) + E_{+,\nu}(n^{-r})\right), \tag{A.18}$$

Thus, Assumption **A**5 is satisfied.

Final Result:

In the previous five steps, we show that Assumption **A**1 - **A**5 are satisfied under the condition (2.6), if C_{ℓ_∞} is sufficiently large. By Theorem A.3.1, with probability $1 - B(\tilde{j})n^{-r}$,

$$\begin{aligned}
&\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*) - \boldsymbol{U}^*\|_{2\to\infty} \\
\lesssim &\frac{\Theta_\nu(n^{-r})}{\lambda_{\min}(\boldsymbol{\Lambda}_\nu^*)}\|\boldsymbol{E}_\nu\boldsymbol{U}^*\|_{2\to\infty} + \left(\frac{E_{+,\nu}^2(n^{-r})}{(\Delta_\nu^*)^2} + \frac{\Theta_\nu(n^{-r})\tilde{\sigma}_\nu(n^{-r})}{\Delta_\nu^*}\right)\|\boldsymbol{U}^*\|_{2\to\infty} \\
&+ \frac{\Theta_\nu(n^{-r})E_{+,\nu}(n^{-r})}{\Delta_\nu^*}\frac{\tilde{b}_{2,\nu}(n^{-r}) + \|\boldsymbol{L}_\nu^* - \boldsymbol{\Sigma}_\nu^*\|_{2\to\infty}}{\lambda_{\min}(\boldsymbol{\Lambda}_\nu^*)}.
\end{aligned} \tag{A.19}$$

To bound $\|\boldsymbol{E}_\nu\boldsymbol{U}^*\|_{2\to\infty}$, we recall Lemma 3.9 of Lei (2019).

Lemma A.3.6. *Let $M(\delta)$ and $R(\delta)$ be defined in (A.10). Then with probability $1-\delta$,*

$$\|\boldsymbol{E}\boldsymbol{U}^*\|_{2\to\infty} \lesssim (M(\delta) + \tilde{j})\|\boldsymbol{U}^*\|_{2\to\infty} + \sqrt{R(\delta)p^*}.$$

Note that $\boldsymbol{E}_\nu = \boldsymbol{E}$. When $\delta = n^{-r}$, by (2.24),

$$\|\boldsymbol{E}_\nu\boldsymbol{U}^*\|_{2\to\infty} \lesssim \sqrt{n\bar{p}^* \log n}\|\boldsymbol{U}^*\|_{2\to\infty} + \sqrt{(\log n)\bar{p}^*} \lesssim \sqrt{n\bar{p}^* \log n}\|\boldsymbol{U}^*\|_{2\to\infty}, \tag{A.20}$$

where the last line uses the fact that $\sqrt{n}\|\boldsymbol{U}^*\|_{2\to\infty} \geq 1$.

Now we derive bounds for other terms. By (A.11) and the definition (2.22) of ν,

$$\frac{\Theta_\nu(n^{-r})}{\lambda_{\min}(\Lambda_\nu^*)} \leq \frac{\Theta_\nu(n^{-r})}{\nu} \lesssim \frac{1}{n(\underline{p}_* - p_\emptyset)} \tag{A.21}$$

Furthermore, by (A.11), (A.13), (A.16) and (A.17),

$$\frac{E_{+,\nu}^2(n^{-r})}{(\Delta_\nu^*)^2} + \frac{\Theta_\nu(n^{-r})\tilde{\sigma}_\nu(n^{-r})}{\Delta_\nu^*} \lesssim \frac{n\bar{p}^* \log n}{(n(\underline{p}_* - p_\emptyset))^2} + \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2} \lesssim \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2}, \tag{A.22}$$

where the last inequality uses (2.24). For the third term, note that

$$\boldsymbol{L}_\nu^* - \boldsymbol{\Sigma}_\nu^* = \boldsymbol{L}^* - \boldsymbol{\Sigma}^* + \nu\boldsymbol{J} - \frac{n-1}{n}\nu\boldsymbol{I}.$$

Thus,

$$\begin{aligned}\sqrt{n}\|\boldsymbol{L}_\nu^* - \boldsymbol{\Sigma}_\nu^*\|_{2\to\infty} \leq& \sqrt{n}\|\boldsymbol{L}^* - \boldsymbol{\Sigma}^*\|_{2\to\infty} + \sqrt{n}\nu\|\boldsymbol{J} - \frac{n-1}{n}\boldsymbol{I}\|_{2\to\infty} \\ \leq& \sqrt{n}\|\boldsymbol{L}^* - \boldsymbol{\Sigma}^*\|_{2\to\infty} + \nu \leq n\bar{p}^* + \nu \lesssim n\bar{p}^*.\end{aligned} \tag{A.23}$$

Furthermore, by (A.14),

$$\sqrt{n}\tilde{b}_{2,\nu}(n^{-r}) \lesssim \sqrt{n\bar{p}^* \log n}. \tag{A.24}$$

Putting (A.20) - (A.24) together and using the fact that $\sqrt{n}\|\boldsymbol{U}^*\|_{2\to\infty} \preceq 1$, we obtain that

$$\begin{aligned}&\sqrt{n}\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*) - \boldsymbol{U}^*\|_{2\to\infty} \\ \lesssim& \frac{\sqrt{n\bar{p}^* \log n}}{n(\underline{p}_* - p_\emptyset)} + \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2} + \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2}\frac{n\bar{p}^*}{n\bar{p}^*} \\ \lesssim& \frac{\sqrt{n\bar{p}^* \log n}}{n(\underline{p}_* - p_\emptyset)} + \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2} \\ \overset{(i)}{\lesssim}& \frac{(n\bar{p}^*)^{3/4}(\log n)^{1/4}}{n(\underline{p}_* - p_\emptyset)} + \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2} \\ \overset{(ii)}{\lesssim}& \frac{n\bar{p}^*\sqrt{n\bar{p}^* \log n}}{(n(\underline{p}_* - p_\emptyset))^2},\end{aligned}$$

where (i) uses (2.24) and (ii) uses the condition (2.6). As a consequence, there exists a constant C that only depends on r and ξ such that

$$\sqrt{n}\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*)-\boldsymbol{U}^*\|_{2\to\infty}\le C\frac{n\bar{p}^*\sqrt{n\bar{p}^*\log n}}{(n(\underline{p}_*-p_\emptyset))^2}.$$

By the condition (2.6) again,

$$\sqrt{n}\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*)-\boldsymbol{U}^*\|_{2\to\infty}\le \frac{C}{\sqrt{C_{\ell_\infty}}}.$$

If $C_{\ell_\infty}\ge C^2/c^2$,

$$\sqrt{n}\|\boldsymbol{U}\mathrm{sign}(\boldsymbol{U}^T\boldsymbol{U}^*)-\boldsymbol{U}^*\|_{2\to\infty}\le c,$$

with probability $1-(B(\tilde{j})+1)n^{-r}\ge 1-(10K+1)n^{-r}$.

Bibliography

Abbe, E. (2017). Community detection and stochastic block models: Recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Abbe, E., Bandeira, A. S., and Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Anandkumar, A., Hsu, D. J., Huang, F., and Kakade, S. M. (2012a). Learning mixtures of tree graphical models. In *NIPS*, pages 1061–1069.

Anandkumar, A., Tan, V. Y., Huang, F., Willsky, A. S., et al. (2012b). High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics*, 40(3):1346–1375.

Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962.

Beinlich, I., Suermondt, H. J., Chavez, R. M., and Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proc. 2nd European Conference on Artificial Intelligence in Medicine*, pages 247–256.

Bresler, G. and Karzand, M. (2020). Learning a tree-structured Ising model in order to make predictions. *The Annals of Statistics*, 48(2):713–737.

Cape, J., Tang, M., and Priebe, C. E. (2019). Signal-plus-noise matrix models: Eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250.

Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90.

Chickering, D. M. (2002a). Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498.

Chickering, D. M. (2002b). Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3(Nov):507–554.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98.

Damle, A. and Sun, Y. (2020). Uniform bounds for invariant subspace perturbations. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1208–1236.

Dasgupta, A., Hopcroft, J., Kannan, R., and Mitra, P. (2006). Spectral clustering by recursive partitioning. In *European Symposium on Algorithms*, pages 256–267. Springer.

Dasgupta, S. (1999). Learning polytrees. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 134–141.

Dasgupta, S. (2016). A cost function for similarity-based hierarchical clustering. In *Proceedings of the Forty-eighth annual ACM symposium on Theory of Computing*, pages 118–127.

Deng, S., Ling, S., and Strohmer, T. (2020). Strong consistency, graph Laplacians, and the stochastic block model. *arXiv preprint arXiv:2004.09780*.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Eldridge, J., Belkin, M., and Wang, Y. (2017). Unperturbed: Spectral analysis beyond Davis-Kahan. *arXiv preprint arXiv:1706.06516*.

Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633.

Foygel, R., Draisma, J., and Drton, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pages 1682–1713.

Gámez, J. A., Mateo, J. L., and Puerta, J. M. (2011). Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22(1-2):106–148.

Ghassami, A., Yang, A., Kiyavash, N., and Zhang, K. (2020). Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, pages 3494–3504. PMLR.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research*, 14(11).

Heinemann, U. and Globerson, A. (2014). Inferning with high girth graphical models. In *International Conference on Machine Learning*, pages 1260–1268. PMLR.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Honorio, J. and Jaakkola, T. (2014). Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392. PMLR.

Huete, J. F. and de Campos, L. M. (1993). Learning causal polytrees. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 180–185. Springer.

Jarrell, T. A., Wang, Y., Bloniarz, A. E., Brittin, C. A., Xu, M., Thomson, J. N., Albertson, D. G., Hall, D. H., and Emmons, S. W. (2012). The connectome of a decision-making neural network. *Science*, 337(6093):437–444.

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

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8(3).

Katiyar, A., Hoffmann, J., and Caramanis, C. (2019). Robust estimation of tree structured Gaussian graphical models. In *International Conference on Machine Learning*, pages 3292–3300. PMLR.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50.

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224.

Le, C. M., Levina, E., and Vershynin, R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561.

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

Lei, L. (2019). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*.

Lei, L., Li, X., and Lou, X. (2020). Consistency of spectral clustering on hierarchical stochastic block models. *arXiv preprint arXiv:2004.14531*.

Li, T., Lei, L., Bhattacharyya, S., Sarkar, P., Bickel, P. J., and Levina, E. (2018). Hierarchical community detection by recursive partitioning. *arXiv preprint arXiv:1810.01509*.

Lou, X., Hu, Y., and Li, X. (2021). Linear polytree structural equation models: Structural learning and inverse correlation estimation. *arXiv preprint arXiv:2107.10955*.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405.

Lyzinski, V., Tang, M., Athreya, A., Park, Y., and Priebe, C. E. (2016). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1):13–26.

Mao, X., Sarkar, P., and Chakrabarti, D. (2017). Estimating mixed memberships with sharp eigenvector deviations. *arXiv preprint arXiv:1709.00407*.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 403–410.

Moitra, A., Perry, W., and Wein, A. S. (2016). How robust are reconstruction thresholds for community detection? In *Proceedings of the Forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841.

Nepusz, T., Petróczi, A., Négyessy, L., and Bazsó, F. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107.

Netrapalli, P., Banerjee, S., Sanghavi, S., and Shakkottai, S. (2010). Greedy learning of Markov network structure. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1295–1302. IEEE.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. (2019). Learning tree structures from noisy data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1771–1782. PMLR.

Nowzohour, C., Maathuis, M. H., Evans, R. J., Bühlmann, P., et al. (2017). Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374.

Ouerd, M., Oommen, B. J., and Matwin, S. (2004). A formal approach to using data distributions for building causal polytree structures. *Information Sciences*, 168(1-4):111–132.

Pearl, J. (2009). *Causality*. Cambridge university press.

Rebane, G. and Pearl, J. (1987). The recovery of causal polytrees from statistical data. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pages 222–228.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. MIT press.

Tan, V. Y., Anandkumar, A., and Willsky, A. S. (2010). Learning Gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701–2714.

Tavassolipour, M., Motahari, S. A., and Shalmani, M.-T. M. (2018). Learning of tree-structured gaussian graphical models on distributed data under communication constraints. *IEEE Transactions on Signal Processing*, 67(1):17–28.

Thomas, M. and Joy, A. T. (2006). *Elements of Information Theory*. Wiley-Interscience.

Verma, T. and Pearl, J. (1991). *Equivalence and Synthesis of Causal Models*. UCLA, Computer Science Department.

Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in artificial intelligence*, pages 323–330. Elsevier.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for Gaussian Markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2):189–202.

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

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, 153(3):707–720.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292.