Efficient and Interpretable Learning on Graph Structures

by

Mingyu Qi

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

(Statistics)

in the University of Virginia

2024

Doctoral Committee:

Assistant Professor Tianxi Li, Advisor

Associate Professor Xiwei Tang, Chair

Associate Professor Civelek Mete

Assistant Professor Shan Yu

Mingyu Qi

mq3sq@virginia.edu

ORCID iD: 0009-0006-0360-4928

Acknowledgements

The biggest thanks go to my advisor, Professor Tianxi Li. I have realized, after numerous attempts, that words fall short in capturing our shared experiences and what I have learned from him. I will nonetheless hold these cherished memories close. I will never forget that day six years ago, when I walked into his office and seized the chance to have him as my advisor. The opportunity remains one of my greatest fortunes. I deeply appreciate his wisdom; it will undoubtedly be a guiding light throughout my life.

I extend special thanks to my other committee members, Professor Mete Civelek, Professor Xiwei Tang, and Professor Shan Yu, for their invaluable assistance throughout my doctoral journey. Additionally, I am grateful to Professor Daniel Keenan, Professor Can Le, Professor Wen Zhou, and Professor Jianhui Zhou for the thoughtful consideration and time they dedicated to my growth and learning.

I hold immense gratitude to my parents for their steadfast support, providing me with unconditional love, heartfelt encouragement, and financial stability. I thank my girlfriend, Zhou Lu, for her love: love is indeed just love. I would be remiss not to also mention my grandparents, who instilled in me the virtues of courage and kindness. Although they departed this world three year ago, I believe they have turned into stars in the night sky, smiling to me even in the deepest dark.

Finally, I want to thank all my friends, especially Bangyi Chen, Olivia Kalafian, Shuangyue Li, Yifan Li, Rihui Ou, Bowei Sun, Wenbo Sun, Xiangnan Xu, and Yuda Shao, who have been there for me without fail. I am so proud to be part of the big family–the Department of Statistics at the University of Virginia. Blue and orange are now my favorite colors. Go, Wahoos!

Abstract

Advances in data collection and the proliferation of social media have brought network data into various fields, including the social sciences, economics, transportation, and biology. Efficient structural learning on networks (e.g., node groupings and the frequency of specific patterns, such as triangles, within graphs) is crucial for understanding complex system dynamics, predicting behavior, and devising interventions to enhance system performance. This thesis presents novel principled statistical tools for structural learning on networks, emphasizing appealing statistical properties and computational efficiency.

The first project features a non-overlapping and separable penalty that approximates the overlapping group lasso penalty. Overlapping group lasso penalty is often used to introduce structured sparsity in statistical learning given the penalty's ability to eliminate predefined groups of parameters. However, when groups overlap, solving the group lasso problem can be time consuming in high-dimensional settings due to groups' non-separability. This issue has constrained the penalty's relevance to cutting-edge computational areas, such as gene pathway selection and graphical model estimation. The proposed approximation greatly improves the computational efficiency of optimization, especially for large-scale and high-dimensional problems. This project confirms the proposed penalty as the tightest separable relaxation of the overlapping group lasso norm within the family of ℓ_{q_1}/ℓ_{q_2} norms. Furthermore, estimators based on the proposed norm are statistically equivalent to those derived from overlapping group lasso in terms of estimation error, support recovery, and the minimax rate under the squared loss.

Scientists have broadly leveraged differential co-expression analysis to clarify diseases' biological mechanisms. Yet the unknown differential patterns tend to be com-

plicated. As such, models based on simplified parametric assumptions may not pinpoint all discrepancies. Meanwhile, biological theory further indicates that, rather than existing in isolation, genes operate in groups to perform biological functions. Thus, differential co-expression analysis becomes more meaningful when genes' co-functioning structure is taken into account. The second project develops a new means of identifying differentially correlated gene groups. This technique enjoys computational efficiency and accuracy while integrating group information in analysis. As a by-product, a parameter-tuning procedure is put forth which considers the structural assumption and outperforms standard methods. Multiple simulation examples and a differential analysis of vascular smooth muscle cell gene expression data substantiate its utility.

Network moments measure the frequency of specific patterns and are instrumental in network analysis. Subsampling techniques serve as tools for approximating network moments' distribution in scenarios where limited networks are available. Although the univariate distribution of network moments can be well approximated, fairly little is known about the consistency for their joint distribution. The third and final project in this thesis focus on estimating the joint distribution of network moments through network node subsampling, thereby aiding in the characterization of the network's distribution. Our contribution opens the door for computationally efficient and interpretable network analysis innovations. For example, a multivariate inference approach is illustrated by analyzing collaboration networks and guppy gene expression networks.

Contents

Chapter 1

Introduction

A network is, in its basic form, a collection of points paired together via lines. Within the network field's nomenclature, a point is called a node or vertex, while a line is termed an edge. A network is a simplistic representation that reduces a system to an abstract structure; only the fundamentals of connections or interaction patterns tend to be captured. Many systems of interest (e.g., in disciplines such as biology, computer science, sociology, and economics) can be framed as networks. Extensive research on network analysis over the past two decades has yielded rich insight into diverse mechanisms, including gene regulation, friendship formation, and ecosystem evolution (Lovász, 2012).

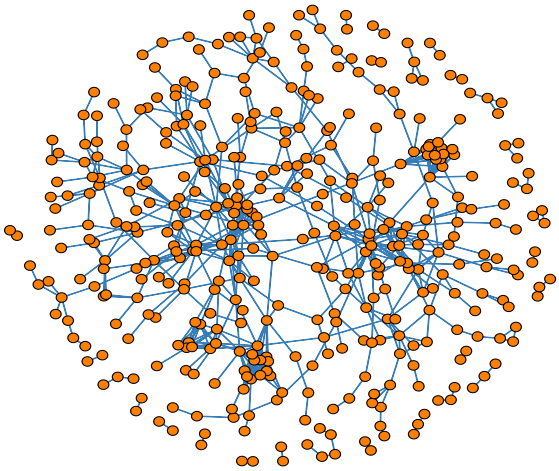


Figure 1.1: Example of a gene regulatory network inferred using gene expression dataset in Fischer et al. (2021) (see details in Section 4.5.2). Each node symbolizes a gene, and each edge represents an interaction between two nodes.

Some networks are directly observable. For instance, Ji and Jin (2016) de-

scribed a collaboration network; each node represents an author; an (undirected) edge exists between two authors if and only if they have coauthored two or more papers. Other networks may be inferred from different analyses. Examples include the protein-protein interaction network, where each node denotes a protein, with proteins' interactions appearing as edges (Perry et al., 2023). Vertices and edges in a network can also be labeled with supplementary information, such as names or strengths, to capture additional system details (Yan and Sarkar, 2021).

Structural information such as node groups or the total number of certain patterns can be useful for investigating and modeling the behavior of real-world systems. Many networks naturally separate themselves into groups or communities: networks of people divide into groups of friends or coworkers; the World Wide Web comprises groups of related web pages; and biochemical networks contain functional modules (Newman, 2018). Integrating network-group information is thereby a fruitful line of inquiry in network analysis. To illustrate, biological theory indicates that, instead of working in isolation, genes execute biological functions in groups. Gene network analysis thus becomes more meaningful when co-functioning groups of genes are considered (Ma and Kosorok, 2010).

A motif refers to a (usually simple) graph, such as an edge (), a 2-star (), or a triangle (). The count of motifs within a network offers crucial insights. For example, real collaboration networks may include more triangles because people introduce pairs of collaborators to each other, and those pairs often go on to cooperate (Newman, 2018). In another study, Milo et al. (2002) investigated various motifs' counts in genetic regulatory networks, suggesting that these motifs act as functional "circuit elements" such as filters or pulse generators. The motifs' counts might be an evolutionary outcome of their usefulness for the involved organisms.

Network moment, informed by motif counts, provides valuable details about

networks' population distribution (Borgs et al., 2010; Bickel et al., 2011). Network moments offer several advantages for network analysis as well; these moments are computationally efficient (Zhang and Xia, 2022) and can flexibly accommodate networks of different sizes (Shao et al., 2022). Network moments, therefore, play key roles in model estimation (Bhattacharyya and Bickel, 2015b) and have emerged as one of the most powerful tools for network comparison (Maugis et al., 2020).

This thesis presents statistical tools for network structural learning, with emphasis on appealing statistical properties and computational efficiency. All results are reproducible, and the corresponding code can be found at the following link: https://github.com/MingyuQi1995. Specifically, Chapter II centers on incorporating group information into analysis by pondering a question: "How can we effectively integrate group information if doing so is time-consuming when groups overlap?" In brief, the chapter suggests a non-overlapping, separable penalty as a statistically equivalent alternative to the overlapping group lasso penalty introduced by Jenatton et al. (2011a) for group selection via overlapping groups. The proposed approximation substantially improves the computational efficiency of optimization, especially for large-scale and high-dimensional problems. This project has been accepted for publication by the Journal of Machine Learning Research.

Chapter III features a novel approach to identifying differentially correlated gene groups. The developed tactic bolsters computational efficiency and accuracy while incorporating group information into analysis. A by-product, a novel parameter-tuning procedure, is also detailed; it accounts for the structural assumption and outperforms standard methods. The proposed method reveals dysregulated metabolic pathways in human vascular smooth muscle cell phenotypic switching, echoing Perry et al. (2023)'s effort.

Chapter IV focuses on characterizing network moments' distribution to gain in-

formation about underlying network models. The high-level question to be broached is "How can we estimate the sampling distribution of network moments if doing so is difficult because of inadequate observations?" The joint sampling distribution of network moments is then approximated based on the moments' joint distribution in subsampled networks. This approach opens the door for network analysis methods that are computationally efficient and interpretable. As an illustration, a multivariate inference is applied to analyze collaboration networks and guppy gene expression networks.

Notation and Preliminaries. Throughout this thesis, given a positive integer n, we define $[n] = \{1, 2, \ldots, n\}$. Given a set $\mathcal{S}$, $|\mathcal{S}|$ is the set's cardinality. When referring to a matrix A, $A_{\mathcal{S}}$ denotes the sub-matrix consisting of columns indexed by $\mathcal{S}$, and $A_{\mathcal{S},\mathcal{S}}$ denotes the sub-matrix induced by both rows and columns indexed by $\mathcal{S}$. Let $\mathbb{R}^{z\times p}$ be the set of all $z \times p$ matrices, and let S_+^p be the set of all $p \times p$ positive definite matrices. For a vector $v \in \mathbb{R}^p$, we define $\|v\|_a = (|v_1|^a + |v_2|^a + \ldots + |v_p|^a)^{\frac{1}{a}}$. Given a symmetric matrix A, denote its spectral norm, Frobenius norm, and operator norm as $\|A\|_2$, $\|A\|_F$, and $\|A\|_{a,b} = \sup_{\|u\|_a \leq 1} \|Au\|_b$, respectively. The largest and smallest singular values of A are denoted by $\gamma_{\max}(A)$ and $\gamma_{\min}(A)$.

Given two sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leqslant Cb_n$ for a sufficiently large n and a universal constant $C > 0$. We write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$. Furthermore, $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold.

Let G be an undirected unweighted graph whose node set is $V(G) = \{v_1, \ldots, v_n\}$ and whose edge set is $\mathscr{E}(G) = \{(v_i, v_j), v_i, v_j \in V(G)\}$. Let $A(G)$ be the $n \times n$ adjacency matrix of G, such that $A(G)_{ij} = 1$ if and only if $(v_i, v_j) \in \mathscr{E}(G)$. We will introduce other notations within the text as needed.

Chapter 2

The Non-Overlapping Statistical Approximation to Overlapping Group Lasso

2.1 Introduction

Grouping patterns of variables are commonly observed in real-world applications. For example, in regression modeling, explanatory variables might belong to different groups with the expectation that the variables are highly correlated within the groups. In this context, variable selection or model regularization should also consider the grouping patterns, and one may prefer to either include the whole group of variables in the selection or completely rule out the group. Group lasso (Yuan and Lin, 2006) is one popular method designed for this group selection task via adding ℓ_1/ℓ_2 regularization, as a broader class for group selection (Bach, 2008; Levina et al., 2008; Meier et al., 2008; Ravikumar et al., 2009; Zhao et al., 2009b; Danaher et al., 2014; Loh, 2014; Basu et al., 2015; Xiang et al., 2015; Campbell and Allen, 2017; Tank et al., 2017; Yan and Bien, 2017; Austin et al., 2020; Yang and Peng, 2020).

While the original group lasso penalty (Yuan and Lin, 2006) focuses on regularizing disjoint parameter groups, overlapping groups appear frequently in many applications such as tumor metastasis analysis (Jacob et al., 2009; Zhao et al., 2009b; Yuan et al., 2011; Chen et al., 2012) and structured model selection problems (Mohan et al., 2014; Cheng et al., 2017b; Yu and Bien, 2017; Tarzanagh and Michailidis, 2018). For example, in tumor metastasis analysis, scientists usually aim to select a small number of tumor-related genes. Biological theory indicates that rather than functioning in isolation, genes act in groups to perform biological functions.

Hence, the gene selection is more meaningful if co-functioning groups of genes are selected together (Ma and Kosorok, 2010). In particular, gene pathways, in the form of overlapping groups of genes, render mechanistic insights into the co-functioning pattern. Applying group lasso with these overlapping groups is then a natural way to incorporate the prior group information into tumor metastasis analysis. For another example, graphical models have been widely used to represent conditional dependency structures among variables. Cheng et al. (2017b) developed a mixed graphical model for high-dimensional data with both continuous and discrete variables. In their model, the groups are naturally determined by groups of parameters corresponding to each edge, and these groups overlap because edges share common nodes. Selecting the graph structures under this class of models requires eliminating groups of parameters from the model, which is achieved by the overlapping group lasso penalty.

The optimization involving the group lasso penalty with non-overlapping groups is efficient (Friedman et al., 2010; Qin et al., 2013; Yang and Zou, 2015). However, the overlapping group lasso problems present more complex challenges despite their convex nature. This is because the non-separability between groups intrinsically increases the problem's dimensionality compared with the non-overlapping situation, as revealed in the study of Yan and Bien (2017). Proposed methods for such optimization problems include the second-order cone program method SLasso (Jenatton et al., 2011a), the ADMM-based methods (Boyd et al., 2011; Deng et al., 2013), and their smoothed improvement, FoGLasso, introduced by Yuan et al. (2011). Nevertheless, these exact solvers of the problem involve expensive gradient calculations when the overlapping becomes severe, which may limit the applicability of the overlapping group lasso penalty in many large-scale applications such as genomewide association studies (Yang et al., 2010; Lee and Xing, 2012, 2014) or graphical model

fitting problems (Cheng et al., 2017b). For instance, Cheng et al. (2017b) showed that overlapping group lasso, though a natural choice for the problem, is infeasible even for moderate-size graphs, and they used a fast lasso approach (Tibshirani, 1996) to solve the graph estimation problem without theory. As we introduce later, our proposed solution includes the method of Cheng et al. (2017b) as a special case, but our method is more general and comes with theoretical guarantees.

In this project, we propose a non-overlapping approximation alternative to the overlapping group lasso penalty. The approximation is formulated as a weighted non-overlapping group lasso penalty that respects the original overlapping group patterns, making optimization significantly easier. The proposed penalty is shown to be the tightest separable relaxation of the original overlapping group lasso penalty within a broad family of penalties. Our analysis reveals that the estimator derived from our method is statistically equivalent to the original overlapping group lasso estimator in terms of estimator error and support recovery. The practical utility of our proposed method is exemplified through simulation examples and its application in a predictive task involving a breast cancer gene dataset. As a high-level summary, our major contribution to this project is the design of a novel approximation penalty to the overlapping group lasso penalty, which enjoys substantially better computational efficiency in optimization while maintaining equivalent statistical properties as the original penalty.

The remainder of this project is organized as follows: Section 2.2 introduces the overlapping group lasso problem and the proposed approximation method. We also establish the optimality of the proposed penalty from the optimization perspective. Section 2.3 details the statistical properties of the penalized estimator based on the proposed penalty. Comparisons between our estimator and the original overlapping group lasso estimator are made to show that they are statistically equivalent with

respect to estimation errors and variable selection performance. Empirical evaluations using simulated and real breast cancer gene expression data are presented in Sections 2.5 and 2.6, respectively. Finally, Section 2.7 concludes the project with additional discussions.

2.2 Methodology

2.2.1 Overlapping Group Lasso

Suppose in a statistical learning problem, the parameters are represented by a vector $\beta \in \mathbb{R}^p$, where β_j denotes the j-th element of β. Let $G = \{G_1, \cdots, G_m\}$ be the m predefined groups for the p parameters, with each group G_g being a subset of $[p]$, and $\cup_{g\in[m]} G_g = [p]$. For each group G_g, $d_g^G = |G_g|$ denoted the group size, and $d_{\max}^G = \max_{g\in[m]} d_g^G$. For any set $T \subset [p]$, β_T denotes the subvector of β indexed by T. Let $w = \{w_1, \cdots, w_m\}$ be the user-defined positive weights associated with the groups. The group lasso penalty (Yuan and Lin, 2006) is defined as

$$\phi^G(\beta) = \sum_{g\in[m]} w_g \left\|\beta_{G_g}\right\|_2. \tag{2.1}$$

We will omit G in all notations when the group structure is clearly given.

In statistical estimation problems involving group selection, the group lasso norm is combined with a convex empirical loss function L_n, and the estimator is determined by solving the following M-estimation problem:

$$\text{minimize}_{\beta\in\mathbb{R}^p} \left\{L_n(\beta) + \lambda_n \phi(\beta)\right\}. \tag{2.2}$$

If the groups are disjoint, then the group lasso penalty will select and elimi-

nate variables by groups. When the groups overlap, the above estimation enforces an "all-out" pattern by simultaneously setting all variables in certain groups to be zero, thus the zero-out variables are form a union of a subset of the groups (Jenatton et al., 2011a). Such a pattern is desirable in many problems, such as graphical models, multi-task learning and gene analysis (Jacob et al., 2009; Zhao et al., 2009b; Mohan et al., 2014; Cheng et al., 2017b; Tarzanagh and Michailidis, 2018). Another generalization of the group lasso for overlapping groups is the latent overlapping group lasso (Jacob et al., 2009; Mairal and Yu, 2013), following an "all-in" pattern by keeping the nonzero patterns as a union of groups. As noted in Yan and Bien (2017), the decision to use an "all-in" or "all-out" strategy depends on the problem and the corresponding scientific interpretations. The comparison between these two strategies is not our objective in this paper. However, both methods suffer from computational difficulties. We focus on introducing an approximation method for the overlapping group lasso penalty (2.1) and will leave the computational improvement of the latent overlapping group lasso for future work.

Problem (2.2) is a non-smooth convex optimization problem (Jenatton et al., 2011a; Chen et al., 2012), and the proximal gradient method (Beck and Teboulle, 2009; Nesterov, 2013) is one of the most general yet efficient strategies to solve it. Intuitively, proximal gradient descent minimizes the objective iteratively by applying the proximal operator of $\lambda_n \phi(\beta)$ at each step.

The proximal operator associated with group lasso penalty in (2.1) is defined as

$$\text{prox}_{\lambda_n}(\mu) = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2}\|\mu - \beta\|^2 + \lambda_n \phi(\beta). \tag{2.3}$$

whose dual problem is shown to be the following by Jenatton et al. (2011b):

$$\underset{\{\xi^g \in \mathbb{R}^p\}_{g\in[m]}}{\text{minimize}} \left(\frac{1}{2} \|\mu - \sum_{g=1}^{m} \xi^g\|_2^2 \right), \quad \text{s.t. } \|\xi^g\|_2 \leq \lambda_n w_g, \text{ and } \xi_j^g = 0 \text{ if } j \notin G_g. \quad (2.4)$$

The proximal operator (2.3) and its dual can be computed using a block coordinate descent (BCD) algorithm, as studied by Jenatton et al. (2011b). We list the procedure in Algorithm 1 for readers' information. The convergence of Algorithm 1 is guaranteed by Proposition 2.7.1 of Bertsekas (1997).

Algorithm 1 BCD for Proximal Operator of Overlapping Group Lasso

Input: Matrix $\mathbf{G}$, weights $\{w_g\}_{g=1}^m$, vector u, penalty λ_n.
Requirement: Weights $\{w_g\}_{g=1}^m > 0$, penalty $\lambda_n > 0$.
Initialization: Initialize $\{\xi^g\}_{g=1}^m$ to a zero vector in $\mathbb{R}^p$.
Output: Optimized coefficients β^*.

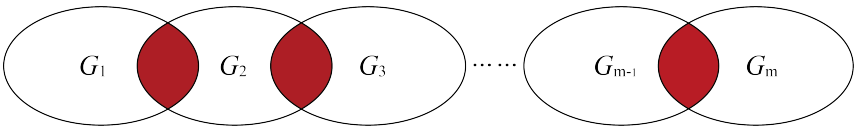
while stopping criterion is not met **do**
  for all $g \in \{1, \ldots, m\}$ **do**
    Compute residual $r^g = u - \sum_{h \neq g} \xi^h$.
    if $\|r^g\|_2 \leq \lambda_n w_g$ **then**
      $\xi_j^g = 0$ for $j \notin G_g$.
      $\xi_j^g = r_j^g$ for $j \in G_g$.
    else
      $\xi_j^g = 0$ for $j \notin G_g$.
      $\xi_j^g = \left(\frac{\lambda_n w_g}{\|r^g\|_2} \right) r_j^g$ for $j \in G_g$.
    end if
  end for
end while
Calculate $\beta^* = u - \sum_{g=1}^m \xi^g$.

Although additional techniques employing smoothing techniques have been developed to improve the optimization (Yuan et al., 2011; Chen et al., 2012), (2.3) and (2.4) continue to offer crucial insights into the computational bottlenecks caused by overlapping groups. Notably, the duality between (2.3) and (2.4) reveals that the overlapping group lasso problem has an intrinsic dimension equal to a $\sum_{g\in[m]} d_g$-

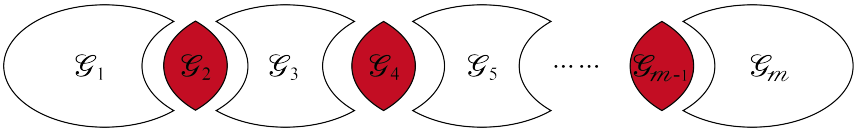
dimensional separable problem. When the groups have a nontrivial proportion of overlapping variables, the computation of the overlapping group lasso becomes substantially more difficult, eventually prohibitive on large-scale problems. This issue significantly limits the applicability of the overlapping group lasso penalty. Next, we introduce our non-overlapping approximation to rectify this challenge.

2.2.2 The Non-overlapping Approximation of the Overlapping Group Lasso

The fundamental challenge in solving overlapping group lasso problems stems from the non-separability of the penalty. To enhance computational efficiency, our approach hinges on introducing separable operators. As a starting point, we will illustrate this concept using a toy example of an interlocking group structure as a special case. In this structure, the groups are arranged sequentially, with each group overlapping with its adjacent neighbors (Figure 2.1a). For simplicity, we consider a uniform weight scenario where $w_g \equiv 1$ for all groups.



(a) Interlocking group structure.



(b) Partitioned group structure.

Figure 2.1: Illustration of proposed group partition in an interlocking group structure. Red regions are the overlapping variables in the original group structure.

We now partition the original overlapping groups in Figure 2.1b into smaller

groups as in Figure 2.1b. This partition identifies intersections as individual groups. We define these new groups as $\mathscr{G} = \{\mathscr{G}_1, \cdots, \mathscr{G}_{\mathscr{m}}\}$, where, in this specific instance, $\mathscr{m} = 2m-1$. Taking G_1 as an example. We have $G_1 = \mathscr{G}_1 \cup \mathscr{G}_2$ and by the triangular inequality,

$$\|\beta_{G_1}\|_2 \leq \|\beta_{\mathscr{G}_1}\|_2 + \|\beta_{\mathscr{G}_2}\|_2.$$

Extending this principle to each group, the norm of the overlapping group lasso based on G can be bounded by a reweighted non-overlapping group norm based on $\mathscr{G}$:

$$\sum_{g\in[m]} \|\beta_{G_g}\|_2 \leq \sum_{\mathscr{g}\in[\mathscr{M}]} h_{\mathscr{g}} \|\beta_{\mathscr{G}_{\mathscr{g}}}\|_2, \tag{2.5}$$

where $h_{\mathscr{g}}$ equals 1 for odd $\mathscr{g}$ and 2 for even $\mathscr{g}$. Consequently, controlling the sum on the right-hand side of (2.5) effectively controls the overlapping group norm on the left-hand side. The key advantage of this approach is the separability of the right-hand side norm, which substantially simplifies and enhances the efficiency of optimization.

While this example is about the interlocking group structures, the whole idea is applicable to any general overlapping pattern, as introduced in the two steps below.

Step 1: overlapping-induced partition construction. Our method starts from constructing a new non-overlapping group structure $\mathscr{G}$ from G, following Algorithm 2. We represent the initial group structure G by an $m \times p$ binary matrix $\mathbf{G}$, where $\mathbf{G}_{gj} = 1$ if and only if the j-th variable is a member of the g-th group, and $\mathbf{G}_{gj} = 0$ otherwise. To clearly differentiate the original group structure G and the derived non-overlapping structure $\mathscr{G}$, we employ standard letters, such as $\{g, d, m, w, G\}$, to represent quantities about the original group structure, while calligraphic letters, like $\{\mathscr{g}, \mathscr{d}, \mathscr{m}, \mathscr{w}, \mathscr{G}\}$, are used for quantities about $\mathscr{G}$. For instance,

$\mathscr{m}$ denotes the number of groups in $\mathscr{G}$, and $\mathscr{g} \in [\mathscr{m}]$ serves as the index for groups within $\mathscr{G}$.

Algorithm 2 Construction of Overlapping-Induced Partition $\mathscr{G}$

Input: Binary matrix $\mathbf{G}$.
Output: New group structure $\mathscr{G}$.

1: Begin with the set of all column indices $C = \{1, \ldots, p\}$.
2: Start with $k = 1$.
3: **while** C is not depleted **do**
4: Select the initial column index j from C, and define I as the set of column indices in $\mathbf{G}$ that match column $G_{\cdot j}$ exactly: $I = \{j' \in C \,|\, G_{\cdot j'} = G_{\cdot j}\}$.
5: Assign the set I as the group $\mathscr{G}_k$ and exclude I from C: $C \leftarrow C \setminus I$.
6: Increment k by one: $k = k + 1$.
7: **end while**
8: Output $\mathscr{G} = \{\mathscr{G}_1, \mathscr{G}_2, \ldots\}$, with each $\mathscr{G}_k$ indicating a distinct group.

Step 2: overlapping-based group weights calculation. Note that each group within $\mathscr{G}$ is a subset of at least one of the original groups in G. Conversely, each group in G can be reconstructed as the union of groups in $\mathscr{G}$. We introduce the following mappings:

$$F(\mathscr{g}) = \{g : g \in [m], \mathscr{G}_{\mathscr{g}} \subset G_g\} \quad \text{and} \quad F^{-1}(g) = \{\mathscr{g} : \mathscr{g} \in [\mathscr{m}], \mathscr{G}_{\mathscr{g}} \subset G_g\}.$$

Given positive weights w of G, we set the weights $\mathscr{w}$ of $\mathscr{G}$ as:

$$\mathscr{w}_{\mathscr{g}} = \sum_{g \in F(\mathscr{g})} w_g, \quad \mathscr{g} \in [\mathscr{m}]. \tag{2.6}$$

With the new partition $\mathscr{G}$ and the new weights $\mathscr{w}$ from the previous two steps, we define the following norm as the proposed alternative to the original overlapping group lasso norm:

$$\psi^{\mathscr{G}}(\beta) = \sum_{\mathscr{g}=1}^{\mathscr{m}} \mathscr{w}_{\mathscr{g}} \left\| \beta_{\mathscr{G}_{\mathscr{g}}} \right\|_2. \tag{2.7}$$

In general, by triangular inequality, the proposed norm is always an upper bound of the original group lasso norm:

$$\phi^G(\beta) = \sum_{g=1}^{m} w_g \left\|\beta_{G_g}\right\|_2 \leqslant \sum_{\mathscr{g}=1}^{\mathscr{m}} \mathscr{w}_{\mathscr{g}} \left\|\beta_{\mathscr{G}_{\mathscr{g}}}\right\|_2 = \psi^{\mathscr{G}}(\beta). \tag{2.8}$$

Our proposed penalty is essentially a weighted non-overlapping group lasso on $\mathscr{G}$. For illustration, Figure 2.2 shows the unit ball of these two norms based on $G_1 = \{\beta_1, \beta_2\}$ and $G_2 = \{\beta_1, \beta_2, \beta_3\}$ in a three dimensional problem. All singular points of the ϕ^G-ball (where exactly zero happens in (2.2)) are also singular points of the $\psi^{\mathscr{G}}$-ball.

Readers may observe that the inequality in (2.8) can also hold for other separable norms. For instance, consider partitioning all p variables into individual groups and employing a weighted lasso norm as another upper bound for ϕ^G, represented by:

$$\sum_{j=1}^{p} \Big(\sum_{\{g|\beta_j \in G_g\}} w_g \Big) |\beta_j|. \tag{2.9}$$

This approach was taken by Cheng et al. (2017b). So what is special about our proposed norm in (2.7)?

Intuitively, as illustrated by our construction process for $\mathscr{G}$ or Figure 2.2, our method introduces additional singular points in the norm only when it is necessary to achieve separability. Unlike the lasso upper bound, this process avoids adding redundancy. As such, our approximation is expected to maintain a certain level of tightness. We now formally substantiate this intuition. Given any group structure G and weights w, following Cai et al. (2022), we define the ℓ_{q_1}/ℓ_{q_2} norm of β for any $0 \leqslant q_1, q_2 \leqslant \infty$ as

$$||\beta_{\{G,w\}}||_{q_1,q_2} = \Big(\sum_{g\in[m]} w_g ||\beta_{G_g}||_{q_2}^{q_1} \Big)^{\frac{1}{q_1}}. \tag{2.10}$$

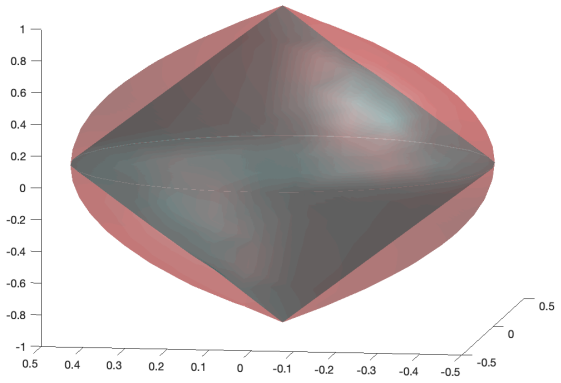


Figure 2.2: Illustration of two norms in $\mathbb{R}^3$: the outer region depicts the unit ball of the overlapping group lasso norm defined by $\{\beta : \phi^G(\beta) \leqslant 1\}$; the inner region represents the unit ball of our proposed separable norm $\{\beta : \psi^{\mathcal{G}}(\beta) \leqslant 1\}$.

This general class of norms potentially includes most commonly used penalties, including the weighted lasso penalty. The subsequent theorem shows that the proposed $\psi^{\mathcal{G}}(\beta)$ is the *tightest separable relaxation* of the original overlapping group lasso norm among all separable ℓ_{q_1}/ℓ_{q_2} norms.

Theorem 1. *Let $\mathbb{G}$ represent the set of all possible partitions of $[p]$. Given the original groups G and their weights w, there does not exist $0 \leqslant q_1, q_2 \leqslant \infty, \tilde{G} \in \mathbb{G}, \tilde{w} \in (0,\infty)^p$ such that:*

$$\begin{cases} \phi^G(\beta) \leqslant ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \psi^{\mathcal{G}}(\beta) & \text{for all } \beta \in \mathbb{R}^p \\ ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} < \psi^{\mathcal{G}}(\beta) & \text{for some } \beta \in \mathbb{R}^p \end{cases}. \tag{2.11}$$

2.3 Statistical Properties

Incorporating the proposed norm $\psi_{\mathcal{G}}$ into an M-estimation procedure leads to the following optimization problem:

$$\text{minimize}_{\beta\in\mathbb{R}^p}\left\{L_n(\beta)+\lambda_n\psi_{\mathcal{G}}\right\}, \tag{2.12}$$

which is different but related to (2.2). In this section, by studying the statistical properties of the regularized estimator based on $\psi_{\mathcal{G}}$ and the estimator based on ϕ_G, we show that $\psi_{\mathcal{G}}$ could be used as an alternative to ϕ_G. Following previous group lasso studies (Huang and Zhang, 2010; Lounici et al., 2011; Chen et al., 2012; Negahban et al., 2012; Dedieu, 2019), our analysis will focus on high-dimensional linear models. Specifically, define the linear model as

$$Y = X\beta^* + \varepsilon, \tag{2.13}$$

where $Y \in \mathbb{R}^{n\times 1}$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the covariate matrix, and $\varepsilon \in \mathbb{R}^{n\times 1}$ is a random noise vector. The overlapping group lasso coefficient estimator under the linear regression model is defined by a solution of (2.2) under the squared loss:

$$\hat{\beta}^G \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \quad \frac{1}{2n}\|Y-X\beta\|_2^2+\lambda_n\phi_G(\beta). \tag{2.14}$$

Correspondingly, we define the regularized estimator by our approximation norm as

$$\hat{\beta}^{\mathcal{G}} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \quad \frac{1}{2n}\|Y-X\beta\|_2^2+\lambda_n\psi_{\mathcal{G}}(\beta). \tag{2.15}$$

The solution uniqueness of (2.14) and (2.15) has been studied by Jenatton et al.

(2011a), and we include their results in Appendix A.1 for completeness. However, our study only requires the estimator to be one solution to the problem, as in Jenatton et al. (2011a); Negahban et al. (2012); Wainwright (2019). So we will not specifically worry about the uniqueness in our discussion.

As a remark, our objective is **not** to present (2.15) as an approximate optimization problem of (2.14). Rather, we focus on the statistical equivalence of the two classes of estimators defined by (2.14) and (2.15) in terms of their statistical properties under sparse regression models when appropriate values of λ_n are chosen (which may differ for each estimator). Our theoretical analysis focuses on three aspects. In Section 2.3.1, we establish that under reasonable assumptions, the ℓ_2 estimation error bound for (2.15) is no larger than that for (2.14). In Section 2.3.2, we present the minimax error rate for the overlapping sparse group regression problem, showing that both (2.14) and (2.15) are minimax optimal under additional requirements of the group structures. Lastly, in Section 2.3.3, we demonstrate that both estimators consistently recover the support of the sparse β^* with high probability under similar sample size requirements.

2.3.1 Estimation Error Bounds

We start by introducing additional quantities. Define the overlapping degree h_j^G as the number of groups in G containing β_j, and $h_{\max}^G = \max h_j$. Given a group index set $I \subseteq [m]$, we use G_I to denote the union $\bigcup_{g \in I} G_g$. Given G and I, following Wainwright (2019), we define two parameter spaces:

$$M(I) = \{\beta \in \mathbb{R}^p \mid \beta_j = 0 \text{ for all } j \in (G_I)^c\},$$
$$M^{\perp}(I) = \{\beta \in \mathbb{R}^p \mid \beta_j = 0 \text{ for all } j \in G_I\},$$

and we further use $\beta_{M(I)}$ to denote the projection of β onto $M(I)$.

Given any set $T \subseteq [p]$, we define the a set of groups $\mathsf{G}_T = \{g \in [m] \mid G_g \cap T \neq \emptyset\}$. Notice that $(G_{\mathsf{G}_T})^c$ is called the hull of T in Jenatton et al. (2011a). Let $supp(\beta) = \{j \in [p] \mid \beta_j \neq 0\}$ denotes the support set. We define the group support set $S^G(\beta) = \mathsf{G}_{supp(\beta)}$, and the augmented group support $\overline{S^G(\beta)} = \{g \in [m] \mid G_g \cap G_{S(\beta)} \neq \varnothing\}$. Furthermore, define $s = |supp(\beta)|$, $s_g = |S(\beta)|$, and $\overline{s_g} = |\overline{S(\beta)}|$. We omit the subscript G in notations when G is clearly given in context. Now we introduce additional assumptions under the regression model (2.13).

Assumption 1 (Sub-Gaussian noise for the response variable)**.** *The coordinates of ε are i.i.d. zero-mean sub-Gaussian with parameter σ. Specifically, there exists $\sigma > 0$ such that $\mathbb{E}[\exp(t\varepsilon)] \leqslant \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$.*

Our theoretical studies also hold for a fixed design of X, with trivial modifications. We prefer to introduce the random design here to make the statements more concise and interpretable, especially for the comparison in Section 2.3.3.

Assumption 2 (Normal random design for covariates)**.** *The rows of the data matrix X are i.i.d. from $N(0, \Theta)$, where $1/c_1 \leqslant \gamma_{\min}(\Theta) \leqslant \gamma_{\max}(\Theta) \leqslant c_1$ for some constant $c_1 > 0$.*

Lastly, we need some mild constraints on the group dimensions.

Assumption 3 (Dimension of the group structure)**.** *The predefined group structure G satisfies $d_{\max} \leqslant c_2 n$ for some constant $c_2 > 0$. In addition, we assume $\log m \ll n$.*

The following theorem establishes the ℓ_2 estimation error bounds for the two estimators.

Theorem 2. *Given G and its induced $\mathscr{G}$ according to Algorithm 2, define $h^g_{\min} = \min_{j \in G_g} h_j$, $h^g_{\max} = \max_{j \in G_g} h_j$. Let $\delta \in (0, 1)$ be a scalar that might depend on n. Under*

Assumptions 1, 2 and 3, for $\hat{\beta}^G$ and $\hat{\beta}^{\mathscr{G}}$ defined in (2.14) and (2.15), we have the following results:

1. *Suppose that β^* satisfies the group sparsity condition*

$$\overline{s_g}(\beta^*) \lesssim \frac{n}{\log m + d_{\max}} \cdot \frac{\min\limits_{g\in[m]}(w_g^2 h_{\min}^g)}{\max\limits_{g\in\overline{S}}(w_g^2 h_{\max}^g)}. \tag{2.16}$$

When $\lambda_n = \frac{c'\sigma}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}\sqrt{\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta}$ for some constant $c' > 0$, we have

$$\left\|\hat{\beta}^G - \beta^*\right\|_2^2 \lesssim \sigma^2 \cdot \frac{\left(\sum\limits_{g\in\overline{S}} {w_g}^2\right) \cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta\right). \tag{2.17}$$

with probability at least $1 - e^{-c_3 n\delta}$ for constant $c_3 > 0$.

2. *Suppose β^* satisfies the group sparsity condition*

$$\overline{s_{\mathscr{g}}}(\beta^*) \lesssim \frac{n}{\log \mathscr{m} + \mathscr{d}_{\max}} \cdot \frac{\min\limits_{\mathscr{g}\in[\mathscr{m}]}(\mathscr{w}_{\mathscr{g}}^2)}{\max\limits_{\mathscr{g}\in S}(\mathscr{w}_{\mathscr{g}}^2)}. \tag{2.18}$$

When $\lambda_n = \frac{c'\sigma}{\min\limits_{\mathscr{g}\in[\mathscr{m}]} \mathscr{w}_{\mathscr{g}}}\sqrt{\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta}$ for some constant $c' > 0$, we have

$$\left\|\hat{\beta}^{\mathscr{G}} - \beta^*\right\|_2^2 \lesssim \sigma^2 \cdot \frac{\sum\limits_{\mathscr{g}\in\{F^{-1}(g)\}_{g\in S}} {\mathscr{w}_{\mathscr{g}}}^2}{\min\limits_{\mathscr{g}\in[\mathscr{m}]}\left(\mathscr{w}_{\mathscr{g}}^2\right)} \cdot \left(\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta\right). \tag{2.19}$$

with probability at least $1 - e^{-c_4 n\delta}$ for constant $c_4 > 0$.

The error bound in (2.17) subsumes the non-overlapping group lasso error bound as a particular instance. When the groups in G are disjoint, the reduced form of

(2.17) matches the bounds studied in Huang and Zhang (2010); Lounici et al. (2011); Negahban et al. (2012); Wainwright (2019). The main difference in the context of overlapping groups is the necessity to account for the overlapping degree and the extension of sparsity requirements to augmented groups. The conditions specified in (2.16) and (2.18) relate to the cardinality of the augmented group support set (the number of non-zero groups in non-overlapping group structure). Although the conditions in (2.16) and (2.18) may initially appear distinct, they generally converge to a similar requirement in many typical cases, which can lead to an informative comparison between the two bounds in (2.17) and (2.19). The following results can characterize this.

Assumption 4. *Assume the predefined group structure G and its induced group structure $\mathscr{G}$ satisfy* $\max\{d_{\max}, m\} \asymp \max\{\mathscr{d}_{\max}, \mathscr{m}\}$.

Proposition 3. *Suppose that* $\max_{g\in\overline{S}} |F^{-1}(g)|$ *is bounded by a constant. Under Assumption 4, the following inequality holds:*

$$\frac{\sum_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}{}^2}{\min_{\mathscr{g}\in[m]} \left(\mathscr{w}_{\mathscr{g}}^2\right)} \cdot \left(\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta\right) \lesssim \frac{\left(\sum_{g\in\overline{S}} w_g{}^2\right) \cdot h_{\max}^{G_{\overline{S}}}}{\min_{g\in[m]} \left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta\right).$$

This implies that the error bound for the estimator $\hat{\beta}^G$ in (2.17) also serves as an upper bound for the error associated with the estimator $\hat{\beta}^{\mathscr{G}}$.

The quantity $|F^{-1}(g)|$ is the number of groups in $\mathscr{G}$ that has intersect with G_g. Proposition 3 requires that every G_g such that $G_g \cap supp(\beta^*) \neq \emptyset$ is partitioned into bounded number of non-overlapping groups. On the other hand, Assumption 4 requires that the maximum of two quantities — the maximum group size and the number of groups in the given group structure G — should have the same

order as those in the induced structure $\mathscr{G}$. The above requirement always holds for interlocking groups with similar groups and overlap sizes (see Figure 2.1). *More importantly, we can always assess the assumption directly on data* by calculating the group sizes and numbers for both G and $\mathscr{G}$. In Section 4.3, we evaluate five group structures from real-world gene pathways and examine the ratio of the maximum of two quantities from each G and $\mathscr{G}$. Assumption 4 looks reasonable in all of these real-world grouping structures. See details in Table 2.1.

2.3.2 Lower Bound of Estimation Error

Proposition 3 compares the two estimators' upper bounds of estimation errors. While the comparison gives intuitive ideas, it does not rigorously establish the statistical equivalence without the tightness of the error bounds. To strengthen our findings, we now investigate the minimax estimation error rate in linear regression models characterized by overlapping group sparsity. We will focus on the following class of group-wise sparse vectors:

$$\Omega(G, s_g) = \left\{ \beta : \sum_{G_g \in G} \mathbb{1}_{\{\|\beta_{G_g}\|_2 \neq 0\}} \leqslant s_g \right\} \tag{2.20}$$

Following the assumption of Cai et al. (2022), we focus on the special case of equal-size groups.

Assumption 5 (Equal size groups)**.** *The m predefined groups of G come with equal group size d, $m \ll p, d \ll \log(p)$.*

Theorem 4. *(Lower bound of estimation error)· Under Assumptions 1,2 and 5, we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega(G, s_g)} E\|\hat{\beta} - \beta\|_2^2 \gtrsim \frac{\sigma^2 \left(s_g(d + \log(\frac{m}{s_g})) \right)}{n}. \tag{2.21}$$

Combining Theorem 2 and Theorem 4, we can see that both estimators attain the minimax error rate and are statistically equivalent, as demonstrated by the following corollary.

Corollary 1. *Under Assumptions 1–4, if* $h_{\max}^{G_{\overline{S}}} \asymp 1$, *both* $\hat{\beta}^G$ *and* $\hat{\beta}^{\mathcal{G}}$ *attain the minimax estimation rate specified in* (2.21).

2.3.3 Support Recovery Consistency

We now proceed to analyze the support recovery consistency of $\hat{\beta}^G$ and $\hat{\beta}^{\mathcal{G}}$. We begin by introducing more quantities for our analysis. For any $\beta \in \mathbb{R}^p$, we define the mapping $r^G(\beta) : \mathbb{R}^p \to \mathbb{R}^p$ as:

$$r^G(\beta)_j = \begin{cases} \beta_j \sum\limits_{g \in \mathsf{G}_{supp(\beta)}, G_g \cap j \neq \emptyset} \frac{w_g}{\|\beta_{G_g \cap supp(\beta)}\|_2}, & \text{if } j \in supp(\beta), \\ 0, & \text{if } j \notin supp(\beta). \end{cases} \tag{2.22}$$

$r^G(\beta)$ is closely related to subgradients of the penalty and is used for determining optimality conditions. In the lasso case, $r^G(\beta)$ is the sign vector, which is exactly the lasso penalty. When focusing on β^*, we write $\mathbf{S} = supp(\beta^*)$, $\mathbf{r}^G = r^G(\beta^*)$, and $\beta^*_{\min} = \min\left\{|\beta^*_j|; \beta^*_j \neq 0\right\}$.

Our analysis essentially follows the strategy in Jenatton et al. (2011a). The major difference is that we study the problem with a more tailored setup for the random design rather than the fixed design as in Jenatton et al. (2011a). Using random designs, as discussed before, is helpful to compare the two estimators $\hat{\beta}^G$ and $\hat{\beta}^{\mathcal{G}}$ directly. We now introduce additional assumptions used to study the pattern consistency, which can be seen as the population-level counterpart of the assumptions in Jenatton et al. (2011a).

Assumption 1' (Gaussian noise for the response variable)**.** *Under model* (2.13)*, the coordinates of* ε *are i.i.d from* $N(0, \sigma^2)$*.*

Assumption 6 (Irrepresentable condition)**.** *For any* $\beta \in \mathbb{R}^p$*, define*

$$\phi^c_{\mathbf{S}}(\beta_{\mathbf{S}^c}) = \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2,$$

and its dual norm

$$(\phi^c_{\mathbf{S}})^*[u] = \sup_{\phi^c_{\mathbf{S}}(\beta_{\mathbf{S}^c}) \leq 1} \beta^\top_{\mathbf{S}^c} u.$$

Assume that there exists $\tau \in (0, \frac{2}{3}]$*, such that*

$$(\phi^c_{\mathbf{S}})^*[\Theta_{\mathbf{S}^c\mathbf{S}}\Theta^{-1}_{\mathbf{S}\mathbf{S}}\mathbf{r}_{\mathbf{S}}] \leqslant 1 - \frac{3\tau}{2}. \tag{2.23}$$

Assumption 1' is widely used to study support recovery consistency of linear regression. For example, in addition to Jenatton et al. (2011a), it is also used in Zhao and Yu (2006); Wainwright (2009, 2019). Assumption 6 is the population-level version of the irrepresentable condition as discussed in Zhao and Yu (2006) and Wainwright (2019).

Theorem 5. *Suppose Assumption 1', Assumption 2 and Assumption 6 hold. Under model* (2.13)*, assume the support of* β^* *is compatible with the overlapping group lasso penalty, such that the zero positions are given by an exact union of groups in* G*. Mathematically, that means*

$$[p] \setminus \{ \bigcup_{G_g \cap \mathbf{S} = \emptyset} G_g \} = \mathbf{S}. \tag{2.24}$$

1. *If*

$$\log(p - |\mathbf{S}|) \geqslant |\mathbf{S}|,$$

$$\lambda_n|\mathbf{S}|^{\frac{1}{2}} \lesssim \min\Big\{\frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum\limits_{g\in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}}\Big\}, \tag{2.25}$$

$$n \gtrsim \max\Big\{\frac{\sigma^2 \log(p-|\mathbf{S}|)}{a^2_{\mathbf{S}^c}\lambda_n^2}, \frac{\max\limits_{j\in\mathbf{S}}\{(\beta^*_j)^2\}\log(p-|\mathbf{S}|)}{a^2_{\mathbf{S}^c}\lambda_n^2}\Big\}, \tag{2.26}$$

where $a_{\mathbf{S}} = \min\limits_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g}{d_g}$, $a_{\mathbf{S}^c} = \min\limits_{g\in\mathsf{G}_{\mathbf{S^c}}} \frac{w_g}{d_g}$, *and* $A_{\mathbf{S}} = h_{\max}(\mathbf{G}_{\mathbf{S}}) \max\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g \|u\|_1$.

Then for the overlapping group lasso estimator $\hat{\beta}^G$, *we have:*

$$\begin{aligned}\mathbb{P}\Big(supp(\hat{\beta}^G) \neq \mathbf{S}\Big) \leqslant & 8\exp\Big(-\frac{n}{2}\Big) + \exp\Big(-\frac{n a^2_{\mathbf{S}}\tau^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\left\|\mathbf{r}_{\mathbf{S}}\right\|_2^2 \gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\Big) \\ & + \exp\Big(-\frac{n\lambda_n^2\tau^2 a^2_{\mathbf{S}^c}}{144\sigma^2}\Big) + 2|\mathbf{S}|\exp\Big(-\frac{nc^2(\mathbf{S},G)}{2\sigma^2}\Big)\end{aligned} \tag{2.27}$$

with

$$c(\mathbf{S},G) \asymp \min\Big\{\frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum\limits_{g\in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}}\Big\}.$$

2. *Furthermore, if* $\max_{g\in\mathsf{G}_{\mathbf{S}}} F^{-1}(g) \asymp 1$, *for the proposed estimator* $\hat{\beta}^{\mathscr{G}}$ *and assuming* $\max_{g\in\mathsf{G}_{\mathbf{S}}} F^{-1}(g) \asymp 1$, *the property holds:*

$$\begin{aligned}\mathbb{P}\Big(supp(\hat{\beta}^{\mathscr{G}}) \neq \mathbf{S}\Big) \leqslant & 8\exp\Big(-\frac{n}{2}\Big) + \exp\Big(-\frac{n a^2_{\mathbf{S}}\tau^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\left\|\mathbf{r}^{\mathscr{G}}_{\mathbf{S}}\right\|_2^2 \gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\Big) \\ & + \exp\Big(-\frac{n\lambda_n^2\tau^2 a^2_{\mathbf{S}^c}}{144\sigma^2}\Big) + 2|\mathbf{S}|\exp\Big(-\frac{nc^2(\mathbf{S},\mathscr{G})}{2\sigma^2}\Big),\end{aligned} \tag{2.28}$$

with

$$c(\mathbf{S},\mathscr{G}) \asymp \min\Big\{\frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum\limits_{g\in \mathscr{G}_{\mathbf{S}}} w_g \sqrt{|\mathscr{G}_g \cap \mathbf{S}|}}\Big\}.$$

The conditions involved in the above theorem can be seen as the population-level counterparts of those used in Jenatton et al. (2011a) for the overlapping group lasso estimator under the fixed design. As an illustration of the conditions, in

the lasso context, (2.25) and (2.26) reduce to the typical scaling of $n \approx \log p$ and $\lambda_n \approx \sigma(\log p/n)^{1/2}$. Together with the requirements on the sample size$|\mathbf{S}| \log(p-|\mathbf{S}|)$ and on $\beta^*_{\min}$, they match the requirements in Wainwright (2009) for the support recovery by the lasso regression. For non-overlapping group lasso estimators, our assumptions align with the conditions outlined in Corollary 9.27 of Wainwright (2019) under the random design.

Theorem 5 shows that both estimators consistently identify the support of the group sparse regression coefficients. Compared to the previous study of the overlapping group lasso estimator of Jenatton et al. (2011a), we switch to the random design of X, because such a setting renders a common basis for the comparison of the two estimators directly. Specifically, comparing (2.27) and (2.28), as well as the common conditions, we can see that the two estimators give comparable performance in support recovery with respect to the sampling complexity.

2.4 Comparison of Computational Complexity

In the previous section, we have shown that the proposed penalty induces a class of estimators statistically equivalent to the original overlapping group lasso estimator. In this section, we demonstrate the advantage of our proposed estimator in computational complexity. Specifically, solving (2.15) admits a lower complexity compared with solving (2.13).

As previously mentioned, the most common strategy for solving overlapping group lasso is based on proximal-based algorithms (Jenatton et al., 2011b; Yuan et al., 2011; Chen et al., 2012). These algorithms involve an outer loop implementing gradient-based steps and an inner loop executing the proximal operator (2.3), as studied in detail by Chen et al. (2012); Yan and Bien (2017). According to Chen

et al. (2012), the per-iteration time complexity for the proximal step is $O(\sum_{g\in[m]} d_g)$, and the proximal gradient method outer loops render a convergence rate of $O(1/\epsilon)$ in scenarios with overlapping groups, where ϵ denotes the desired accuracy.

In contrast, the proposed penalty converts the optimization of the overlapping group lasso problem to a non-overlapping group lasso problem. For any available proximal gradient algorithm, as the groups are disjoint, the proximal operator in (2.3) can be computed in closed form with the complexity of $O(p)$ for each iteration (Yuan and Lin, 2006), which gives a substantial reduction compared with the overlapping group lasso, especially when the groups in the original structure heavily overlap. Moreover, in non-overlapping scenarios, the outer loop enjoys an improved convergence rate of $O(1/\sqrt{\epsilon})$ (Liu et al., 2009a; Mairal et al., 2010). Therefore, solving (2.15) by proximal gradient methods enjoys better efficiency in both per-iteration complexity and number of iterations.

Furthermore, there are even more efficient strategies (Friedman et al., 2010; Qin et al., 2013; Yang and Zou, 2015) to solve the non-overlapping group lasso problem than the proximal gradient methods. These methods offer further improvements in the computational complexity. However, to our knowledge, these improvement options are unavailable for the overlapping group lasso problem. Hence, the proposed method can enjoy the benefits of these more efficient strategies, further amplifying its computational advantage.

2.5 Simulation

In this section, we assess the performance of the proposed estimator to demonstrate our claimed properties. At a high level, we want to use the simulation experiments to show that the proposed estimator based on (2.7) gives similar statis-

tical performance to the overlapping group lasso estimator while admitting much better computational efficiency. Our estimator achieves this primarily because of the tightest separable relaxation property of Theorem 1, which can be attributed to two designs of the norm (2.7): the induced partition $\mathscr{G}$ and the corresponding overlapping-based weights w. Therefore, in our simulation experiments, we will also evaluate the effects of these two designs by comparing the proposed estimator with other benchmark estimators. In Sections 2.5.1–2.5.3, we evaluate the performance of the proposed estimator and compare it with the weighted lasso estimator with overlapping-based weights, as discussed in (2.9), under various configurations. This sequence of experiments will demonstrate the importance of our proposed partition $\mathscr{G}$. In Section 2.5.4, we compare the proposed estimator with two other group lasso estimators, using the same $\mathscr{G}$ but overlapping-ignorant weights, under the same set of configurations. The results will demonstrate the importance of using the proposed overlapping-based weights w.

Two MATLAB-based solvers for the overlapping group lasso problems are employed. The first solver (Yuan et al., 2011) is from the SLEP package (Liu et al., 2009b). It can handle general overlapping group structures. The second solver is from the SPAM package (Mairal et al., 2014), which is designed to solve the overlapping group lasso problem when the groups can be represented by tree structures, formally defined in Section 2.5.2. Therefore, the SPAM solver is used only for the experiment in Section 2.5.2. The SLEP solver is more general, but using the two solvers can provide a more thorough evaluation across multiple implementations. For a fair comparison, the SLEP and SPAM package solvers were also applied to solve lasso and non-overlapping group lasso estimators in our benchmark set to ensure that the timing comparison implementation is consistent.

As an important side note, SLEP is widely acknowledged as one of the most effi-

cient solvers for the overlapping group lasso problem (Yuan et al., 2011; Chen et al., 2012; Cheng et al., 2017b). Still, for non-overlapping group lasso problems, alternative solvers, such as Yang and Zou (2015), may offer much better computational efficiency. For example, Yang and Zou (2015) reported that their solver is about 10–30 times faster than the SLEP package when solving non-overlapping group lasso problems. Such solvers are available because of the separability in non-overlapping groups and are not available for overlapping problems. For a fair comparison to avoid implementation bias, we use SLEP to solve for our estimator. Therefore, the computational advantage we demonstrate will be conservative. In practice, with the better solvers used, our method would enjoy an even more substantial computational advantage over the original overlapping group lasso than reported in the experiments.

Evaluation criterion. For each configuration, we generate 50 independent replicates and report the average result. The performance assessment is conducted in three aspects:

- **Regularization path computing time.** We begin by performing a line search to determine two pivotal values: $\lambda_{\max}$ and $\lambda_{\min}$. The search for $\lambda_{\max}$ starts at 10^8 and decreases progressively, multiplying by 0.9 at each iteration, until reaching the first value at which no variables are selected. In contrast, the determination of $\lambda_{\min}$ starts from 10^{-8} and increases incrementally, multiplying by 1.1 each time, until the first value is found that retains the entire set of variables. Following this, we select 50 values in log-scale within the range $[\lambda_{\min}, \lambda_{\max}]$. Subsequently, We compute the entire regularization path using these λ values and record the computation time associated with this process as a performance metric. The computing time evaluation mimics the most

practical situation where the whole regularization path is solved for tuning purposes.

- **Relative ℓ_2 estimation error**: From the entire regularization path, we select the smallest relative estimation error, defined as $\|\hat{\beta} - \beta^*\|_2 / \|\beta^*\|_2$, as the estimation error for the method. This serves as the measure of the ideally tuned performance.

- **Support discrepancy**: From the entire regularization path, we select the smallest support discrepancy, defined as $|\{i \in [p] : |\text{sign}(\hat{\beta}_i)| \neq |\text{sign}(\beta_i^*)|\}|/p$. Such a (normalized) Hamming distance is commonly used as a performance metric for support recovery (Grave et al., 2011; Jenatton et al., 2011a) to quantify the accuracy of pattern selection.

2.5.1 Interlocking group structure

In the first set of experiments, we evaluate the performances based on interlocking group structure (Figure 2.1a). This group structure exhibits a relatively low degree of overlap and is frequently used for evaluating overlapping group lasso methods (Yuan et al., 2011; Chen et al., 2012). Specifically, we set m interlocked groups with d variables in each group and $0.2d$ variables in each intersection. For example, $G_1 = \{1, \cdots, 10\}, G_2 = \{8, 9, \cdots, 17\}, \cdots, G_{10} = \{33, 34, \cdots, 42\}$ when $m = 5$ and $d = 10$. In the experiment, we will vary m and d to evaluate their impacts on the performance.

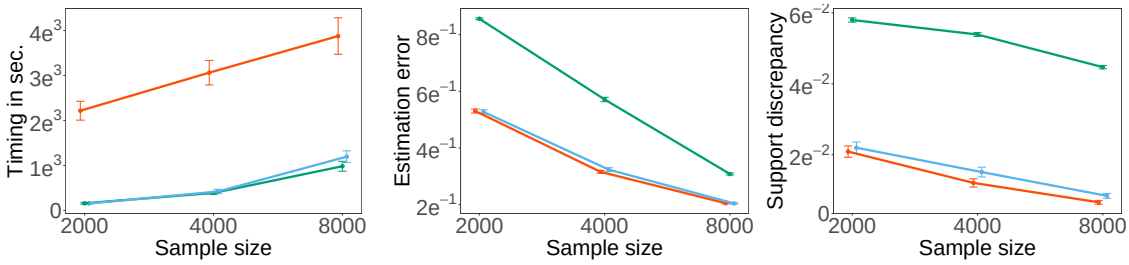
Following the strategy of Yan and Bien (2017), we generate the data matrix X from a Gaussian distribution $N(0, \Theta)$, where Θ is determined to match the correlations within the specified group structure. Initially, we construct a matrix $\tilde{\Theta}$ as

follows:

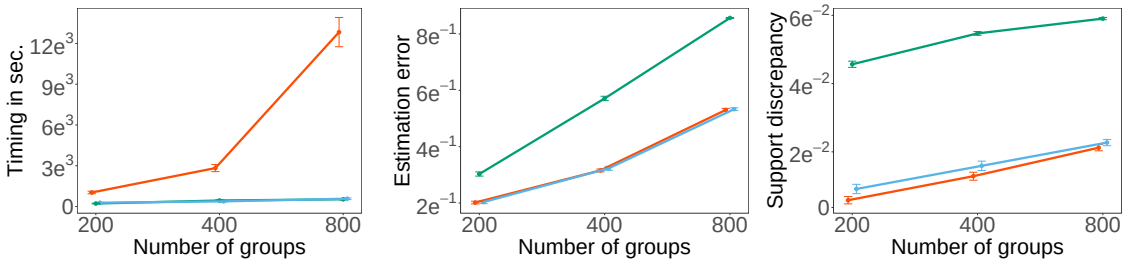
$$\tilde{\Theta}_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } \beta_i \text{ and } \beta_j \text{ belong to different groups in } G, \\ 0.6, & \text{if } \beta_i \text{ and } \beta_j \text{ are in the same group in } \mathscr{G}, \\ 0.36, & \text{if } \beta_i \text{ and } \beta_j \text{ are in the same group in } G \text{ but different groups in } \mathscr{G}, \end{cases}$$

and then Θ is derived as the projection of $\tilde{\Theta}$ onto the set of symmetric positive definite matrices with a minimum eigenvalue of 0.1. Such strong within-group correlation patterns have also been used in Zhao et al. (2009a); Yang and Zou (2015).

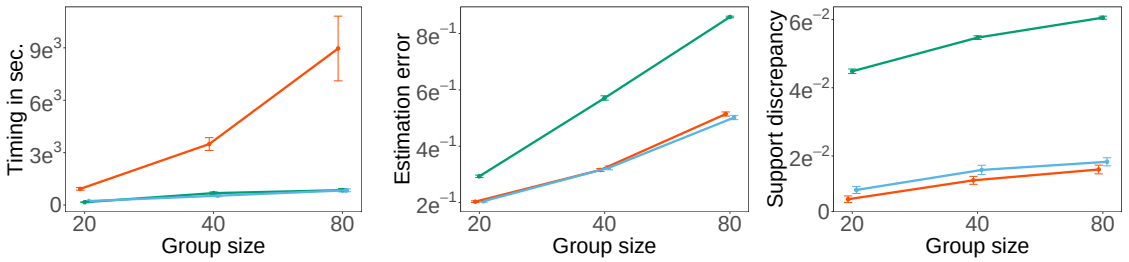
We generate β^* by first sampling its p coordinates from the normal distribution $N(10, 16)$, then randomly flipping signs of the covariates and randomly setting 90% of the groups to be zero. This setup aligns with the setting in Bach (2008); Friedman et al. (2010); Huang and Zhang (2010). The response variable Y is generated from $Y = X\beta^* + \epsilon$, where ϵ follows a normal distribution with mean 0 and variance σ^2, and we set $\sigma^2 = 3$ following Yang and Zou (2015). The group weight in the overlapping group lasso problem is $w_g = \sqrt{d_g}$, as is usually used in practice. We used the absolute difference in function values between iterations for all methods as the stopping criterion, with a tolerance set at 10^{-5}.



(a) Performance vs. Sample size



(b) Performance vs. Number of groups



(c) Performance vs. Groups size

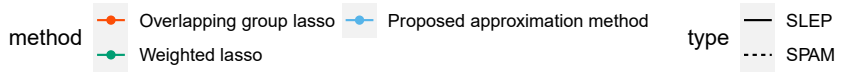


Figure 2.3: Regularization path computing time, ℓ_2 estimation error, and support discrepancy under different configurations of interlocking groups. (a) Varying sample size n when fixing $m = 400$ and $d = 40$ (p = 12808); (b) Varying number of groups m when fixing $n = 4000$ and $d = 40$; (c) Varying group size d when fixing $n = 4000$ and $m = 400$.

Figure 2.3 presents the average computation times, estimation errors, and support discrepancy with 95% confidence intervals (CIs). The result highlights the

significant computational advantage of the proposed method over the original overlapping group lasso. Specifically, our method is 5–20 times faster than the original overlapping group lasso.

Even though the overlap is not severe within the interlocking group structure, solving the overlapping group lasso problem carries a more substantial computational burden due to the non-separable structure within its penalty term. The computational time increases with larger sample sizes, a greater number of variables, and larger group sizes, and the computational disadvantage of the overlapping group lasso is more substantial as the problem scales up. In contrast, our proposed method consistently achieves accuracy similar to the overlapping group lasso estimator in both the estimation error and support discrepancy. This consistency in performance, observed across a spectrum of configurations, serves as an empirical confirmation of the validity of our theoretical findings.

On the other hand, the weighted lasso approximation is slightly faster than our method. This is expected from the optimization perspective. However, the weighted lasso approximation exhibits much higher errors than the overlapping group lasso estimator and our estimator across all configurations, revealing that the weighted also gives a poor approximation to the overlapping group lasso. This is because the weighted lasso fails to leverage the group information, different from the induced groups $\mathscr{G}$ used in our estimator.

In summary, our proposed estimator achieves comparable statistical performance to the original overlapping group lasso estimator while significantly enhancing computational efficiency. In contrast, although computationally efficient, the weighted lasso yields notably poor estimations, rendering it an uncompetitive alternative for approximating the original problems.

2.5.2 Nested tree structure of overlapping groups

In this experiment, we evaluate the performance of the estimators under a configuration of the tree-group structures introduced in Jenatton et al. (2011b), as below.

Definition 1. *(Jenatton et al., 2011b) A set of groups $G = \{G_1, \cdots, G_m\}$ is said to be tree-structured in $[p]$ if $\cup_{g\in[m]} G_g = [p]$ and if for all $g, g' \in [m]$. $G_g \cap G_{g'} \neq \emptyset$ implies either $G_g \subset G_{g'}$ or $G_{g'} \subset G_g$.*

In particular, we consider the special case of the tree groups, the nested group structure where all groups are nested. This configuration is interesting as it represents an extreme setting of overlapping groups – the overlapping degree is maximized in a certain sense and we hope to evaluate the methods in this extreme scenario. The nested group structure was also used in a few previous studies (Kim and Xing, 2012; Nowakowski et al., 2023). In this experiment, the SPAM solver, designed for the tree group structures, is also used to provide a more thorough evaluation across different implementations. We consider the following nested group configuration: 800 groups $G = \{G_1, \ldots, G_{800}\}$ are established, where $G_g \subset G_{g+1}$ and $|G_g| = g \times 4$, $g = 1, \cdots, 800$ with $p = 3200$ in total. The sample size varies from 600 to 2400. The data matrix X is generated from $N(0, \Theta)$, where Θ is generated by first constructing the matrix $\tilde{\Theta}$ as

$$\tilde{\Theta}_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0.6, & \text{if } \beta_i \text{ and } \beta_j \text{ belong to the same group in } \mathscr{G}, \\ 0.36, & \text{if } \beta_i \text{ and } \beta_j \text{ are in one group in } G \text{ but in two groups in } \mathscr{G}, \end{cases} .$$

and then projecting $\tilde{\Theta}$ onto the set of symmetric positive definite matrices with min-

imum eigenvalue 0.1. The generative process for β^* and y remains nearly identical as before, where the only difference is that the first 90% of the groups are set to zero following the hierarchical structure. The group weights are set to $w_g = 1/d_g$ as suggested (Nowakowski et al., 2023). For a fair comparison of the two solvers, in this experiment, we adopt the stopping criterion provided in the SPAM package (Mairal et al., 2014) with a convergence tolerance 10^{-5}.

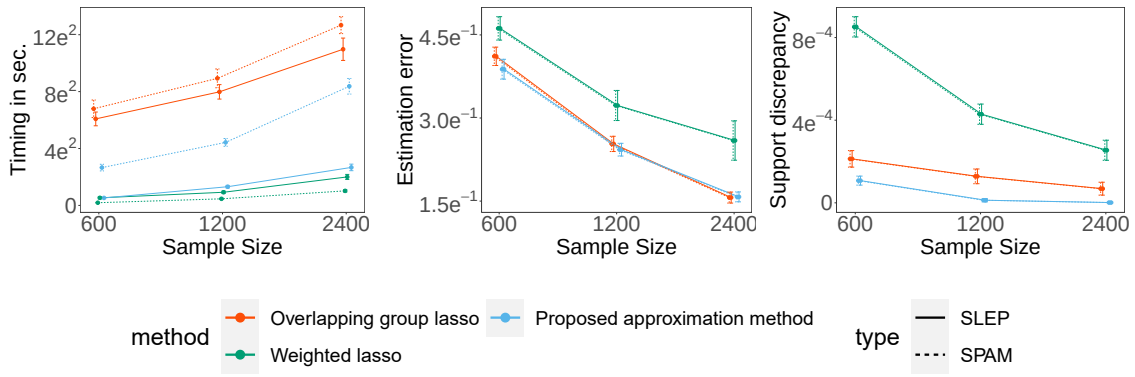


Figure 2.4: Regularization path computing time, ℓ_2 estimation error, and support discrepancy across various sample sizes under the nested tree group structure.

Figure 2.4 shows the performance of the three methods based on both solvers. SLEP is generally faster than SPAM, but the two solvers give consistent conclusions about the estimators. As studied by Jenatton et al. (2011b), solving the overlapping group lasso problem becomes highly efficient under such a nested group structure because, under a tree structure, a single iteration over all groups is adequate to obtain the exact solution of the proximal operator. Our timing results support this statement. Compared with the previous setting, the timing advantage of our method is reduced. However, our method is still at least twice as fast as the overlapping group lasso. When considering estimation error and support discrepancy, our proposed estimator consistently delivers similar results compared to the overlapping group lasso estimator. The comparison with the weighted lasso remains similar to

the previous experiment; while the lasso estimator is also fast to compute, it delivers very poor approximation.

In summary, solving overlapping group lasso problems exhibits efficiency when applied to tree structures. However, even in such cases, our proposed estimator maintains reasonable computational advantage and similar statistical estimation performance compared to the original overlapping group lasso estimator.

2.5.3 Group structures based on real-world gene pathways

Table 2.1: Summary information for the gene pathways: the mean and standard deviation of both group size ($\bar{d}/\mathrm{sd}(d)$) and the overlapping degree ($\bar{h}/\mathrm{sd}(h)$), the number of genes (p), and the ratio required in Assumption 4.

Pathways	$\bar{d}/\mathrm{sd}(d)$	$\bar{h}/\mathrm{sd}(h)$	p	$\max\{m, d_{\max}\}/\max\{m, d_{\max}\}$
BioCarta (Kong et al., 2006)	15.4/ 8.71	3.25/ 5.56	1129	2.35
PID (Schaefer et al., 2008)	38.51/ 19.59	3.28/ 5.09	2297	5.95
KEGG (Kanehisa et al., 2015)	58.48/ 47.36	2.58/ 3.39	4207	3.61
WIKI (Slenter et al., 2017)	38.17/ 44.10	4.35/ 7.70	6242	4.94
Reactome (Gillespie et al., 2021)	45.31/ 54.10	8.78/ 13.26	8331	2.35

The previous two sets of experiments are based on human-designed group structures. To reflect more realistic situations, in this set of experiments, we use five gene pathway sets from the Molecular Signatures Database (Subramanian et al., 2005) as group structures, summarized in Table 2.1. Each gene pathway represents a collection of genes united by common biological characteristics. These pathways have been widely adopted in studies of cancer and biological mechanisms (Menashe et al., 2010; Yuan et al., 2011; Livshits et al., 2015; Chen et al., 2020).

In particular, this data set can be used to assess the empirical applicability of Assumption 4 in our theory. The last column of Table 2.1 shows the ratio between $\max\{m, d_{\max}\}$ and $\max\{m, d_{\max}\}$. All values are within the range of [2,6],

indicating that the two terms can be treated as terms in the same order.

We use the gene expression data from Van De Vijver et al. (2002) as the covariate matrix X, which can be accessed through the R package `breastCancerNKI` (Schroeder et al., 2021). This design matrix has 295 observations and 24,481 genes. We perform gene filtering for each gene pathway set to exclude genes not defined within any pathways, a data processing step commonly used in similar studies (Jacob et al., 2009; Lee and Xing, 2014; Chen et al., 2012). The data-generating procedure for β^* and y remains almost the same as before, except that we use a much sparser model because of the smaller sample size of the data. Specifically, we randomly sample $0.05m$ active groups and set the coefficients in other groups to zero. The weights in overlapping group lasso are set to be $\sqrt{d_g}$.

Table 2.2: Comparison of the average computing time (in seconds) and the corresponding 95% confidence intervals for each pathway group structure.

Group Structure	Overlapping group lasso	Weighted lasso	The proposed approximation
BioCarts	67.18 [62.28, 72.08]	6.22 [5.99, 6.45]	16.03 [15.17, 16.89]
KEGG	287.27 [267.18, 307.36]	28.77 [26.42, 31.12]	48.32 [45.12, 51.52]
PID	445.99 [420.56, 471.42]	10.27 [9.74, 10.80]	31.25 [29.43, 33.07]
WIKI	1279.22 [1214.34, 1344.10]	63.56 [57.36, 69.76]	132.79 [121.82, 143.76]
Reactome	3739.97 [3569.27, 3910.67]	116.34 [106.32, 126.36]	194.61 [181.31, 207.91]

Table 2.2 displays the computing time, and Table 2.3 displays the estimation error results for the five pathway group structures. The high-level message remains consistent. Both our proposed group lasso approximation and the lasso approximation could substantially reduce the computing time. Across all settings, the proposed method reduces the computation time by 4 - 20 times and is more than

Table 2.3: Comparison of the relative ℓ_2 estimation errors and the corresponding 95% confidence intervals for each group structure.

Group Structure	Overlapping group lasso	Lasso	Proposed approximation
BioCarts	0.22 [0.20, 0.24]	0.28 [0.24, 0.32]	0.25 [0.22, 0.28]
KEGG	0.52 [0.47, 0.57]	0.80 [0.76, 0.84]	0.54 [0.51, 0.57]
PID	0.23 [0.21, 0.25]	0.50 [0.44, 0.56]	0.25 [0.23, 0.28]
WIKI	0.55 [0.49, 0.61]	0.65 [0.58, 0.72]	0.55 [0.49, 0.61]
Reactome	0.66 [0.63, 0.69]	0.85 [0.83, 0.87]	0.65 [0.62, 0.68]

Table 2.4: Comparison of the support discrepancy and the corresponding 95% confidence intervals for each group structure.

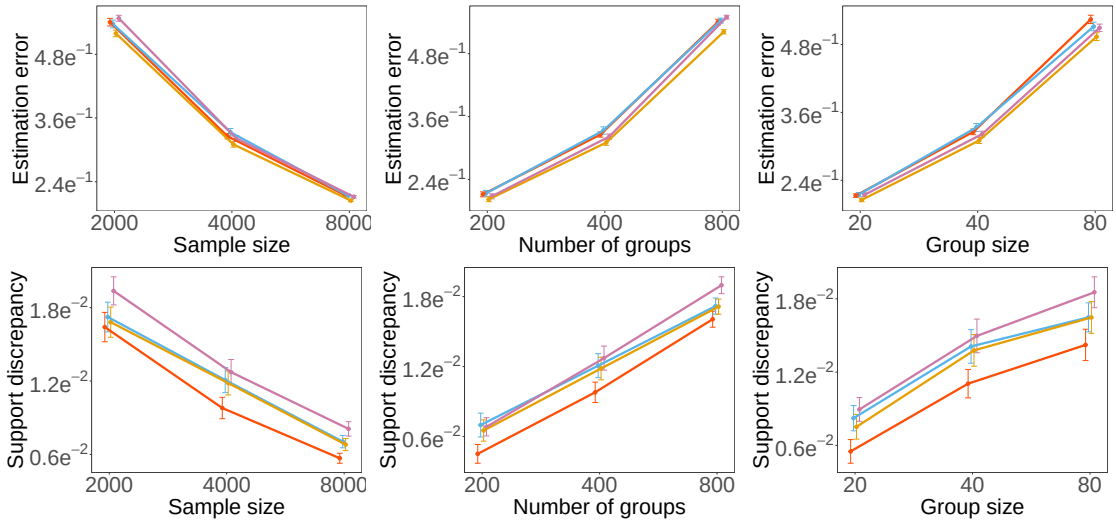
Group Structure	Overlapping group lasso	Lasso	Proposed approximation
BioCarts	0.041 [0.039, 0.043]	0.043 [0.040, 0.046]	0.041 [0.039, 0.043]
KEGG	0.023 [0.021, 0.025]	0.026 [0.024, 0.028]	0.023 [0.021, 0.025]
PID	0.033 [0.031, 0.035]	0.033 [0.031, 0.035]	0.033 [0.031, 0.035]
WIKI	0.013 [0.012, 0.014]	0.013 [0.011, 0.015]	0.013 [0.012, 0.014]
Reactome	0.012 [0.011, 0.013]	0.020 [0.019, 0.021]	0.012 [0.010, 0.014]

10 times faster in all settings with higher dimensions. Meanwhile, the proposed estimator delivers statistical performance similar to that of the original overlapping group lasso estimator. In contrast, the lasso approximation fails to leverage the group information effectively and yields inferior estimation results.

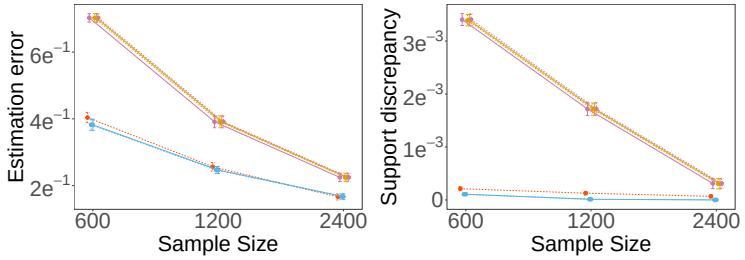
2.5.4 Comparison of different weighting choices

In addition to the partitioned groups, the overlapping-based weight defined in (2.6) for each partitioned group $\mathscr{g}$ is another crucial component to ensure the tightness of (2.7). We will demonstrate this aspect by experiments here to compare the proposed weights (2.6) with two other commonly used choices of weights that do not consider the original overlapping pattern: the uniform weights and group size-dependent weights (Yuan and Lin, 2006), on the same induced groups $\mathscr{G}$. Specifically, uniform weighting is the setting when all groups share the same weight while the size-dependent weighting uses the weight $\sqrt{d_{\mathscr{g}}}$ if $w_g = \sqrt{d_g}$ (interlocking and

gene pathway groups) and is $1/d_g$ if $w_g = 1/d_g$ (nested groups). The comparative analysis is performed under all group structures in the previous simulations, maintaining consistent simulation settings.



(a) Performance under interlocking group structure



(b) Performance vs. Sample size under nested tree structure

Figure 2.5: Regularization ℓ_2 estimation error and support discrepancy of the proposed method using different choices of weights under interlocking group structure and nested tree structure. Figure 2.5a is an extension to Figure 2.3, and Figure 2.5b is an extension to Figure 2.4.

Figure 2.5a and Figure 2.5b illustrate the weigh effects comparison in the settings

of Figure 2.3 and Figure 2.4, respectively. Under the interlocking group structure (Figure 2.5a), three weighting themes deliver similar performance in terms of estimation errors. Still, the size-dependent weighting leads to a larger support discrepancy. This interlocking group structure is not very distinctive for the three weights themes because the overlapping degree is nearly uniform. The nested group structures (Figure 2.5b) more effectively highlight the importance of the proposed weights. Our method significantly outperforms the other two weighting themes and aligns well with the original overlapping group lasso estimator. The weight design comparison on the gene pathway group structure is shown in Tables 2.5–2.6. The proposed estimator gives a close approximation to the original overlapping group lasso, but the other two weighing designs lead to significantly different performances in several settings.

Table 2.5: Comparative analysis of average estimation errors and the corresponding 95% confidence intervals for three weighting designs. The $*$ indicates that the error is statistically different from that of overlapping group lasso by a paired t-test.

Group Structure	Proposed weight	Uniform weight	Group size-dependent weight
BioCarts	0.25 [0.22, 0.28]	0.28 [0.26, 0.30]*	0.35 [0.30, 0.40]*
KEGG	0.54 [0.51, 0.57]	0.80 [0.77, 0.83]*	0.58 [0.51, 0.65]*
PID	0.25 [0.23, 0.27]	0.24 [0.21, 0.27]	0.39 [0.36, 0.42]*
WIKI	0.55 [0.49, 0.61]	0.83 [0.80, 0.86]*	0.74 [0.67, 0.81]*
Reactome	0.65 [0.62, 0.68]	0.58 [0.55, 0.61]*	0.69 [0.63, 0.75]

In summary, the experiments demonstrate that the weights designed in our penalty also serve as an indispensable part of a successful approximation to the overlapping group lasso estimation, which is another aspect of the tightest separable relaxation property in Theorem 1.

Table 2.6: Comparative analysis of average support discrepancy and the corresponding 95% confidence intervals for three weighting designs. The $*$ indicates that the value is statistically different from that of overlapping group lasso by a paired t-test.

Group Structure	Proposed weight	Uniform weight	Group size-dependent weight
BioCarts	0.041 [0.039, 0.043]	0.045 [0.042, 0.048]*	0.042 [0.039, 0.045]
KEGG	0.023 [0.021, 0.025]	0.059 [0.055, 0.063]*	0.024 [0.022, 0.026]
PID	0.033 [0.031, 0.035]	0.037 [0.035, 0.039]*	0.030 [0.027, 0.033]*
WIKI	0.013 [0.012, 0.014]	0.025 [0.023, 0.027]*	0.013 [0.012, 0.014]
Reactome	0.012 [0.010, 0.014]	0.010 [0.008, 0.012]*	0.022 [0.021, 0.023]*

2.6 Application Example: Pathway Analysis of Breast Cancer Data

In this section, we demonstrate the proposed method by predictive tasks on the breast cancer tumor data, as previously used in Section 2.5.3. This time, unlike the previous simulation studies, we use the complete data set with tumor labels for each observation. Specifically, each observation is labeled according to the status of the breast cancer tumors, with 79 classified as metastatic and 216 as non-metastatic. These labels serve as the response variable for our analysis.

Gene pathways have been widely used to key gene groups in cancer studies. In particular, Yuan et al. (2011); Chen et al. (2012); Lee and Xing (2014) used the overlapping group lasso techniques to exclude less significant biological pathways in cancer prediction. As a detailed example, Chen et al. (2012) leveraged the overlapping group lasso penalty to pinpoint biologically meaningful gene groups. Their analysis revealed multiple groups of genes associated with essential biological functions, such as protease activity, protease inhibitors, nicotine, and nicotinamide metabolism, which turned out to be important breast cancer markers (Ma and Kosorok, 2010). This evidence highlights the potential of using the overlapping group lasso penalty

in cancer analysis. On the other hand, another way to incorporate gene pathway information in such analysis is to retain genes by entire pathways. Jacob et al. (2009) used the latent overlapping group lasso penalty to achieve this while Mairal and Yu (2013) introduced an ℓ_∞ variant further. The success of all these previous studies reveals the potential of the gene pathway information in cancer prediction. They also show that the proper way to use the pathways (e.g., either eliminating-by-group, as in overlapping group lasso, or including-by-group, as in latent overlapping group lasso) highly depends on the data set and genes.

In our analysis, we use regularized logistic regression to build a classifier with the overlapping group lasso penalty (OGL), our proposed group lasso approximation penalty (Proposed approximation), the standard lasso penalty, the latent overlapping group lasso penalty (LOG) (Jacob et al., 2009), and the ℓ_∞ latent overlapping group lasso penalty of (Mairal and Yu, 2013). As mentioned in previous sections, our focus is not on justifying the overlapping group lasso should be used. Instead, **our primary objective is to demonstrate that when an overlapping group lasso penalty is used, our method provides a good approximation to the overlapping group lasso (with a much faster computation) across various pathway sets** (Table 2.1), whether or not the overlapping group lasso penalty is the best option for the problem.

Two additional aspects can also be evaluated as by-products of our analysis. First, as the lasso penalty does not consider the pathway information, comparing the performance of the group-based penalty and the lasso penalty in this problem would verify whether a specific gene pathway set contains predictive grouping information for breast cancer tumor type. Second, by assessing the predictive performances among the overlapping group lasso classifier and the latent overlapping group lasso classifiers, we can verify whether a specific gene pathway set is more suitable for

eliminating-by-group or including-by-group strategies for prediction.

Table 2.7: Computing time (in seconds) under different pathway databases.

Method / Database	OGL	Lasso	Proposed approximation
BioCarts	732	26	75
KEGG	2468	102	225
PID	1231	41	107
WIKI	5172	170	395
Reactome	11356	321	1186

We adopt the evaluation procedure of Lee and Xing (2014), where we randomly split the data set into 200 training observations and 95 test observations. All methods are tuned by 5-fold cross-validation on the training data. We calculate the area under the receiver operating characteristic (AUC) curve, a commonly used metric for classifying accuracy (Hanley and McNeil, 1982), on the test data. The total time for the entire cross-validation process is recorded as computation time. The experiment is repeated 100 times independently. Table 2.7 and Table 2.8 show the average computing time and AUC, respectively.

The following can be summarized from the results:

- First and foremost, the proposed estimator acts as an effective and computationally efficient approximation for the overlapping group lasso estimator. The results evidently support this claim. The proposed estimator delivers predictive performance that is (the most) similar to the overlapping group lasso estimator across various pathway datasets while significantly reducing the computing time by roughly ten times.

- Second, the lasso classifier performs best only on the WIKI pathway set, suggesting that the pathways in the WIKI database might not be sufficiently informative for cancer prediction.

- Third, the superiority between the overlapping group lasso regularizations and the latent overlapping group lasso regularizations depends on the specific group information. Among the four pathway sets with useful group information, the overlapping group lasso delivers superior predictive performance for the Biocarts and PID databases, while the latent overlapping group lasso classifiers provide better predictions on the KEGG and Reactome databases.

Table 2.8: Predictive AUC results of the three methods under different pathway databases.

Method / Database	OGL	Lasso	Proposed approximation	LOG	LOG ∞
BioCarts	0.7103	0.6989	0.7242	0.6888	0.6995
KEGG	0.7021	0.6862	0.7081	0.7390	0.7333
PID	0.7475	0.7004	0.7301	0.6881	0.6891
WIKI	0.6862	0.7282	0.6893	0.7149	0.7207
Reactome	0.6921	0.7301	0.7053	0.7463	0.7438

As a remark, while our evaluation is based on prediction accuracy, it is not the only criterion to determine if a method is proper for the dataset. For example, Mairal and Yu (2013) found that neither the overlapping group lasso model nor the latent overlapping group lasso model outperformed simple ridge regularization in prediction. The value of structured penalties also lies in their ability to identify potentially more interpretable genes, depending on the biological interpretations.

2.7 Discussion

We have introduced a separable penalty as an approximation to the group lasso penalty when groups overlap. The penalty is designed by partitioning the original overlapping groups into disjoint subgroups and reweighing the new groups according to the original overlapping pattern. The penalty is the tightest separable relaxation of the overlapping group lasso among all ℓ_{q_1}/ℓ_{q_2} norms. We have also shown that for linear problems, the proposed estimator is statistically equivalent to the original overlapping group lasso estimator but enjoys significantly faster computation for large-scale problems.

Several interesting directions could be considered for future research. The overlapping group lasso penalty presents a variable selection by eliminating variables by entire groups. A counterpart selection procedure can include variables by entire groups, which is achieved by the latent overlapping group lasso (Jacob et al., 2009). This penalty also suffers from a non-separability computational bottleneck. It would be valuable to investigate whether a similar approximation strategy could be designed to boost the computational performance in this scenario. More generally, the introduced concept of "tightest separable relaxation" might be a promising direction for optimizing non-separable functions. Studying the more general form and corresponding properties of this concept may generate fundamental insights about optimization.

Chapter 3

Compressed spectral screening with overlapping groups for large-scale differential correlation analysis

3.1 Introduction

High-throughput RNA sequencing (RNA-seq) has emerged as a powerful tool in molecular biology studies (Stark et al., 2019) given its ability to provide quantitative insights into gene expression levels within biological samples (Conesa et al., 2016). Moreover, RNA-seq enables researchers to explore transcriptomic landscapes with unprecedented depth and resolution. Scholars are then better able to elucidate complex gene regulatory networks and biological processes (Deshpande et al., 2023).

RNA-seq is often applied for differential expression, which identifies genes expressed at different levels across two or more biological conditions (Soneson and Delorenzi, 2013; Stark et al., 2019). Multiple methods have been proposed for such analysis; examples include edgeR (Robinson et al., 2010), DEseq2 (Love et al., 2014), and limma (Ritchie et al., 2015). These approaches use a negative binomial distribution or a weighted linear model to address the overdispersed count data commonly seen in RNA-seq experiments. The techniques also offer robust tools for identifying genes that are differentially expressed between experimental conditions, thus shedding light on the biological processes and pathways underlying apparent changes.

It is similarly important to examine changes in correlation patterns across diverse conditions. This task, often called differential correlation analysis, serves several

purposes: it reveals details about genome architecture (Zhou et al., 2021), uncovers regulatory biomarkers (McKenzie et al., 2016), and helps predict gene functions (Barter et al., 2014; Ghazanfar et al., 2019; Miller and Bishop, 2021). Various approaches have been developed for this type of analysis and can be broadly classified into two categories, namely testing conditional correlations via graphical models and testing marginal correlations (Li et al., 2023). Conditional correlation–based methods aim to infer a differential network from observed data (Shojaie, 2021). For instance, Yuan et al. (2017) suggested using the d-trace loss to model a precision matrix. However, methods grounded in graphical models come with challenges, especially regarding their direct biological interpretability (Li et al., 2023). Techniques that center on marginal correlations (Schott, 2007; Li and Chen, 2012; Cai and Zhang, 2016) seek to identify differential correlation patterns by conducting statistical tests within certain structural frameworks (Chang et al., 2017; Zhu et al., 2017). For example, the spectral screening process which Li et al. (2023) devised leverages spectral structures and random sampling. This approach enables precise discernment of features exhibiting complex differential patterns. It is also scalable and capable of handling datasets encompassing thousands of genes.

The aforementioned methods primarily concentrate on identifying genes on an individual level instead of a group level. Yet biological theory indicates that, rather than behaving in isolation, genes operate in groups to perform biological functions. Gene selection is hence more meaningful if co-functioning groups are chosen together (Ma and Kosorok, 2010). Therefore, in this study, our goal is to incorporate group information into differential gene selection in order to pinpoint gene groups with differential correlation patterns.

Group lasso (Yuan and Lin, 2006) is a popular method designed to integrate group information tasks by adding ℓ_1/ℓ_2 regularization (Bach, 2008; Ravikumar

et al., 2009; Zhao et al., 2009b; Tank et al., 2017; Yan and Bien, 2017; Austin et al., 2020; Yang and Peng, 2020). However, the original group lasso is limited to non-overlapping groups. Overlapping groups frequently appear in cases such as tumor metastasis analysis (Jacob et al., 2009; Zhao et al., 2009b; Chen et al., 2012) and structured model selection problems (Mohan et al., 2014; Cheng et al., 2017b; Yu and Bien, 2017; Tarzanagh and Michailidis, 2018). Latent overlapping group lasso is an extension of group lasso that handles these groups (Jacob et al., 2009). By applying ℓ_1/ℓ_2 regularization to a set of latent variables, each supported by one of the groups, latent overlapping group lasso can recover the underlying groups of variables under mild conditions Jacob et al. (2009). This method has proven valuable in pathway analysis and has enjoyed widespread adoption in diverse fields Deng et al. (2021); Zeng and Breheny (2016); Perry et al. (2023).

In this paper, we introduce a novel approach based on spectral screening method (Li et al., 2023) and the overlapping group lasso penalty (Jacob et al., 2009) to identify differentially correlated gene groups. Our technique capitalizes on spectral screening's computational efficiency and accuracy while using the overlapping group lasso penalty to incorporate group information into the analysis. As a by-product, we propose a novel parameter-tuning procedure that accounts for the structural assumption and outperforms standard methods such as the Akaike information criterion (AIC) and cross-validation.

The remainder of this paper is organized as follows. We introduce our differential correlation analysis method and tuning approach in Section 3.2. In Section 3.3, we conduct experiments to demonstrate the efficiency of overlapping group lasso–based spectral screening (OGSS) and the advantages of our tuning method. Section 3.4 presents an analysis of real gene expression data. Finally, Section 3.5 concludes with a discussion.

3.2 Methodology

We will describe our algorithms by taking the covariance matrix as the statistic of interest. Assume we have a size–n_1 random sample $x_i, i = 1, \ldots, n_1$ from a distribution with mean μ_1 and covariance Σ_1. We also have a size–n_2 sample $y_i, i = 1, \ldots, n_2$ from a distribution with mean μ_2 and covariance Σ_2. Here, $\mu_1, \mu_2 \in \mathbb{R}^p$ and $\Sigma_1, \Sigma_2 \in S_+^p$. We assume that only a small set of coordinates $S \subseteq [p]$ leads to different correlations, with $|S| = s \ll p$. That is,

$$\Sigma_{1,ij} \neq \Sigma_{2,ij} \quad \text{only if} \quad i, j \in S.$$

Let $G = \{G_1, \cdots, G_m\}$ be the m predefined groups for the p parameters, with each group G_g being a subset of $[p]$, and $\cup_{g\in[m]} G_g = [p]$. For each group G_g, $d_g^G = |G_g|$ denotes the group size. We further assume that S is formed by a combination of groups:

$$S = \bigcup_{g\in\mathscr{G}} G_g,$$

where $\mathscr{G} \subset [m]$. Our main objective is to identify S. Let $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ be the sample covariance matrices of X_1 and X_2, respectively. We further define $D = \Sigma_1 - \Sigma_2$ and $\hat{D} = \hat{\Sigma}_1 - \hat{\Sigma}_2$. Our method is introduced in Section 3.2.1; our tuning procedure is described in Section 3.2.2.

3.2.1 Overlapping group lasso–based spectral screening

We denote the rank of D as K, and we have $K \leq s$ in the current context. Let $D = U\Lambda U^T$ be the eigen-decomposition of D, where Λ is a square diagonal matrix of all the nonzero eigenvalues of D (with non-increasing magnitude). $U = (u_1, \ldots, u_K)$

consists of the corresponding eigenvectors. Without loss of generality, throughout this paper, we assume S is the set of the first s variables; that is, $S = [s]$. In this case, let U_1 be the matrix of the first s rows in U and U_2 be the matrix of the last $p - s$ rows. We then have

$$D = \begin{pmatrix} D_{S,S} & \mathbf{0}_{|S|\times(p-|S|)} \\ \mathbf{0}_{(p-|S|)\times|S|} & \mathbf{0}_{(p-|S|)\times(p-|S|)} \end{pmatrix} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \Lambda U^T = \begin{pmatrix} U_1 \Lambda U^T \\ U_2 \Lambda U^T \end{pmatrix}.$$

Following Li et al. (2023), such a pattern indicates that

$$0_{(p-s)\times p} = U_2 \Lambda U^T$$

and further leads to

$$U_2 = 0_{(p-s)\times m} U \Lambda^{-1} = 0_{(p-s)\times K}.$$

This relation suggests a simple strategy to identify S based on the rows in U, as these rows fully capture the sparsity pattern of the differential correlation structure (Li et al., 2023). Based on this idea, we propose estimating S with the following optimization problem:

$$\underset{U\in\mathbb{R}^{p\times\hat{K}}}{\arg\min} \frac{1}{2}||U_0 - U||_F^2 + \lambda\Omega^G(U), \tag{3.1}$$

where

$$\Omega^G(U) = \min \sum_{g=1}^{m} w_g \left\|U_g\right\|_2, \text{ s.t } \sum_{g=1}^{m} U_g = U.$$

If the groups are disjointed, then the group lasso penalty will select and eliminate variables by groups. The penalty $\Omega^G(U)$ enforces an "all-in" pattern when the

groups overlap by keeping the nonzero patterns as a combination of groups. This type of pattern is desirable for many problems, especially in gene analysis (Park et al., 2015; Zeng and Breheny, 2016). Another generalization of group lasso for overlapping groups is the mixed ℓ_1/ℓ_2 regularization proposed by Jenatton et al. (2011a): it simultaneously sets all variables in certain groups to zero, such that the zeroed-out variables represent a combined subset of groups. As noted in Yan and Bien (2017), the decision to use an "all-in" or "all-out" strategy depends on the problem and corresponding scientific interpretations. Comparing these two tactics is beyond the scope of this paper.

Algorithm 3 Estimate K with Compressed Spectral Screening

Input: Observation matrices X_1, X_2, sampling proportion p, validation proportion τ, and $K_l < K_u$.
Output: Zero sparse matrices $\hat{D}, \Omega, \hat{\Omega} \in \mathbb{R}^{p \times p}$.
Initialize $\hat{D}, \Omega, \hat{\Omega}$ as zero matrices.
for each $(i, j) \in [p] \times [p]$ with $i < j$ **do**
    Sample $\hat{\xi}_{ij} \sim \text{Bernoulli}((1+\tau)p)$ and $\hat{\phi}_{ij} \sim \text{Bernoulli}\left(\frac{1}{1+\tau}\right)$.
    Set $\hat{\xi}_{ij} = \hat{\xi}_{ij} \cdot \hat{\phi}_{ij}$.
    if $\hat{\xi}_{ij} = 1$ **then**
        Calculate $\hat{D}_{ij} = \text{Cov}(X_{1,i}, X_{1,j}) - \text{Cov}(X_{2,i}, X_{2,j})$.
        Set $\hat{D}_{ij} = \hat{D}_{ij}/p$.
    end if
end for
Compute the partial eigendecomposition of $\hat{D}$ up to rank K_u.
for $K_l \leq K \leq K_u$ **do**
    Approximate entries for (i, j) where $\hat{\xi}_{ij} = 1$.
    Calculate the loss L_K.
end for
Select $\hat{K}$ that minimizes L_K.
return $\hat{D}$ and $\hat{K}$.

Two inputs, U_0 and $\hat{K}$, are required for the optimization problem (3.1). In line with Li et al. (2023), we perform the singular value decomposition $U_0 U_0^\top = \hat{D}$ to obtain U_0. We next use Algorithm 3 proposed in Li et al. (2023), which is built on

cross-validation with basic random sampling validation to get $\hat{K}$:

The optimization problem (3.1) has a closed-form solution, which can be solved by Algorithm 4:

Algorithm 4 Differential Correlation Analysis with Group Regularization

Input: Estimated U_0, group structure G, regularization parameter λ, penalty factors w
Output: $\hat{U}$
for each group g in G **do**
    Compute adaptive $\lambda_g = \lambda \times w_g$
    Extract subset $U_g = (U_0)_g$
    Calculate Frobenius norm $\|U_g\|_F$
    if $\|U_g\|_F = 0$ **then**
        $\theta_g = 0$
    else
        $\theta_g = 1 - \frac{\lambda_g}{\|U_g\|_F}$
    end if
    Update U_g to $\theta_g \times U_g$
end for
Aggregate $\hat{U} = \sum_{g \in G} U_g$
return $\hat{U}$

3.2.2 Parameter tuning and average spectral norm (ASN) criterion

The implementation of OGSS requires an input parameter λ, which is closely related to the number of chosen groups and needs to be selected through a tuning procedure. Existing tuning approaches, including the AIC-based method (Akaike, 1998; Ding et al., 2018) and the cross-validation-based method (Cao et al., 2019), often fail due to the selection of overly sparse models. To address this issue, we introduce a novel tuning strategy named the average spectral norm (ASN) method. Our approach returns more favorable results by considering the structure of D.

3.2.3 ASN

The ASN method draws inspiration from the elbow method (Murphy, 2022), which focuses on identifying a sharp drop in estimated risk. Let $\lambda_{seq} = \{\lambda_1, \cdots, \lambda_T\}$ be a sequence of candidate parameters. Each $\lambda_t \in \lambda_{seq}$ corresponds to an estimated support $\hat{S}_t$. Intuitively, we want the estimated $\hat{S}$ to yield a large $\gamma(\hat{D}_{\hat{S},\hat{S}})_{\max}$ and a small $\gamma(\hat{D}_{\hat{S}^c,\hat{S}^c})_{\max}$ because the underlying differential covariance submatrix D_{S^c,S^c} contains all zeros. In addition, the following proposition causes us to consider the ASN: $\theta(\hat{S}) = \gamma(\hat{D}_{\hat{S},\hat{S}})_{\max}/|\hat{S}|$ and $\theta(\hat{S}^c) = \gamma(\hat{D}_{\hat{S}^c,\hat{S}^c})_{\max}/|\hat{S}^c|$.

Proposition 6. *Let* $n = \min\{n_1, n_2\}/2$. *If* $\gamma_{\max}(\Sigma_1) \leq c_1$ *and* $\gamma_{\max}(\Sigma_2) \leq c_2$. *Let* $n = \min\{n_1, n_2\}/2$, *then or any* $A \subset [m]$ *and* $\delta > 0$,

$$\Pr\left\{\theta(A) \leq \frac{1}{n} + \frac{(c_1 + c_2)(1+\delta)}{|A|}\right\} \geq 1 - 2\exp(-n\delta^2/2).$$

Proposition 6 indicates that the upper bound of $\gamma_{\max}(\hat{D}_{A,A})$ is of the order $O(|A|/n)$ with high probability. Thus, the upper bound of $\gamma_{\max}(\hat{D}_{A,A})$ can become quite large, even if $\gamma_{\max}(D)$ is a constant. This possibility warrants the contemplation of $\theta(A)$ in our analysis. Furthermore, the upper bound of $\theta(A)$ is related to $|A|$. As λ_t increases, a series of estimated supports emerges: $\hat{S}_1 \supset \hat{S}_2 \supset \ldots \supset S \ldots$, with the corresponding upper bound of $\theta(\hat{S})$ not decreasing for a fixed δ before meeting the true support S. Relatedly, the complementary sets $\hat{S}_1^c \subset \hat{S}_2^c \subset \ldots \subset S^c \ldots$ show that the upper bound of $\theta(\hat{S}^c)$ will be non-increasing until it meets the S^c.

As per the elbow method (Murphy, 2022), we determine $\lambda = \lambda_{\min\{\hat{t}_1,\hat{t}_2\}}$, where $\hat{t}_1 = \{t : \theta(\hat{S}_t) - \theta(\hat{S}_{t+1}) > \alpha_1\}$ and $\hat{t}_2 = \{t : \theta(\hat{S}_{t+1}^c) - \theta(\hat{S}_t^c) > \alpha_2\}$. This choice follows the logic that a substantial gap between $\theta(\hat{S}_t)$ and $\theta(\hat{S}_{t+1})$ signals a notable drop in $\gamma(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})_{\max}$ due to some differential variables not being selected.

Similarly, a sizable gap between $\theta(\hat{S}^c_{t+1})$ and $\theta(\hat{S}^c_t)$ implies the inclusion of non-differential variables.

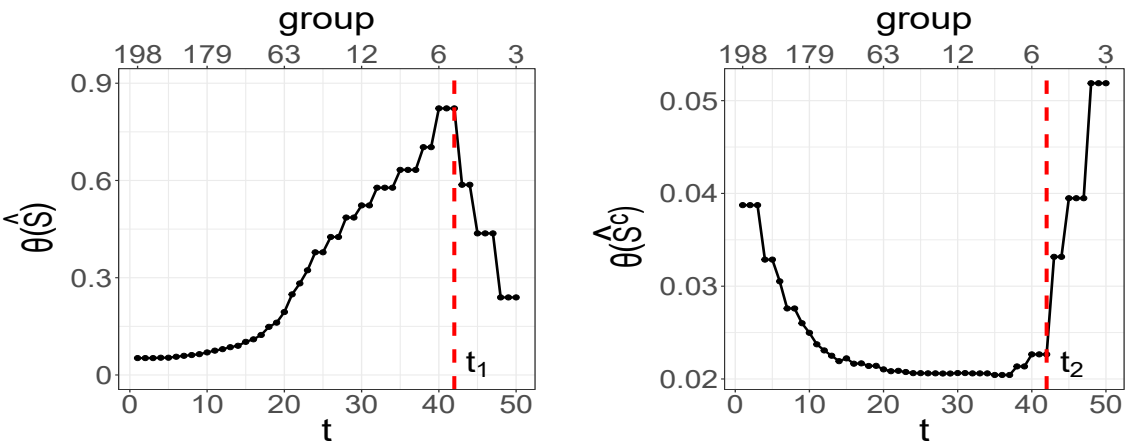


Figure 3.1: Estimated scores across different tuning parameters.

We illustrate this idea in Figure 3.1, where the underlying 6 groups are differentially correlated (for details about dataset generation, please see 3.3). The left panel displays the estimated $\theta(\hat{S})$ change across 50 candidate tuning parameters and is similar to the right panel of the estimated $\theta(\hat{S}^c)$. This figure shows that, as λ increases, $\hat{S}$ varies from all of the 200 groups to empty. The $\theta(\hat{S})$ increases initially but later decreases when a true group is excluded from $\hat{S}$, which corresponds to λ_{t_1}. A similar observation applies to $\theta(\hat{S}^c)$, and we can obtain λ_{t_2}. The ASN criterion ultimately selects $\lambda_{\hat{t}}$; $\hat{t} = \min\{t_1, t_2\}$ produces 6 groups, correctly recovering the true underlying differentially correlated groups.

In practice, we recommend $\alpha_1 = 1/|\hat{S}_t|$ and $\alpha_2 = 1/|\hat{S}^c_{t+1}|$, as suggested by Proposition 7. Proposition 7 indicates that, when $\theta(\hat{S}_t) - \theta(\hat{S}_{t+1}) > O(1/|\hat{S}_t|)$, there is a difference between $\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t})$ and $\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})$. In turn, $\hat{S}_t \supset S \supset \hat{S}_{t+1}$ because the difference between $\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t})$ and $\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})$ is closely related to that between $\gamma_{\max}(D_{S_t,S_t})$ and $\gamma_{\max}(D_{S_{t+1},S_{t+1}})$ (Wainwright, 2019). Using the same principle, we get $\alpha_2 = O(1/|\hat{S}^c_{t+1}|)$. Although there is no theoretical guarantee, this

choice of α_1 and α_2 works well in practice. It nonetheless stands to be analyzed more robustly in future work.

Proposition 7. *Suppose that* $\hat{S}_{t+1} = \hat{S}_t \setminus \Delta_t$. *Then*

$$\theta(\hat{S}_t) - \theta(\hat{S}_{t+1}) = \frac{\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t}) - \gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})}{|\hat{S}_t|} + \frac{\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})|\Delta_t|}{|\hat{S}_t|^2} + o(|\Delta_t|).$$

3.3 Simulation

In this section, we assess the performance of OGSS using simulated data and while considering tuning methods and scenarios with the difference matrix D being both low-rank and high-rank. We also demonstrate OGSS's effectiveness by comparing it with popular methods. Throughout our experiments, we focus on the following interlocking group structure with $m = 200$ groups, $d = 20$ variables per group, and $0.1d$ variables in each red intersection.

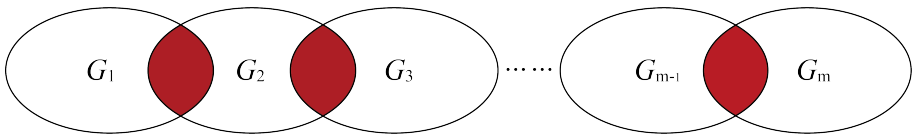


Figure 3.2: Interlocking group structure.

Among these m groups, we set the differential groups $\mathscr{G} = \mathscr{G}_1 \cup \mathscr{G}_2$, where each $\mathscr{G}_1$, $\mathscr{G}_2$ contains three elements randomly selected from $[m]$ with replacement. The data generation procedures are detailed in each configuration. To determine the lambda sequence for the tuning procedure, we start by performing a line search to identify two pivotal values: $\lambda_{\max}$ and $\lambda_{\min}$. The search for $\lambda_{\max}$ begins at 10^8 and progressively decreases, multiplying by 0.9 at each iteration, until reaching the first value at which no variables are selected. By contrast, the determination of $\lambda_{\min}$ starts at 10^{-8} and increases incrementally, multiplying by 1.1 each time, until

locating the first value that retains the entire set of variables. We subsequently choose 50 values in log-scale within the range $[\lambda_{\min}, \lambda_{\max}]$.

Evaluation criterion. For each configuration, we generate 50 independent replicates and report the average result. Our performance appraisal pertains to three aspects:

- To measure support recovery, we adopt the Jaccard index defined as follows:

$$SR = \frac{|\hat{S} \cap S|}{|\hat{S} \cup S|}. \tag{3.2}$$

- To measure group support recovery, we consider the group-level Jaccard index defined as follows:

$$GR = \frac{|\hat{\mathscr{G}} \cap \mathscr{G}|}{|\hat{\mathscr{G}} \cup \mathscr{G}|}. \tag{3.3}$$

- Following Cheng et al. (2017a); Fang et al. (2017); Li et al. (2023), we adopt the area under the ROC curve (AUC) (Hanley and McNeil, 1982) to measure the selection accuracy of the true differential variables. For OGSS, we generate the full ROC curve with respect to sensitivity and specificity by varying the tuning parameters, which control the number of variables chosen. For SS and DGCA, we obtain the full ROC curve by directly changing the number of selected variables as done by Li et al. (2023).

3.3.1 Evaluation of tuning methods under low-rank D

In the first simulation, we analyze OGSS based on multiple tuning methods. We next consider the problem scale $n_1 = n_2 = 40$ and $p = 3602$ and develop two samples x_i and y_i from multivariate normal distributions with zero means and the

following spiked covariance matrices:

$$\Sigma_1 = I + v_1 v_1^T, \;\; \Sigma_2 = I + v_2 v_2^T, \tag{3.4}$$

where v_1 is a p-dimensional vector with nonzero entries in $\{G_g\}_{g \in \mathscr{G}_1}$ from $N(0.4, 0.2)$ and zero entries elsewhere; v_2 contains nonzero entries in $\{G_g\}_{g \in \mathscr{G}_2}$ from $N(0.4, 0.2)$ and zero entries elsewhere. This data generation process aligns with that in Li et al. (2023).

Tuning method	GR (mean/sd)	SR (mean/sd)
ASN	0.796/0.034	0.798/0.034
CV	0.371/0.047	0.373/0.047
AIC	0.169/0.001	0.170/0.001
BIC	0.169/0.001	0.170/0.001

Table 3.1: Performance across various tuning methods.

As depicted in Table 3.1, our tuning approach outperforms the other three methods. This enhanced performance is due to ASN accounting for the structure of D. Additionally, both AIC and the Bayesian information criterion impose excessive penalties on the number of chosen variables, leading to overly sparse results.

3.3.2 Evaluation of tuning methods under high-rank D

The rank of matrix D was consistently equal to 2 under the previous setting. In this series of experiments, we increase the rank of D to further demonstrate OGSS's flexibility. In doing so, we substitute v_1 and v_2 with the matrices V_1 and V_2. Matrix V_1 contains nonzero entries in rows corresponding to the elements in $\mathscr{G}_1$. In the same vein, V_2 includes nonzero entries in rows aligned with $\mathscr{G}_2$. These entries are sampled from a multivariate normal distribution, with the mean and covariance adjusted to ensure that the Frobenius norm $|D|_F$ and the ratio $|D|_F/|\Sigma_1|_F$ are comparable

to those in the low-rank case. All other data generation steps are as described in Section 3.3.1.

Tuning method	GR (mean/sd)	SR (mean/sd)
ASN	0.848/0.031	0.851/0.031
CV	0.477/0.050	0.481/0.050
AIC	0.169/0.001	0.170/0.001
BIC	0.169/0.001	0.170/0.001

Table 3.2: Performance under full-rank condition across various tuning methods.

As shown in Table 3.2, despite this task being slightly simpler due to the stronger signals in D, the core message remains consistent: our tuning method outperforms others thanks to considering the pattern in D.

3.3.3 Misspecified group structures

In real-world applications, researchers may not always have access to accurate group information for analysis. Discrepancies between the underlying true group structure and the group structure used in an analysis may substantially influence outcomes. In pondering this prospect, we employ the setup in Section 3.3.1 to generate data and then consider the following group structures, which we enter into Algorithm 4:

- The true group structure G used in Section 3.3.1 to generate data: We represent G with an $m \times p$ binary matrix $\mathbf{G}$ where $\mathbf{G}_{gj} = 1$ if and only if the j-th variable belongs to the g-th group; $\mathbf{G}_{gj} = 0$ otherwise.
- Fully misspecified G_f: Randomly reassign all columns in $\mathbf{G}$.
- Partially misspecified G_{p_t}: Randomly reassign all columns in $\mathbf{G}_{,\mathscr{G}}$.
- Partially misspecified G_{p_f}: Randomly reassign all columns in $\mathbf{G}_{,\mathscr{G}^c}$.

As summarized in Table 3.3, OGSS performs relatively similarly across G, G_{pt}, and G_{pf}. This observation indicates that even when the group information is inaccurate, if incorrect patterns either exist only among differential variables or exclusively within non-differential variables, then the results remain largely unaffected (i.e., the findings closely mirror those acquired from the accurate group structure). However, the performance of OGSS under G_f is far inferior: inaccuracies in the group structure that span differential and non-differential variable patterns (i.e., when the true non-differential group contains differential variables or the true differential group contains non-differential variables) produce marked deviations from the outcomes expected with accurate group information.

Tuning method	**AUC (mean/sd)**	**SR (mean/sd)**
G_f	0.659/0.008	0.054/0.001
G_pt	0.953/0.014	0.803/0.035
G_pf	0.947/0.014	0.758/0.031
G	0.952/0.014	0.797/0.034

Table 3.3: Comprehensive performance comparison across various tuning methods.

3.3.4 Comparative analysis with other methods

In the last set of experiments, we compare OGSS with two widely used methods: compressed spectral screening (SS) (Li et al., 2023) and differential gene correlation analysis (DGCA)(McKenzie et al., 2016). The data generation process remains consistent with that in Section 3.3.1.

Method	AUC (mean/sd)	SR (mean/sd)
DGCA	0.655/0.004	0.029/0.001
SS	0.848/0.013	0.243/0.013
Our	0.952/0.013	0.798/0.034

Table 3.4: Performance of different methods.

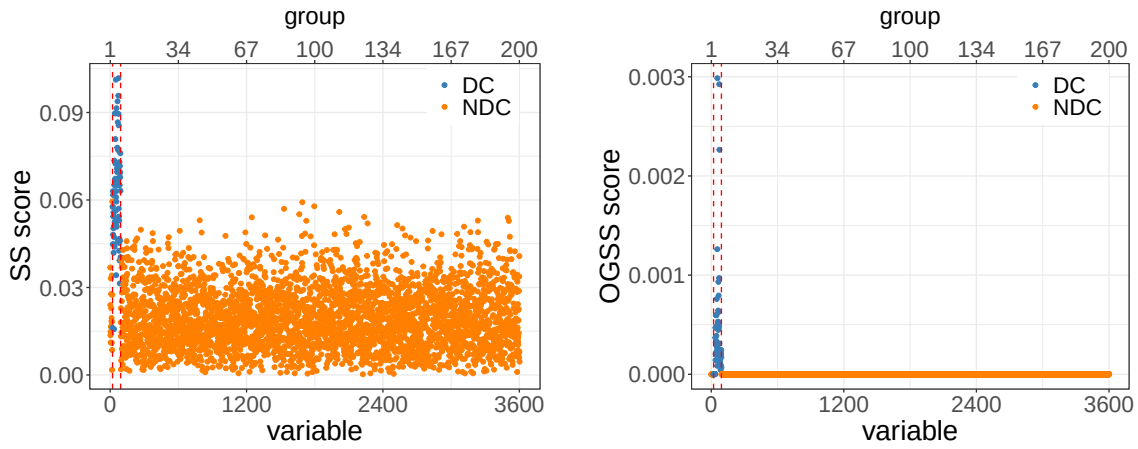


Figure 3.3: Estimated scores from Li et al. (2023)'s method and OGSS. Blue for true non-zero scores.

As shown in Table 3.4, our method delivers superior performance than the other two methods in terms of both AUC and support recovery. Further looking at each method, we noticed that DGCA turns out to have lots of false discoveries. This is due to the DGCA test whether the correlation between each pair of genes is significantly different, then using a BH procedure to control the false discovery rate (McKenzie et al., 2016). However, such a strategy inherently needs each test to be independent (Benjamini and Yekutieli, 2001). However, in our settings, the entries have correlations, and thus DGCA does not give good performance.

Our method also has a notable advantage over SS. Both approaches leverage the idea of spectral screening, which involves the nonzero rows U. We compare the

estimated scores. As pictured in Figure 3.3, SS turns out to be noisy on the non-differential variables and suffers from false discovery in the selection process. By contrast, our method reduces the scores of non-differential variables to zero, thus enabling more precise group-level selection.

3.4 Differential correlation analysis of vascular smooth muscle cell gene expression data

Coronary artery disease (CAD) is the leading cause of mortality in the United States. Vascular smooth muscle cells (VSMCs), a type of cell that constitutes the medial layer of the vessel wall, are known to influence each phase of atherosclerosis– the underlying cause of CAD (Perry et al., 2023).

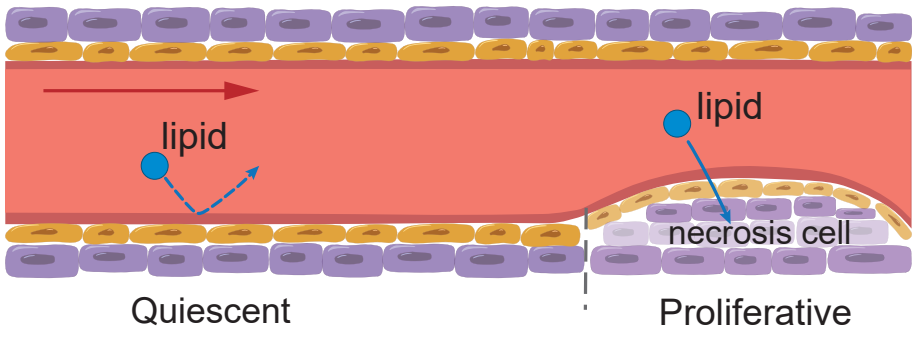


Figure 3.4: Illustration of the migration of VSMCs.

VSMCs show remarkable plasticity in response to vascular injury. As shown in Figure3.4, quiescent VSMCs can shift to a highly proliferative and migratory phenotype that encourages VSMCs to migrate into the intimal layer of the vessel wall. VSMCs in this layer produce extracellular matrix components and promote fibrous cap stability, thereby protecting against plaque rupture. As VSMCs undergo pheno-

typic switching, they can lose the expression of traditional VSMC marker genes and dedifferentiate into atheroprotective (fibroblast-like) and atherogenic (macrophage-like) cell types.2,3 Macrophage-like VSMCs become proinflammatory and release cytokines, facilitate phagocytic cells' migration, and accelerate cell necrosis and plaque growth. Despite extensive research, the mechanisms driving VSMCs' structural and functional phenotypic transformations are not fully understood. Unveiling how gene expression in VSMCs is being reprogrammed could prompt the discovery of novel treatments.

In this section, we analyze gene expression data processed by Perry et al. (2023). These RNA-sequencing data contain gene expression information from aortic SMCs isolated from 151 heart transplant donors of distinct genetic ancestries cultured under quiescent and proliferative conditions. Moreover, Perry et al. (2023) identified two gene group structures: one corresponding to the quiescent condition (quiescent group structure), comprising 41 gene pathways for 10,764 genes; and another for the proliferative condition (proliferative group structure), encompassing 45 gene pathways for 8,422 genes.

For quiescent group structure, we identified two modules: "orange" and "steelblue". These two modules are representative of quiescent gene expression profiles that are co-expressed only in the quiescent condition representing rewired biological pathways between that of the proliferative condition. These two modules are enriched for biological pathways of cell cycle and cellular metabolism processes (FDR < 0.05) and are representative of major biological features expected to be operating in phenotypic specific natures given the differences in experimental design between cells cultured to be in a resting, quiescent state and cells to be in an active proliferative state.

On the other hand, we identified two modules: "red" and "turquoise" for the

proliferative group structure that are enriched for biological processes more representative of disease biology relevant to the role SMCs play in contributing to atherosclerosis. Here, these two modules are enriched for cholesterol biosynthesis and lipid metabolism (FDR < 0.05). It's been shown by several groups that exposure to cholesterol or oxidized phospholipids triggers phenotypic switching of vascular smooth muscle cells to a proliferative, more macrophage/fibroblast-like cell state. In this model, however, we did not expose SMCs to such stimulus, yet we see that the metabolic pathways associated with this phenomenon are being rewired in a heightened proliferative state and that potentially the presence of lipids may not be sufficient in driving phenotypic changes, but maybe that increased proliferation of VSMCs results in metabolic processes for lipids/cholesterol that interact with lipids in an atherogenic manner.

3.5 Discussion

In this paper, we have devised a novel method named OGSS to identify differentially correlated groups of genes. OGSS manages to handle the large-scale problem as in SS and to integrate prior group information; doing so improves results' interpretability. We also propose a new criterion, ASN, and demonstrate its properties when choosing tuning parameters. A simulation study confirmed OGSS's superiority in selecting informative groups of genes with differential correlation patterns. Applying our method to vascular smooth muscle cell gene expression data further identified reasonable gene modules.

Several avenues could be considered for future research. The latent overlapping group lasso penalty allows for variable selection by choosing entire groups of variables. A complementary selection procedure could eliminate variables in the same

fashion (i.e., by group), namely via overlapping group lasso (Jenatton et al., 2011a). It would be interesting to investigate whether OGSS would be useful in this scenario. Another intriguing direction concerns our proposed method's theoretical properties. Several pertinent matters merit attention, including OGSS's consistency and the ASN tuning procedure. Our approach could also be extended to more complex group structures, such as hierarchical structures in Li et al. (2022). Multi-omics cases would be worthwhile to examine as well; pathways could be chosen to gain a holistic sense of underlying biological processes.

Last but not least, building on the promising results presented in this project, it would be insightful to delve deeper into the characterization of individual genes within the identified modules and to extend the scope of future studies to explore the potential relationships among these groups. Such discussions could illuminate how specific gene interactions contribute to the physiological processes in vascular smooth muscle cells and their implications in disease contexts like atherosclerosis. Additionally, considering the individual gene roles could help refine the OGSS methodology further, enhancing its applicability and accuracy in uncovering critical gene pathways across diverse biological systems. This focused approach on individual genes and their interconnections within groups could potentially uncover novel biomarkers and therapeutic targets, offering new avenues for treating vascular and other complex diseases.

Chapter 4

Multivariate Inference of Network Moments by Subsamping

4.1 Introduction

Networks, spanning diverse fields such as social sciences, biology, and computer science, are widely used as data structures facilitating the exploration of complex systems. Statistical network analysis serves as a powerful toolset for uncovering patterns, structures, and dynamics within these networks, enabling insights into phenomena ranging from social interactions to biological processes (Barabási, 2013; Newman, 2018). In this paper, we are interested in characterizing a population of networks based on a single observed network, which allows us to understand the underlying structure and dynamics inherent in complex systems in a broader scope.

In particular, network motif counts, such as the number of triangles or stars, play a crucial role in providing insights into the local structure and connectivity patterns within networks. By quantifying the prevalence of these motifs across a population of networks, we can discern common structural motifs and infer underlying mechanisms governing network formation and functions (Borgs et al., 2010; Bickel et al., 2011). For instance, a high number of triangles in a social network might indicate the presence of tightly-knit communities or cliques, while an abundance of stars could suggest influential nodes or hubs connecting disparate parts of the network (Wasserman and Faust, 1994). Furthermore, characterizing networks using motifs provides statistical advantages. Local motif counts reveal global properties and facilitate inference across networks of varying sizes but within the same

population. Consequently, motif counts are crucial in goodness of fit testing and model selection (Gao and Lafferty, 2017; Klusowski and Wu, 2020; Yuan et al., 2022), and network comparison tasks like two-sample tests and correlation analysis (Ghoshdastidar et al., 2017; Mao et al., 2021; Maugis et al., 2020; Shao et al., 2022).

At a high level, the statistical task of the paper can be stated as follows. Given a network G from an underlying generating model which will be assumed to be a graphon model Bickel and Chen (2009), and a set of motifs $R_1, \cdots, R_m$ of interest, we seek to characterize the distribution of the properly rescaled counts (i.e., network moments) of the motifs for the random networks (potentially with different size from G) under the true model.

One seemingly possible approach for this task is to estimate the true graphon model and then derive or sample the required distribution directly. However, identifying the graphon function is challenging without making restrictive assumptions (Yang et al., 2014; Chan and Airoldi, 2014) or resorting to computationally infeasible methods (Olhede and Wolfe, 2014; Choi and Wolfe, 2014; Gao et al., 2015). While there are computationally feasible and accurate methods for estimating the connection probabilities of the given network G (Chatterjee, 2015; Zhang et al., 2017; Li and Le, 2023), they cannot handle the population distributional studies at the graphon level. Additionally, all these estimation approaches still rely on certain smoothness assumptions. Hence, we alternatively turn to resampling strategies, generally regarded as flexible and versatile approaches for characterizing distributions. There have been many studies of resampling inference methods in network problems such as cross-validation (Chen and Lei, 2018; Li et al., 2020), bootstrap Green and Shalizi (2022); Levin and Levina (2019), subsampling (Bhattacharyya and Bickel, 2015b; Zhang and Xia, 2022; Lunde and Sarkar, 2023) and conformal inference (Lunde et al., 2023).

Among those, the most relevant ones to our current study are the subsampling method of Zhang and Xia (2022); Lunde and Sarkar (2023) and the bootstrap method of Green and Shalizi (2022); Levin and Levina (2019), which also focus on the distribution of motif counts from the population model. However, all of the above studies focus on the resampling approximation of the marginal distribution for a single motif, which offers a limited view of network structure by isolating individual aspects without considering interactions between motifs. Such marginal distributions fall short of capturing the full complexity of network interdependencies, potentially leading to less robust inferences about the network. Consider comparing two co-expression networks, G_1 and G_2, from different gene sets. Suppose we examine the counts (properly normalized for network size) of two motifs, V-shape ($\vee$) and 3-star ($\triangle$), separately. We might find G_1 has a statistically higher proportion of both motifs compared to G_2. However, because a V-shape is a subgraph of a 3-star, these counts are highly dependent. A greater number of V-shapes generally means more 3-stars, which might not provide additional insights into the networks' differences beyond what's indicated by the V-shapes. Consequently, ignoring this dependence can lead to numerous false positives and redundant comparisons, and thus, basing analysis and inference solely on separate marginal distributions can produce misleading scientific conclusions. This issue is exemplified in our examples later (Section 4.5). This example, along with similar cases, underlines the need to explore the joint distribution of motif counts as a crucial tool to understand multivariate objectives like dependence structures and conditional distributions.

In this paper, we introduce the use of node subsampling to characterize the joint distribution of multiple network moments. We show that subsampling provides an asymptotically accurate approximation of these joint distributions, extending the known effectiveness of subsampling from marginal to joint distributions. Our

findings enable more flexible approaches to network inference tasks. Specifically, we highlight its utility in two real-world network comparison studies: one examines collaboration patterns within the statistics community, including temporal changes and comparisons to high-energy physics, and the other compares gene networks across different gene sets. Both examples showcase the salient insights gained from multivariate network moment inference, which are not evident when only marginal distributions are considered. Further details are provided in Section 4.5.

4.2 Notations, motif counts and network moments

Throughout this paper, we denote the set $\{1, \cdots, n\}$ for any positive integer n by $[n]$. We use $|\cdot|$ to denote the cardinality of a set. Let G be an undirected unweighted graph whose node set is $V(G) = \{v_1, \ldots, v_n\}$ and edge set is $\mathscr{E}(G) = \{(v_i, v_j), v_i, v_j \in V(G)\}$.

A graph S is a subgraph of G, written as $S \subset G$, if $V(S) \subset V(G)$ and $\mathscr{E}(S) \subset \mathscr{E}(G)$. In particular, a subgraph $S \subset G$ is called an induced subgraph of G, denoted by $S \Subset G$, if for any $v_i, v_j \in V(S)$, $(v_i, v_j) \in \mathscr{E}(S)$ whenever $(v_i, v_j) \in \mathscr{E}(G)$. Lastly, two graphs S and G are isomorphic, denoted by $S \cong G$, if there exists a bijective function ϕ: $V(S) \to V(G)$ such that $(v_i, v_j) \in \mathscr{E}(S)$ if and only if $[\phi(v_i), \phi(v_j)] \in \mathscr{E}(G)$.

A motif refers to a (usually simple) graph, such as an edge (), a V-shape (), a triangle () or a 3-start , which constitutes the building blocks of larger graphs. In this study, we denote a motif by R, where $|V(R)| = r$ represents the number of vertices and $|\mathscr{E}(R)| = \mathfrak{r}$ is the number of edges. Our analysis exclusively considers connected motifs, aligning with previous research (Bickel et al., 2011; Bhattacharyya and Bickel, 2015b; Lunde and Sarkar, 2023). For a network G and motif R, we define

the motif count of R in G as the number subgraphs of G that are isomorphic to R:

$$X_R(G) = \big|\{S : S \subset G, S \cong R\}\big|.$$

This functional has received considerable attention in network analysis literature (Cook, 1971; Milo et al., 2002; Maugis et al., 2020; Bhattacharya et al., 2022) and is known as the non-induced motif count, indicating that the subgraph S need not be an induced subgraph of G. In contrast, the induced motif count is defined as

$$\tilde{X}_R(G) = \big|\{S : S \subset\!\!\!\subset G, S \cong R\}\big|$$

which requires that the subgraph S in the calculation must be an induced subgraph. Despite their mathematical equivalence via a linear mapping Bickel et al. (2011), non-induced counts offer a more streamlined theoretical framework (Zhang and Xia, 2022). Thus, following Bickel et al. (2011), Bhattacharyya and Bickel (2015b) and Zhang and Xia (2022), we focus on the non-induced motif count in our theoretical studies, while our data examples employ induced counts for enhanced interpretability.

The scale of motif counts is influenced by network size and motif size, making direct comparisons across networks of different sizes less informative. To address this, it is common to analyze a rescaled version of the motif count. Specifically, for a given motif R, the (sample) network moment of R in a graph G is defined as follows:

$$U_R(G) = \binom{n}{r}^{-1} X_R(G).$$

Several efficient computation strategies for network moments are outlined in Ribeiro and Silva (2010), Gonen et al. (2011), and Maugis et al. (2020). Additional

properties regarding network models are detailed in Section C.1.

4.3 Node subsampling and its properties

Our main goal for network inference is to characterize the properties of the population of networks, given one observation, which is the network G. Therefore, we shall first introduce the probabilistic framework we use to define the population. In this paper, we will use the following graphon framework (Hoover, 1979; Aldous, 1981; Bickel and Chen, 2009) for our study.

Definition 8. *[Sparse graphon model] Let the graphon function $w : [0,1]^2 \to [0,1]$ be a nonnegative Lebesgue measurable function, such that $w(u,v) = w(v,u)$ for any $u, v \in [0,1]$, such that $\int_0^1 \int_0^1 w(u,v)\, du\, dv = 1$. Furthermore, define a sequence of scalars $\rho_n \in [0,1]$. A random network is written to be $\mathbb{G}_n \sim \rho_n w(u,v)$ if it is generated as follows.*

1. *Generate $\{\xi_i\}_{i=1}^n$ independently as*

$$\xi_i \sim_{i.i.d} \text{Uniform}(0,1) \tag{4.1}$$

2. *For each node pair $(i,j), i < j$, connect the two nodes independently with probability $\rho_n w(u,v) \mathbb{1}_{\{\rho_n w(u,v) \le 1\}}$.*

The parameter ρ_n, facilitating network sparsity, typically tends towards 0 at a specified rate. Therefore, similar to Bickel et al. (2011), we always assume that $\rho_n w(u,v) \le 1$ in our analysis and ignore the constraint $\rho_n w(u,v) \le 1$.

We assume that the observed network, denoted as G, follows the sparse graphon model $\mathbb{G}_n$. From G, our aim is to infer the distributional properties of network

moments derived from the graphon model. Specifically, given a set of motifs $R_j, j \in [m]$, and a sample size b where $b < n$[1], we seek to characterize the distribution of network moments $U_{R_j}(\mathbb{G}_b)$, for $\mathbb{G}_b$ drawn from $\rho_b w$. Our primary emphasis, as discussed earlier, lies in the joint distribution of $U_{R_j}(\mathbb{G}_b), j \in [m]$, rather than their individual distributions.

Consider an ideal situation where we have knowledge of the true graphon model $\rho_n w$. We can then approximate the distribution of $U_{R_j}(\mathbb{G}_b), j \in [m]$ directly using the Monte Carlo method: by sampling numerous $\mathbb{G}_b$ from the model and computing their corresponding network moments to construct empirical CDFs or employing more advanced techniques for approximation. However, in our context, the graphon model is unknown thus the above procedure is not applicable. Nevertheless, if n is sufficiently large, we may consider the graph G as a discretized approximation of the true graphon, making a suitable sampling procedure based on G still feasible to mimic the Monte Carlo strategy. This insight motivates the subsequent subsampling algorithm.

Algorithm 5 Uniform Node Subsampling for Multivariate Network Moments

1: **Input:** Network G of size n; motifs $R_1, \ldots, R_m$; replication number N_{sub}; subsampling size b.
2: **Calculate** the edge density $\widehat{\rho}_G = |\mathscr{E}(G)|/[n(n-1)]$.
3: **for** $i = 1$ to N_{sub} **do**
4: Randomly sample b nodes (without replacement) from $[n]$ to form the subsampled set $\mathcal{S}$.
5: Define $G_b^{*(i)}$ as the induced subgraph of G by $\mathcal{S}$.
6: Compute network moments $U_{R_j}[G_b^{*(i)}]$ for each $j \in [m]$.
7: Construct the m-dimensional vector $Y_b^{(i)} = (U_{R_1}[G_b^{*(i)}], \ldots, U_{R_m}[G_b^{*(i)}])$.
8: **end for**
9: **Output:** Edge density $\widehat{\rho}_G$; Set of vectors $\{Y_b^{(i)}\}_{i=1}^{N_{\text{sub}}}$ for downstream inference tasks.

[1]In practical scenarios, n tends to be large in observations, making computation of network moments on $b \geq n$ infeasible, even without advanced inference processes. Hence, we focus on $b < n$.

A crucial aspect of the subsampling approach is its emphasis on computing network moments within networks of size b rather than n during the generation of $Y_b^{(i)}$. Considering that motif counting complexity typically increases superlinearly with network size (Ribeiro and Silva, 2010), this subsampling method emerges as a pivotal technique for addressing scalability in network inference tasks to handle a large network G with the inference of a much smaller b. Additionally, it's worth noting that we keep the specific inference method for downstream tasks open in the algorithm subsequent to obtaining the sample $\{Y_b^{(i)}\}_{i=1}^{N_{\text{sub}}}$. Consequently, the process remains flexible for any inference method the user may be interested in, ranging from intuitive visualization to more sophisticated testing procedures.

The aforementioned subsampling procedure for network moments has been explored in Zhang and Xia (2022); Lunde and Sarkar (2023). However, as mentioned in Section 4.1, these studies primarily focused on inferring a single network moment at a time, specifically concerning the marginal distribution of individual motifs R_j. Yet, relying solely on marginal network moment distributions often falls short of providing insightful inference results for practical problems. Therefore, we proceed to introduce our study of the subsampling correctness of the joint distribution, laying the groundwork for flexible multivariate inference regarding joint or conditional distributions. Such a generalization involves precisely characterizing the dependence between network moments, which is nontrivial from the marginal cases.

To initiate our discussion, Algorithm 5 operates based on the observed network G, and $\{Y_b^{(i)}\}_{i=1}^{N\text{sub}}$ constitutes a random sample from the subsampling distribution conditioning on $\mathbb{G}_n = G$, where $\mathbb{G}_n \sim \rho_n w$. Our objective is to illustrate that the subsampling distribution, as a random probability distribution (with respect to the randomness of $\mathbb{G}_n$), effectively approximates the multivariate network moments distribution for $\mathbb{G}_b$ from the graphon model. We denote $\mathbb{G}_b^{(*G)}$ as a randomly

induced subgraph of G from the node subsampling procedure. When discussing distributional quantities such as the expectation or variance of $\mathbb{G}_b^{(*G)}$, conditioning on $\mathbb{G}_n = G$, we use $(*G)$ in our notation. For instance, $\mathrm{var}(*G)$ represents the variance of $\mathbb{G}_b^{(*G)}$ conditioning on $\mathbb{G}_n = G$. We write $(*G)$ as $*$ when the context clearly identifies G. When we study the asymptotic properties, we are considering a sequence of random networks $\{\mathbb{G}_n\}$, with $n \to \infty$.

Fix a set of motifs $\{R_1, \cdots, R_m\}$, with $r_j = |V(R_j)|$ and $\mathfrak{r}_j = |\mathscr{E}(R_j)|$ for $j \in [m]$, define the following cumulative distribution functions (CDFs):

$$J_{*,n,b}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) = \mathrm{pr}_*\Big\{\sqrt{b}\big[\widehat{\rho}_G^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b^*) - \widehat{\rho}_G^{-\mathfrak{r}_1}U_{R_1}(G)\big] \le t_1, \\ \cdots, \sqrt{b}\big[\widehat{\rho}_G^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b^*) - \widehat{\rho}_G^{-\mathfrak{r}_m}U_{R_m}(G)\big] \le t_m\Big\}, \tag{4.2}$$

$$J_{b,c}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) = \mathrm{pr}\Big\{\sqrt{bc}\big\{\widehat{\rho}_{\mathbb{G}_b}^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b) - E[\rho_b^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b)]\big\} \le t_1, \\ \cdots, \sqrt{bc}\big\{\widehat{\rho}_{\mathbb{G}_b}^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b) - E[\rho_b^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b)]\big\} \le t_m\Big\}. \tag{4.3}$$

We call $J_{*,n,b}^{\{R_1,\cdots,R_m\}}$ the subsampling distribution (conditioning on G) and $J_{b,c}^{\{R_1,\cdots,R_m\}}$ the graphon sampling distribution. The term c will be used to correct the sampling sizes between b and n, the format of which will be clear soon. Our theoretical analysis is established under the following assumptions.

Assumption 1 (Subsampling size)**.** $\lim_{n\to\infty} b/n = c_2$ *for a constant* $c_2 \in [0, 1)$.

Assumption 2 (Sparsity level)**.** *Define* $r = \max\{r_1, \cdots, r_m\}$ *and* $\mathfrak{r} = \max\{\mathfrak{r}_1, \cdots, \mathfrak{r}_m\}$. *There exists a constant* $c_1 > 1$ *such that* $n\rho_n^{4\mathfrak{r}} \geqslant c_1 \log(n)$ *for sufficiently large* n. *Furthermore,* $b\rho_n^{r/2} \to \infty$, *and* $b\rho_n^{2\mathfrak{r}} \to \infty$ *as* $n \to \infty$.

Similar assumptions have been made in Green and Shalizi (2022), Zhang and Xia (2022) and Lunde and Sarkar (2023) when they studied univariate marginal network moment distributions. Zhang and Xia (2022) allows a sparser regime, but their results also require an additional Cramer-type condition to ensure the regularity for network moments. Lunde and Sarkar (2023) has a slightly stronger requirement of $b = o(n)$, and their sparsity assumption is implicit. Under the above assumptions, we have the following property for the multivariate subsampling distribution.

Theorem 9. *Under Assumptions 1 and 2, with probability one (with respect to the random sequence* $\{\mathbb{G}_n\}$*),*

$$\sup_{(t_1,\cdots,t_m)\in\mathbb{R}^m} \left| J_{*,n,b}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) - J_{b,(1-\frac{b}{n})}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) \right| \to 0, \tag{4.4}$$

Theorem 9 is the first result that shows the first-order consistency of the subsampling joint distribution of network moments. For subsampling marginal distributions of network moments, Zhang and Xia (2022) have established the second-order accuracy through their Edgeworth expansion. However, whether such higher-order accuracy is attainable for multivariate joint distributions remains unclear. We defer the exploration of this direction to future endeavors.

4.4 Simulation

We now employ simulated data to assess the accuracy of approximating subsampling distributions by evaluating the finite sample approximation error given by the righthand side of (4.4). Specifically, using networks generated from graphon models, we calculate the empirical Kolmogorov-Smirnov distance between $\widehat{J}_{*,n,b}^{R_1,\cdots,R_m}$ and $\widehat{J}_{b,(1-b/n)}^{R_1,\cdots,R_m}$, which are the empirical CDFs corresponding to (4.2) and (4.3), re-

spectively. We focus on the performance for $m = 1$ (marginal distribution) and 2 (bi-variate joint distribution), considering three basic motifs: $\curlyvee$ (2-star), $\triangle$ (triangle), and $\curlyvee$ (3-star). The experimental setups are detailed below:

- The true network models: two graphons from previous studies (Zhang and Xia, 2022; Green and Shalizi, 2022; Lunde and Sarkar, 2023) are used.

  1. Graphon 1 (smooth): $w(u, v) \propto \exp\{-25(u - v)^2/2\}$.
  2. Graphon 2 (nonsmooth): $w(u, v) \propto 0.5 \cos[0.1\{(u - 0.5)^2 + (v - 0.5)^2\} + 0.01] \cdot \max(u, v)^{2/3} + 0.4$.

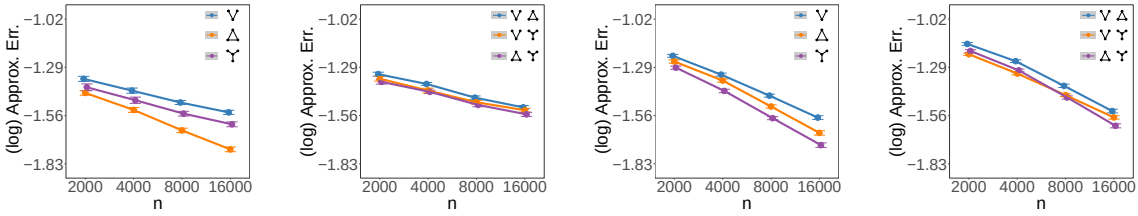
- The network and subsampling sizes: n varies from 2000 to 16000 and $b = \lceil n^{2/3} \rceil$.

- Sparsity levels: Two sparsity levels are considered $\rho_n = 0.25n^{-0.1}$ and $\rho_n = 0.25n^{-0.25}$.

For each configuration, the true CDF is approximated by the empirical CDF from network moments of size-b networks sampled from the actual model. To assess the approximation error (Kolmogorov-Smirnov distance) for each configuration, we generate a size-n network from the true model and use the empirical CDF of the subsampled $\{Y_b^{(i)}\}_{i=1}^{N_{\text{sub}}}$ from Algorithm 5 with $N_{\text{sub}} = 2000$. This method is replicated 50 times, and we report the average approximation error from these replications as the performance metric.

Figure 4.1 displays the log-scale approximation errors for both the marginal and pairwise joint distributions under two graphon models at a sparsity level of $\rho_n = 0.25n^{-0.1}$. Across all evaluated CDFs, there is a clear decreasing trend in errors. With both axis ticks labeled in the log scale, the decreasing trend is close to linear. The error-decreasing rate of the marginal distributions roughly aligns with

the findings of Zhang and Xia (2022). The joint distributions seem to have a slightly slower decrease, but the pattern remains the same. Both graphons (smooth vs. non-smooth) demonstrate consistent decreasing patterns, highlighting the subsampling method's potential robustness to graphon smoothness.

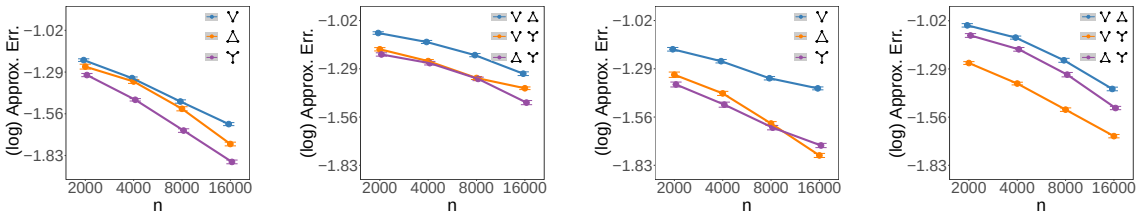


(a) Graphon 1: marginal CDFs **(b)** Graphon 1: joint CDFs **(c)** Graphon 2: marginal CDFs **(d)** Graphon 2: joint CDFs

Figure 4.1: Empirical approximation errors of the CDFs under the sparsity level $\rho_n = 0.25n^{-0.1}$.

Figure 4.2 displays the evaluation under a sparser setting with $\rho_n = 0.25n^{-0.25}$. The pattern remains consistent with previous results, although the errors are slightly higher due to the increased sparsity. The variation in numerical values across different motifs is more pronounced, yet follows the same trend. It's important to note that excessive sparsity can weaken the signal-to-noise ratio to a point where the approximation fails, as known for network resampling methods (Zhang and Xia, 2022; Green and Shalizi, 2022; Lunde and Sarkar, 2023). We explore such an overly sparse scenario in Section C.9.

Additional results for experiments with a subsampling size of $b = \lceil 2n^{1/2} \rceil$ are also available in Section C.9 of the supplementary material.



(a) Graphon 1: marginal CDFs

(b) Graphon 1: joint CDFs

(c) Graphon 2: marginal CDFs

(d) Graphon 2: joint CDFs

Figure 4.2: Empirical approximation errors of the CDFs under the sparsity level $\rho_n = 0.25n^{-0.25}$.

4.5 Statistical application: unmatchable network comparison

4.5.1 Network comparisons based on subsampling distributions

To illustrate subsampling inference, we focus on a key application: network comparison. This involves determining if two network data sets originate from the same underlying population, a question that has gained significant attention recently. For instance, studies like Ghoshdastidar and Von Luxburg (2018); Maugis et al. (2020); Yuan and Wen (2023) have examined how to compare multiple networks within each set, especially when these sets are large. Conversely, research such as Tang et al. (2017a); Li and Li (2018); Liu et al. (2021); Chatterjee et al. (2023); Du and Tang (2023) has explored comparisons between just two networks that share the same nodes, known as "matchable networks." A more complex case arises with "unmatchable" networks, which differ in size and node matchability. Our focus will be on these unmatchable network comparisons. Various methods (Tang et al., 2017b; Agterberg et al., 2020; Alyakin et al., 2024) have been developed for these situations under the random dot product graph model (Young and Scheinerman, 2007). Under the more general graphon model, Ghoshdastidar et al. (2017) and Shao et al. (2022)

have introduced hypothesis testing procedures based on network moments, which can be embedded into resampling methods. However, these methods compare only the marginal distribution of network moments. We now explore the significance of applying multivariate inference to network moments.

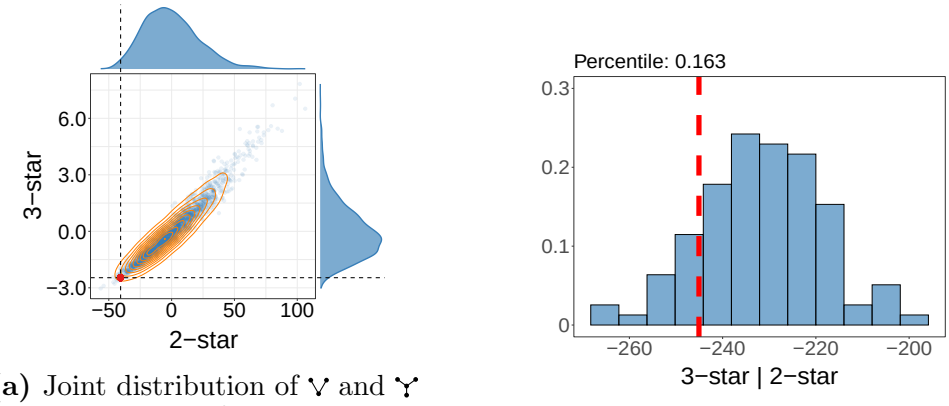
Consider this scenario: two unmatchable networks, G and G', of sizes n and n', are observed instantiations of $\mathbb{G}$ and $\mathbb{G}'$. We model $\mathbb{G} \sim \rho_n w$ and $\mathbb{G}' \sim \rho_{n'} w'$. We aim to determine if $w = w'$ using network moments as multivariate statistics. The strength of subsampling inference is its adaptability to various comparison scenarios. We will illustrate this flexibility with two specific comparison cases. Again, assume we already have a preset of motifs $R_1, \cdots, R_m$ of interest.

Case 1: comparison with highly imbalanced sizes. Suppose $n \gg n'$, and n' is sufficiently large. In this scenario, set $b = n'$. By Theorem 9, we can subsample from G using Algorithm 5 to approximate the true distribution of network moments (4.3) from the graphon w. We then compare whether the observed network moments in G' match this distribution. The comparison step can be done by more rigorous methods such as outlier detection or simple visualization, depending on users' preferences, on either the joint distribution or proper conditional distributions.

Case 2: comparison with comparable sizes. Suppose n and n' are comparable, and both are sufficiently large. In this case, directly subsampling one network based on the size of the other is not effective. We might choose a smaller subsampling size b, less than both n and n'. Using this approach, Algorithm 5 can be applied separately on G and G' to generate two sets of multivariate data: $\{Y_b^{(i)}\}_{i=1}^{N_{\text{sub}}}$ from G and $\{Y_b^{'(i)}\}_{i=1}^{N_{\text{sub}}}$ from G'. According to Theorem 9, these data sets represent random samples from the joint moment distributions of w and w', respectively. This allows for directly comparing the distributions using rigorous hypothesis testing or simple visualization methods.

In the upcoming examples, we will explore both cases mentioned above. Additionally, we'll highlight the value of using multivariate inference of network moments over univariate inference through two examples. In the first example, while the marginal distributions of the network moments show statistically significant differences between two networks, the conditional distribution reveals no differences. In the second example, the marginal distributions of the network moments provide no significant indicators, but the conditional distribution uncovers notable differences between the two networks under comparison.

4.5.2 Guppies gene network comparison for colonization behaviors



(a) Joint distribution of $\curlyvee$ and $\curlyvee$ ($\times 10^2$).

(b) Conditional distribution of $\curlyvee \mid \curlyvee$.

Figure 4.3: Comparison of network moments between two genetic networks: the blue points and bars illustrate the subsampling distributions from the reference gene network while the red point and dotted line denote the observed values in the target gene network.

The study by Fischer et al. (2021) presents a gene expression dataset from guppies in Trinidad and Tobago's Aripo River. The dataset includes a subset of genes, termed target genes, which are believed to be linked with colonization behaviors in

new environments. Our scientific inquiry focuses on how the co-expression patterns of these target genes differ from those of other (reference) genes in the dataset.

The method of Cai and Liu (2016) is used to construct co-expression networks with the FDR control at level 0.05. This results in a large network G between 16,485 reference genes and a smaller G' between 618 target genes. Given their imbalanced sizes, we use the case 1 strategy introduced in the previous section to compare two networks. In particular, N_{sub} is taken to be 2000 for subsampling from the reference gene network, and we use the 2-star (⩛) and 3-star (Y) as the motifs, which intuitively depicting the 3rd-order centralized interactions and the 4th-order centralized interactions. To demonstrate how easy it is to use the subsampling for comparison, we directly use visualization for comparison without resorting to more sophisticated testing procedures.

Figure 4.3a displays 2000 subsampled network moments from the reference gene network, illustrated with fitted contours. The red dot represents the network moment vector of the target gene network. At first glance, it's clear that the target gene network exhibits significantly fewer counts in both 2-stars and 3-stars by their marginal distributions. However, since the 2-star is an induced subgraph of a 3-star, these counts are highly correlated, and the observed differences might simply reflect redundant information. Given that a 3-star represents a higher-order interaction pattern, it is pertinent to examine its conditional distribution based on the 2-star value, particularly when ⩛ matches the value in the current target network (marked by the vertical line at the red point in Figure 4.3a).

This analysis is further detailed in Figure 4.3b. When comparing the conditional distribution of 3-stars to the observed level in the target gene network, we find that the 3-star count in the target network aligns with what is expected from the reference gene network. This suggests that the lower level of 3-stars in the target network can

be fully explained by its reduced 2-star count. After accounting for this, there is no evidence of differences in the higher-order centralized interaction patterns between the networks. This type of analysis provides deeper insights than a mere marginal analysis.

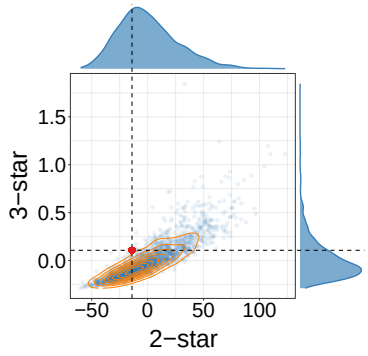
4.5.3 Analysis of collaboration patterns in statistical research

Ji et al. (2022) collected a data set of publications, citations, and collaborations in statistical research based on more than 80,000 papers spanning over 40 years. This data set provides a valuable test base for statistical analysis of text data and network data. In this example, we use the collaboration relations from this data set to analyze the collaboration pattern between statisticians by comparing them with the collaborations in high energy physics Newman (2001) and also dig its temporal variations.

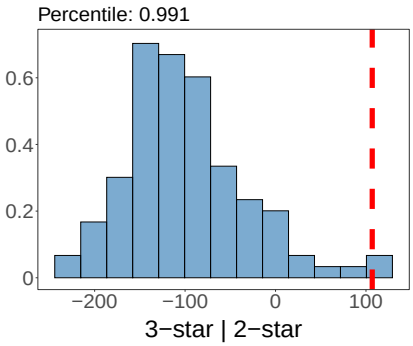
Collaboration comparison between statistics and high energy physics. We focus on the period of 1995–1999 to align with the high energy physics study of Newman (2001). The dataset from Ji et al. (2022) spans a wide range of publications, including various interdisciplinary journals. To concentrate on statistical research, we adopt Ji and Jin (2016)'s approach, limiting our scope to four prominent statistical journals: Annals of Statistics, Biometrika, Journal of the American Statistical Association, and the Journal of the Royal Statistical Society. In the collaboration network, two authors are connected if they have coauthored at least one paper in the data set. The high energy physics collaboration network is already processed by Newman (2001). From both, we extract the largest connected components as is common practice in the literature (Karrer and Newman, 2011; Amini et al., 2013; Miao and Li, 2023), resulting in networks with 750 and 5835 nodes for statistics

and high energy physics, respectively. We aim to compare collaboration patterns in these fields using network moments. This can be done by the strategy outlined in Case 1 of Section 4.5.1, following the method demonstrated in the guppy gene network example.

Figure 4.4 presents the comparison between the statistics and high energy physics networks. The network moments $\curlyvee$ and $\curlyvee$ for the statistics network fall within the expected range when compared to the high energy physics network's marginal distributions. However, a closer look at the joint distribution in Figure 4.4a and the conditional distribution in Figure 4.4b indicates a higher-than-expected number of 3 stars in the statistics network. This suggests a tendency toward more centralized collaboration within the statistics research community as opposed to high energy physics. Such an insightful understanding emerges only from comparing the network moments jointly but not individually.



(a) Joint distribution of $\curlyvee$ and $\curlyvee$ ($\times 10^3$).



(b) Conditional distribution of $\curlyvee \mid \curlyvee$.

Figure 4.4: Comparison of network moments between statistics network and high energy physics network: the blue points and bars illustrate the subsampling distributions from the high energy physics network, while the red point and dotted line denote the observed values in the statistics network.

Temporal comparison of statistics collaborations over time. We next compare the collaboration patterns in statistics over three periods: 1990–1994, 1995–

1999, and 2000–2004. After processing the data, the networks contain 601, 750, and 750 nodes, respectively. There are 186 common nodes between the first two periods and 215 between the last two, with only 68 nodes present throughout all three periods. Due to the small overlap, we consider the three networks unmatchable. With their sizes being similar, we employ the comparison method from Case 2 in Section 4.5.1. For each network, Algorithm 5 is used to generate 2000 subsampled network moments. We then visually compare the three multivariate distributions, focusing on the motifs $\vee$ and $\curlyvee$.

Figure 4.5 displays the three bivariate distributions, showing similar contour shapes and a nonlinear positive relationship between $\vee$ and $\curlyvee$. The distribution for 1990-1994 is noticeably more concentrated, indicating less variability in network moments than in the subsequent periods. The latter two periods show a marked increase in variability yet appear quite similar to each other. Thus, Figure 4.5 implies consistent overall dependence between $\vee$ and $\curlyvee$ across the three periods. From 1990 to 1999, there was an observable growth in the diversity of collaboration patterns, as evidenced by the larger variance. Between 1995 and 2004, however, the pattern of collaboration seems stable regarding these motifs.

4.6 Discussion

We have demonstrated that network node subsampling provides asymptotically valid inference for the joint distributions of multiple network moments. Through multiple examples, we have shown that the joint distribution derived from subsampling offers more utility and deeper insights than the marginal distributions previously studied. Several avenues could extend this work. Notably, examining if a certain type of higher-order accuracy of the joint distribution exists for the node

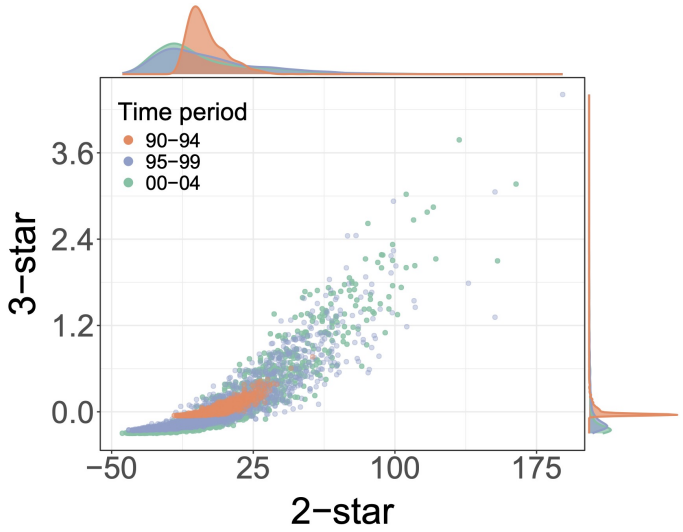


Figure 4.5: Joint distribution of network moments for $\curlyvee$ ($\times 10^3$) and $\curlyvee$ from different statistical collaboration networks.

subsampling is the natural next step. Furthermore, calculating network moments can be computationally intensive, restricting their practical use. Developing efficient methods to calculate network moments in large networks, possibly with some approximations, is a crucial next step. Additionally, how these approximations affect the inferences drawn is another critical problem to study. Advancements here could significantly enhance the scalability of the subsampling inference in network analysis tasks.

Appendices

A Appendix for Chapter II

A.1 Uniqueness of the Overlapping Group Lasso Problem

The group lasso penalization problems (2.14) and (2.15) are generally convex, but may not be strictly convex. The uniqueness of these problems has been studied by Jenatton et al. (2011a). Here we introduce their results for completeness. Note that our theoretical properties in Section 2.3 do not rely on such uniqueness.

Lemma 10. *(see Jenatton et al., 2011a, Proposition 1) If the gram matrix $Q = X^\top X/n$ is invertible, or if there exists $g \in [m]$ such that $G_g = [p]$, then the optimization problem specified in (2.14), with $\lambda_n > 0$, is guaranteed to have a unique solution. The same property holds for problem (2.15) with G replaced by $\mathscr{G}$.*

A.2 Additional Theoretical Results

To begin with, we introduce our proposed upper bound for the dual norm of the overlapping group lasso.

Proposition 1. *The sharp upper bound for ϕ^* (the dual norm of overlapping group lasso penalty in (2.1)) is*

$$\max_{g\in[m]} \frac{1}{w_g} \left\|(H\beta)_{G_g}\right\|_2,$$

where H is a diagonal matrix with diagonals $(\frac{1}{h_1}, \cdots, \frac{1}{h_p})$.

Assumption 7. *Under model (2.13), we assume*

1. *(Sub-Gaussian noises) The coordinates of ε are i.i.d zero mean sub-Gaussian random variable denote with parameter σ, which means that there exist $\sigma > 0$*

such that

$$E[e^{t\varepsilon})] \leqslant \frac{e^{\sigma^2 t^2}}{2}, \text{ for all } t \in \mathbb{R}.$$

2. *(Group normalization condition)* $\sqrt{\gamma_{\max}(X_{G_g}^\top X_{G_g}/n)} \leqslant c$ *for some constant* c.

3. *(Restricted strong convexity condition) For some* $\kappa > 0$,

$$\frac{\left\|X\left(\bar{\beta}-\beta^*\right)\right\|_2^2}{n} \geqslant \kappa\left\|\bar{\beta}-\beta^*\right\|_2^2, \text{ for all } \bar{\beta} \in \left\{\beta \mid \phi\left(\left(\beta-\beta^*\right)_{M^\perp(\overline{S})}\right) \leqslant 3\phi\left(\left(\beta-\beta^*\right)_{M(\overline{S})}\right)\right\}.$$

Remark: The assumption requires an upper bound for the quadratic form associated with each group. This type of assumption is commonly used for developing the upper estimation error bound for non-overlapping group lasso (Lounici et al., 2011; Huang and Zhang, 2010; Dedieu, 2019; Negahban et al., 2012; Wainwright, 2019). Additionally, the restricted curvature conditions have been well discussed by Wainwright (2019). The curvature κ in Assumption 7 is a parameter measuring the convexity. Generally speaking, the restricted curvature conditions state the loss function is locally strongly convex in a neighborhood of ground truth and thus guarantees that a small distance between the estimate and the true parameter implies the closeness in the loss function. However, such a strong convexity condition cannot hold in the high-dimensional setting. So, we focus on a restrictive set of estimates. Restricted curvature conditions are milder than the group-based RIP conditions used in (Huang and Zhang, 2010; Dedieu, 2019), which require that all submatrices up to a certain size are close to isometries (Wainwright, 2019). Based on Assumption 7, Theorem 11 gives ℓ_2 norm estimation upper error bound for overlapping group lasso.

Theorem 11. *Define* $h_{\min}^g = \min_{j \in G_g} h_j$, $d_{\max} = \max_{g \in [m]} d_g$, *and* $\mathscr{d}_{\max} = \max_{g \in [m]} \mathscr{d}_g$. *Suppose Assumption 7 holds, for any* $\delta \in [0,1]$,

1. *with* $\lambda_n = \frac{8c\sigma}{\min_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, *the following bound hold for* $\hat{\beta}^G$ *in* (2.14)

$$\left\|\hat{\beta}^G - \beta^*\right\|_2^2 \lesssim \frac{\sigma^2}{\kappa^2} \cdot \frac{\left(\sum_{g\in\overline{S}} {w_g}^2\right) \cdot h_{\max}^{G_{\overline{S}}}}{\min_{g\in[m]}\left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right). \tag{A.5}$$

with probability at least $1 - e^{-2n\delta}$.

2. *with* $\lambda_n = \frac{8c\sigma}{\min_{g\in[m]} w_g}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, *the following bound hold for* $\hat{\beta}^{\mathcal{G}}$ *in* (2.15)

$$\left\|\hat{\beta}^{\mathcal{G}} - \beta^*\right\|_2^2 \lesssim \frac{\sigma^2}{\kappa^2} \cdot \frac{\sum_{g\in F^{-1}(S)} {w_g}^2}{\min_{g\in[m]}\left(w_g^2\right)} \cdot \left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right). \tag{A.6}$$

Following the framework in (Negahban et al., 2012; Wainwright, 2019), we further study the applicability of the restricted curvature conditions in terms of a random design matrix. Given a group structure G, Theorem 11 is developed based on the assumption that the fixed design matrix X satisfies the restricted curvature condition. In practice, verifying that a given design matrix X satisfies this condition is difficult. Indeed, developing methods to "certify" design matrices this way is one line of ongoing research (Wainwright, 2019). However, it is possible to give high-probability results based on the following assumptions.

Theorem 12. *Under Assumptions 1,2, and 3, we have*

1. *With probability at least* $1 - e^{-c'n}$, $\max_{g\in[m]}\sqrt{\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n})} \leqslant c$ *for some constants* $c, c' > 0$, *as long as* $\log m = o(n)$.

2. *The restricted strong convexity condition, which is*

$$\frac{\left\|X\left(\bar{\beta}-\beta^{*}\right)\right\|_{2}^{2}}{n} \geqslant \kappa\left\|\bar{\beta}-\beta^{*}\right\|_{2}^{2},$$

for all

$$\bar{\beta} \in\left\{\beta \mid \phi\left(\left(\beta-\beta^{*}\right)_{M^{\perp}(\overline{S})}\right) \leqslant 3 \phi\left(\left(\beta-\beta^{*}\right)_{M(\overline{S})}\right)\right\}.$$

hold with probability at least $1-\frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{64}}}$ *for some constant* $\kappa>0$.

A.3 Proofs

A.3.1 Proof of Theorem 1

Lemma 13. *For any norm* $||\cdot_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ *satisfying the conditions in* (2.11)*, the following two statements hold:*

1. *For any* $g \in [m]$*, there exists a* $\tilde{g} \in [|\tilde{G}|]$ *such that* $\mathcal{G}_g \subseteq \tilde{G}_{\tilde{g}}$.

2. *For any* $\tilde{g} \in [|\tilde{G}|]$*, there exists a* $g \in [m]$ *such that* $\tilde{G}_{\tilde{g}} = \mathcal{G}_g$.

Proof. Based on lemma 13, if a norm $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ satisfies (2.11), then it must be that $\tilde{G} = \mathcal{G}$. Consequently, any disparity between $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ and our proposed norm could only be due to differences in weights or the values of q_1 or q_2. Consequently, for any β with non-zero elements solely in the gth group $\mathcal{G}_g$, we have:

$$\sum_{g\in[m]} w_g||\beta_{G_g}||_2 = \sum_{g\in[m]}\Big(\sum_{g\in F(g)} w_g\Big)||\mathcal{G}_g||_2 \leqslant ||\beta_{\{\mathcal{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \sum_{g\in[m]} w_g||\beta_{\mathcal{G}_g}||_2. \quad \text{(A.7)}$$

This implies that $(\tilde{w}_g||\beta_{\mathcal{G}_g}||_{q_2}^{q_1})^{\frac{1}{q_1}} = w_g||\beta_{\mathcal{G}_g}||_2$. By setting one element in $\mathcal{G}_g$ to 1, and other elements to 0, it follows that $\tilde{w}_g = w_g$. Since this holds for any group in $\mathcal{G}$, we have $\tilde{w} = w$.

From (A.7), it is evident that $(w_g||\beta_{\mathcal{G}_g}||_{q_2}^{q_1})^{\frac{1}{q_1}} = w_g||\beta_{\mathcal{G}_g}||_2$ for any β with non-zero elements only in $\mathcal{G}_g$. This suggests that $q_1 = 1$ and $q_2 = 2$. Therefore, the existing norm $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ does not satisfy the second condition in (2.11). □

Proof of Lemma 13

Proof. We begin by proving the first item. Recall that $\mathbb{G}$ represents the space of all possible partitions of $[p]$. Given that $\tilde{G} \in \mathbb{G}$, for an arbitrary $g \in [m]$, suppose

$\mathscr{G}_{\mathscr{g}} \nsubseteq \tilde{G}_{\tilde{g}}$ for any $\tilde{g}$. Then, we can identify the smallest set T such that:

$$\mathscr{G}_{\mathscr{g}} \subseteq \bigcup_{\tilde{g} \in T} \tilde{G}_{\tilde{g}}.$$

Let $T = \{t_1, t_2, \cdots, t_{|T|}\}$. Select one element $\beta_j \in \mathscr{G}_{\mathscr{g}} \cap \tilde{G}_{t_1}$ and another $\beta_k \in \mathscr{G}_{\mathscr{g}} \cap \tilde{G}_{t_2}$. Since both β_j and β_k belong to $\mathscr{G}_{\mathscr{g}}$, if an original group includes β_j, it also contains β_k. Let β be a vector where only β_j and β_k are non-zero, then we have:

$$\begin{aligned}\sum_{g \in [m]} w_g ||\beta_{G_g}||_2 = \big(\sum_{\{g | \beta_j \in G_g\}} w_g \big) \sqrt{\beta_j^2 + \beta_k^2} \leqslant ||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2} \\ \leqslant \sum_{\mathscr{g} \in [m]} \mathscr{w}_{\mathscr{g}} ||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 = \big(\sum_{\{g | \beta_j \in G_g\}} w_g \big) \sqrt{\beta_j^2 + \beta_k^2},\end{aligned}$$

which further leads to

$$||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2} = \Big((\tilde{w}_{t_1} |\beta_j|)^{q_1} + (\tilde{w}_{t_2} |\beta_k|)^{q_1} \Big)^{\frac{1}{q_1}} = \tilde{w}_{t_1}^{\frac{1}{q_1}} |\beta_j| + \tilde{w}_{t_2}^{\frac{1}{q_1}} |\beta_k| = \Big(\sum_{\{g | \beta_j \in G_g\}} w_g \Big) \sqrt{\beta_j^2 + \beta_k^2},$$

for any $0 \leqslant q_1, q_2 \leqslant \infty$. However, by setting

$$\begin{cases} \beta_j = \beta_k = 1, \beta_{\{[p] \setminus \{j,k\}\}} = 0 & \text{if } w_{t_1}^{\frac{1}{q_1}} + w_{t_2}^{\frac{1}{q_1}} \neq \sqrt{2} \left(\sum_{\{g | \beta_j \in G_g\}} w_g \right) \\ \beta_j = 2, \beta_k = 1, \beta_{\{[p] \setminus \{j,k\}\}} = 0 & \text{if } w_{t_1}^{\frac{1}{q_1}} + w_{t_2}^{\frac{1}{q_1}} = \sqrt{2} \left(\sum_{\{g | \beta_j \in G_g\}} w_g \right) \end{cases},$$

we arrive at a contradiction. Thus, we demonstrate that if a norm $|| \cdot_{\{\tilde{G}, w\}} ||_{q_1, q_2}$ exists, then each group in $\tilde{G}$ is a union of groups in $\mathscr{G}$.

Now, we continue to prove the second item. Given that the first part establishes each group in $\tilde{G}$ is a union of groups in $\mathscr{G}$, consider a specific group $\tilde{g} \in [|\tilde{G}|]$. Assume there is an index set $V \subseteq [\mathscr{m}]$ such that $\tilde{G}_{\tilde{g}} = \bigcup_{\mathscr{g} \in V} \mathscr{G}_{\mathscr{g}}$ with $|V| > 1$. Denote $V = \{v_1, \cdots, v_{|V|}\}$. We analyze two scenarios:

- **Case I:** $\nexists a \in [m]$ s.t. $(\mathscr{G}_{v_1} \cup \mathscr{G}_{v_2}) \subseteq G_a$.
- **Case II:** $\exists a \in [m]$ s.t. $(\mathscr{G}_{v_1} \cup \mathscr{G}_{v_2}) \subseteq G_a$.

Under case I, if only $\mathscr{G}_{v_1}$ and $\mathscr{G}_{v_2}$ have non-zero values in β, we obtain:

$$\begin{aligned}\sum_{g\in[m]} w_g||\beta_{G_g}||_2 &= \Big(\sum_{g\in F(v_1)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{v_1}}} + \Big(\sum_{g\in F(v_2)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{v_2}}} \\ &\leqslant ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \sum_{\mathscr{g}\in[m]} \boldsymbol{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 \\ &= \boldsymbol{w}_{v_1}\sqrt{\beta^2_{\mathscr{G}_{v_1}}} + \boldsymbol{w}_{v_2}\sqrt{\beta^2_{\mathscr{G}_{v_2}}} \\ &= \Big(\sum_{g\in F(v_1)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{v_1}}} + \Big(\sum_{g\in F(v_2)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{v_2}}},\end{aligned}$$

which leads to

$$\boldsymbol{w}_{v_1}\sqrt{\beta^2_{\mathscr{G}_{v_1}}} + \boldsymbol{w}_{v_2}\sqrt{\beta^2_{\mathscr{G}_{v_2}}} = \tilde{w}_{\tilde{g}}\Big(\sum_{j\in\tilde{G}_{\tilde{g}}} |\beta_j|^{q_2}\Big)^{\frac{1}{q_2}} = \tilde{w}_{\tilde{g}}\Big(\sum_{j\in\{\mathscr{G}_{v_1}\cup\mathscr{G}_{v_2}\}} |\beta_j|^{q_2}\Big)^{\frac{1}{q_2}}.$$

This equation does not hold by picking $j \in \mathscr{G}_{v_1}, k \in \mathscr{G}_{v_2}$, and setting

$$\begin{cases}\beta_j = \beta_k = 1, \beta_{\{[p]\setminus\{j,k\}\}} = 0 & \text{if } \boldsymbol{w}_{v_1} + \boldsymbol{w}_{v_2} \neq \tilde{w}_{\tilde{g}} \cdot 2^{\frac{1}{q_2}} \\ \beta_j = 2, \beta_k = 1, \beta_{\{[p]\setminus\{j,k\}\}} = 0 & \text{if } \boldsymbol{w}_{v_1} + \boldsymbol{w}_{v_2} = \tilde{w}_{\tilde{g}} \cdot 2^{\frac{1}{q_2}}\end{cases}.$$

Therefore, $|V| > 1$ cannot happen.

Under case II, let $\beta_j \in \mathscr{G}_{v_1}$ and $\beta_k \in \mathscr{G}_{v_2}$. Define β^j as the vector with 1 at the j-th element and 0 elsewhere, and β^k as the vector with 1 at the k-th element and 0 elsewhere, with $j \neq k$.

When $\beta = \beta^j$, we have:

$$\sum_{g\in[m]} w_g||\beta_{G_g}||_2 = \Big(\sum_{g\in F(v_1)} w_g\Big) \leqslant \tilde{w}_{\tilde{g}} \leqslant \sum_{g\in[m]} w_g||\beta_{\mathcal{G}_g}||_2 = w_{v_1},$$

indicating that $\tilde{w}_{\tilde{g}} = w_{v_1}$ for all q_1, q_2. Similarly, for $\beta = \beta^k$, we have:

$$\sum_{g\in[m]} w_g||\beta_{G_g}||_2 = \Big(\sum_{g\in F(v_2)} w_g\Big) \leqslant \tilde{w}_{\tilde{g}} \leqslant \sum_{g\in[m]} w_g||\beta_{\mathcal{G}_g}||_2 = w_{v_2},$$

indicating that $\tilde{w}_{\tilde{g}} = w_{v_2}$ for all q_1, q_2.

If $w_{v_1} \neq w_{v_2}$, then such a weight assignment is not feasible. Assuming $w_{v_1} = w_{v_2} = w_{\tilde{g}} = k$, then for any β with non-zero values only in $\mathcal{G}_{v_1}$, we have $w_{\tilde{g}}||\beta_{\mathcal{G}_{v_1}}||_2 = (w_{\tilde{g}}||\beta_{\mathcal{G}_{v_1}}||_{q_2}^{q_1})^{\frac{1}{q_1}}$, implying that if a norm satisfies (2.11), it must be an ℓ_1/ℓ_2 norm.

Since $\mathcal{G}_{v_1}$ and $\mathcal{G}_{v_2}$ are different groups, there is at least one original group that contains variables in $\mathcal{G}_{v_1}$ but not in $\mathcal{G}_{v_2}$, and vice versa. Taking β with non-zero values in both $\mathcal{G}_{v_1}$ and $\mathcal{G}_{v_2}$, we have:

$$\sum_{g\in[m]} k||\beta_{G_g}||_2 > k||\beta_{\mathcal{G}_{v_1}} \cup \beta_{\mathcal{G}_{v_2}}||_2 = ||\beta_{\{\tilde{G},\tilde{w}\}}||_{1,2},$$

which is a contradiction. Hence, in both cases, $|V| > 1$ is not possible, implying that there exists a $g \in [m]$ such that $\tilde{G}_{\tilde{g}} = \mathcal{G}_g$. □

A.3.2 Proof of Theorem 2

Proof. We begin by examining the bound for the estimator $\hat{\beta}^G$. Considering a fixed design matrix X and a group structure G that comply with Assumption 7, and selecting an appropriate λ_n, Theorem 11 asserts that both inequalities (2.17) and (2.19) hold with a probability of at least $1 - e^{-2n\delta}$.

Under Assumptions 1,2, and 3, Theorem 12 establishes that Assumption 7 is valid with a probability of at least $1 - e^{-c_2 n\delta^2} - \frac{e^{-\frac{n}{32}}}{1-e^{\frac{n}{64}}}$, where c_2 is a positive constant.

Considering these two theorems together, we conclude that under Assumptions 1,2, and 3, both (2.17) and (2.19) are satisfied with a probability of at least $1 - e^{-c_2 n\delta^2} - e^{-2n\delta} - \frac{e^{-\frac{n}{32}}}{1-e^{\frac{n}{64}}}$. This probability can be further bounded below by $1 - e^{-c' n\delta}$ for some suitable constant c'.

The bound for $\hat{\beta}^{\mathscr{G}}$ can be directly derived, noting that it represents a group lasso estimator with group $\mathscr{G}$ and weights $\mathscr{w}$.

□

A.3.3 Proof of Corollary 3

Proof. Assuming that $\max\{d_{\max}, m\} \asymp \max\{\mathscr{d}_{\max}, \mathscr{m}\}$, then we have

$$\Big(\frac{\mathscr{d}_{\max}\log 5}{n} + \frac{\log \mathscr{m}}{n} + \delta\Big) \asymp \Big(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\Big).$$

Let $\mathscr{w}_{\mathscr{g}} = \sum_{g\in F(\mathscr{g})} w_g$, by the Cauchy–Schwarz inequality, we have

$$\mathscr{w}_{\mathscr{g}}^2 = \Big(\sum_{g\in F(\mathscr{g})} w_g\Big)^2 \leqslant \mathscr{h}_{\mathscr{g}}\Big(\sum_{g\in F(\mathscr{g})} w_g^2\Big).$$

Therefore,

$$\sum_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}^2 \leqslant \sum_{\mathscr{g}\in F^{-1}(S)} \mathscr{h}_{\mathscr{g}}\Big(\sum_{g\in F(\mathscr{g})} w_g^2\Big) \leqslant h_{\max}^{G_S}\Big(\sum_{\mathscr{g}\in F^{-1}(S)}\sum_{g\in F(\mathscr{g})} w_g^2\Big).$$

Let's introduce k_g as the number of non-overlapping groups from G into which the gth group is partitioned in the new structure $\mathscr{G}$. We also define K as the maximum number of such partitions, i.e., $K = \max_g k_g$ and $K \leqslant \infty$. Now we want

to show that

$$\sum_{\mathscr{g}\in F^{-1}(S)} \sum_{g\in F(\mathscr{g})} w_g^2 \leqslant \sum_{g\in\overline{S}} k_g w_g^2.$$

Recall the definition of $F^{-1}(S)$ as:

$$F^{-1}(S) = \{\mathscr{g} \mid \mathscr{g} \in F^{-1}(g), g \in S\}.$$

For each $\mathscr{g} \in F^{-1}(g)$ that also belongs to $F^{-1}(S)$, we add w_g^2 to the summation. Therefore, the maximum contribution from each original group g to the sum $\sum_{\mathscr{g}\in F^{-1}(S)} \sum_{g\in F(\mathscr{g})} w_g^2$ is $k_g w_g^2$.

Given that

$$\{g|g \in F(\mathscr{g}) \text{ and } \mathscr{g} \in F^{-1}(S)\} = \overline{S},$$

we have

$$h_{\max}^{G_{\overline{S}}}\Big(\sum_{\mathscr{g}\in F^{-1}(S)} \sum_{g\in F(\mathscr{g})} w_g^2\Big) \leqslant h_{\max}^{G_{\overline{S}}} \sum_{g\in\overline{S}} k_g w_g^2 \leqslant h_{\max}^{G_{\overline{S}}} K \sum_{g\in\overline{S}} w_g^2.$$

On the other hand, we have

$$\begin{aligned}\min_{\mathscr{g}\in[m]}\left(\boldsymbol{w}_{\mathscr{g}}^2\right) &= \min_{\mathscr{g}\in[m]}\Big(\sum_{g\in F(\mathscr{g})} w_g\Big)^2 \geqslant \min_{\mathscr{g}\in[m]}\Big(\sum_{g\in F(\mathscr{g})} \min_{g\in[m]}\{w_g\}\Big)^2 \\ &\geqslant \min_{\mathscr{g}\in[m]}\left(h_{\min}^{g}\min_{g\in[m]}\{w_g\}\right)^2 = \left(h_{\min}^{g}\min_{g\in[m]}\{w_g\}\right)^2 \geqslant \min_{g\in[m]}\left(w_g^2 h_{\min}^{g}\right).\end{aligned}$$

Therefore,

$$\frac{\sum_{\mathscr{g}\in F^{-1}(S)} \boldsymbol{w}_{\mathscr{g}}{}^2}{\min_{\mathscr{g}\in[m]}\left(w_{\mathscr{g}}^2\right)} \leqslant \frac{K\Big(\sum_{g\in\overline{S}} w_g^2\Big)\cdot h_{\max}^{G_{\overline{S}}}}{\min_{g\in[m]}\left(w_g^2 h_{\min}^{g}\right)}.$$

Consequently, if K is upper bounded by a constant, then

$$\frac{\sigma^2}{\kappa^2}\cdot\frac{\sum\limits_{g\in F^{-1}(S)} w_g^2}{\min\limits_{g\in[m]}\left(w_g^2\right)}\cdot\left(\frac{d_{\max}\log 5}{n}+\frac{\log m}{n}+\delta\right)\lesssim\frac{\sigma^2}{\kappa^2}\cdot\frac{\left(\sum\limits_{g\in\overline{S}} w_g^2\right)\cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}\cdot\left(\frac{d_{\max}\log 5}{n}+\frac{\log m}{n}+\delta\right).$$

□

A.3.4 Proof of Proposition 1

Proof. Let H_{G_g} be the sub-matrix of H consisting of the columns indexed by G_g. Let u_{G_g}, v_{G_g} be the sub-vectors of u, v indexed by G_g respectively. Given two vectors $u, v \in \mathbb{R}^p$, we have

$$\begin{aligned}
\phi^*(v) &= \sup_{\phi(u)\leqslant 1}\left\{u^T v\right\} = \sup_{\phi(u)\leqslant 1}\left\{u_1 v_1 + u_2 v_2 + \cdots + u_p v_p\right\} \\
&= \sup_{\phi(u)\leqslant 1}\left\{\frac{v_1}{h_1}\cdot h_1\cdot u_1 + \cdots + \frac{v_p}{h_p}\cdot h_p\cdot u_p\right\} \\
&= \sup_{\phi(u)\leqslant 1}\left\{\sum_{g=1}^m \left(H_{G_g} v_{G_g}\right)^T u_{G_g}\right\} = \sup_{\phi(u)\leqslant 1}\left\{\sum_{g=1}^m \frac{\left((Hv)_{G_g}\right)}{w_g}\cdot w_g\cdot u_{G_g}\right\} \\
&\leqslant \sup_{\phi(u)\leqslant 1}\left\{\sum_{g=1}^m \frac{\left\|(Hv)_{G_g}\right\|_2}{w_g}\cdot\left\|w_g u_{G_g}\right\|_2\right\} \leqslant \left(\max_{g\in[m]}\frac{1}{w_g}\cdot\left\|(Hv)_{G_g}\right\|_2\right)\cdot\phi(u) \\
&\leqslant \max_{g\in[m]}\frac{1}{w_g}\cdot\left\|(Hv)_{G_g}\right\|_2,
\end{aligned}$$

where the first inequality is achieved by using Cauchy's inequality.

Let $g_0 = \arg\max\limits_{g\in[m]} \frac{1}{w_g}\left\|(Hv)_{G_g}\right\|_2$ and $h_{\max}^{g_0} = 1$. Define $u \in \mathbb{R}^p$ as

$$u_j = \begin{cases} 0 \ \textit{for}\ j \notin G_{g_0} \\ \frac{1}{w_{g_0}}\cdot\frac{v_j}{h_j^2}\cdot\frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \ \textit{for}\ j\in G_{g_0}, \end{cases}$$

then we have

$$\begin{aligned}\phi(u) &= \sum_{g=1}^{m} w_g \left\|u_{G_g}\right\|_2 = w_{g_0} \cdot \frac{1}{w_{g_0}} \cdot \frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \cdot \sqrt{\sum_{j \in G_{g_0}} \frac{{v_j}^2}{{h_j}^4}} \\ &= \frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \sqrt{\sum_{j \in G_{g_0}} \frac{{v_j}^2}{{h_j}^2}} = 1,\end{aligned}$$

where the last equality holds due to the fact that $h_j = 1$ for any $j \in G_{g_0}$, and we also have

$$\begin{aligned}u^T v &= \frac{1}{w_{g_0}} \frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \cdot \sum_{j \in G_{g_0}} \frac{{v_j}^2}{{h_j}^2} = \frac{1}{w_{g_0}} \frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \cdot \left\|(Hv)_{G_{g_0}}\right\|_2^2 \\ &= \frac{1}{w_{g_0}} \left\|(Hv)_{G_{g_0}}\right\|_2 = \max_{g \in [m]} \frac{1}{w_{g_0}} \left\|(Hv)_{G_g}\right\|_2 = \phi^*(v).\end{aligned}$$

Therefore, this is a sharp bound.

□

A.3.5 Proof of Theorem 11

Proof. This section mostly follow the proof in Wainwright (2019, Chap. 14). For simplicity, we write $S = S(\beta^*)$ and $\overline{S} = \overline{S}(\beta^*)$. From the optimality of $\hat{\beta}^G$, we have

$$
\begin{aligned}
0 &\geqslant \frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&= \frac{1}{n}\left(Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} - Y^TY + 2Y^TX\beta^* - \beta^{*T}X^TX\beta^*\right) + \lambda_n\left(\phi(\hat{\beta}) - \phi\left(\beta^*\right)\right) \\
&= \frac{1}{n}\left((2X^TX\beta^* - 2X^TY)^T(\hat{\beta} - \beta^*) + (\hat{\beta} - \beta^*)^TX^TX(\hat{\beta} - \beta^*)\right) + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&= \left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right)\right\rangle + \frac{\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2}{n} + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&\geqslant \left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right)\right\rangle + \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&\geqslant -\left|\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right)\right\rangle\right| + \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right),
\end{aligned}
$$

where the penultimate step is valid due to the assumption of restrictive strong convexity.

By applying Holder's inequality with the regularizer ϕ and its dual norm ϕ^*, we have

$$
\left|\left\langle \bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right)\right\rangle\right| \leqslant \phi^*\Big(\bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}\Big)\phi\big(\hat{\beta} - \beta^*\big). \tag{A.8}
$$

Next, we have

$$
\begin{aligned}
\phi(\hat{\beta}) = \phi\left(\beta^* + (\hat{\beta} - \beta^*)\right) &= \phi\left(\beta^*_{M(S)} + \beta^*_{M^\perp(S)} + (\hat{\beta} - \beta^*)_{M(\overline{S})} + (\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) \\
&\geqslant \phi\left(\beta^*_{M(S)} + (\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi(\beta^*_{M^\perp(S)}) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) \\
&= \phi(\beta^*_{M(S)}) + \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi(\beta^*_{M^\perp(S)}) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right).
\end{aligned}
$$

The inequality holds by applying the triangle inequality on $\phi(\hat{\beta})$, and the last

step holds by applying lemma 15. Consequently, we have

$$\begin{aligned}\phi(\hat{\beta}) - \phi\left(\beta^*\right) &\geqslant \phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) - \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right) - 2\phi(\beta^*_{M^{\perp}(S)}) \\ &= \phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) - \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right),\end{aligned} \tag{A.9}$$

where $\phi\left(\beta^*_{M^{\perp}(S)}\right) = 0$ as $\beta^*_{M^{\perp}(S)}$ is a zero vector.

Based on (A.8) and (A.9), we have

$$\begin{aligned}&\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\ &\geqslant -\left|\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta}-\beta^*\right)\right\rangle\right| + \kappa\left\|\left(\hat{\beta}-\beta^*\right)\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\ &\geqslant \kappa\left\|\left(\hat{\beta}-\beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) - \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right)\right) - \left|\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta}-\beta^*\right)\right\rangle\right| \\ &\geqslant \kappa\left\|\left(\hat{\beta}-\beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) - \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right)\right) - \phi^*\left(\bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}\right)\phi\left(\hat{\beta}-\beta^*\right) \\ &\geqslant \kappa\left\|\left(\hat{\beta}-\beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) - \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right)\right) - \frac{\lambda_n}{2}\phi\left(\hat{\beta}-\beta^*\right),\end{aligned}$$

where the last step is valid because lemma 14 implies that we can guarantee $\lambda_n \geqslant 2\phi^*\left(\bigtriangledown\frac{\left\|Y-X\beta^*\right\|_2^2}{n}\right)$ with high probability by taking appropriate λ_n. Moreover, lemma 16 implies that

$$\hat{\beta} \in \left\{\beta \in \mathbb{R}^p \mid \phi\left(\left(\beta-\beta^*\right)_{M^{\perp}(\overline{S})}\right) \leqslant 3\phi\left(\left(\beta-\beta^*\right)_{M(\overline{S})}\right)\right\}.$$

By the triangle inequality, we have

$$\phi(\hat{\beta}-\beta^*) = \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})} + \left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right) \leqslant \phi\left(\left(\hat{\beta}-\beta^*\right)_{M(\overline{S})}\right) + \phi\left(\left(\hat{\beta}-\beta^*\right)_{M^{\perp}(\overline{S})}\right),$$

and hence we have

$$
\begin{aligned}
&\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
\geqslant\ & \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) - \frac{\lambda_n}{2}\phi\left(\hat{\beta} - \beta^*\right) \\
\geqslant\ & \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) \\
&\qquad - \frac{\lambda_n}{2}\left(\phi\left((\hat{\beta} - \beta^*)_{M(S)}\right) + \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right)\right) \\
\geqslant\ & \kappa\left\|\hat{\beta} - \beta^*\right\|_2^2 + \frac{\lambda_n}{2}\left(\phi(\hat{\beta} - \beta^*)_{M^\perp(\overline{S})} - 3\phi(\hat{\beta} - \beta^*)_{M(\overline{S})}\right) \\
\geqslant\ & \kappa\left\|\hat{\beta} - \beta^*\right\|_2^2 - \frac{3\lambda_n}{2}\phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right).
\end{aligned}
$$

By definition, we have $\phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) = \sum_{g\in\overline{S}} w_g\left\|\left(\hat{\beta} - \beta^*\right)_{G_g}\right\|_2$, and by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\sum_{g\in\overline{S}} w_g\left\|\left(\hat{\beta} - \beta^*\right)_{G_g}\right\|_2 &\leqslant \sqrt{\sum_{g\in\overline{S}} w_g^2}\cdot\sqrt{h_{\max}^{G_{\overline{S}}}\cdot\max_{g\in\overline{S}}\left\|\left(\hat{\beta} - \beta^*\right)_{G_g}\right\|_2^2} \\
&\leqslant \sqrt{\sum_{g\in\overline{S}} w_g^2}\cdot\sqrt{h_{\max}^{G_{\overline{S}}}\cdot\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2} \\
&= \sqrt{\sum_{g\in\overline{S}} w_g^2}\cdot\sqrt{h_{\max}^{G_{\overline{S}}}}\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2.
\end{aligned}
$$

On the other hand, since $\kappa\left\|\hat{\beta} - \beta^*\right\|_2^2 - \frac{3\lambda_n}{2}\sqrt{\sum_{g\in\overline{S}} w_g^2}\cdot\sqrt{h_{\max}^{G_{\overline{S}}}}\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2 \leqslant 0$,

we have

$$\begin{aligned}\left\|\hat{\beta}-\beta^*\right\|_2^2 &\leqslant \frac{9\lambda_n^2}{4\kappa^2}\sum_{g\in\overline{S}} w_g^2\cdot h_{\max}^{G_{\overline{S}}}\\ &\leqslant \frac{9}{4\kappa^2}\cdot\frac{64c^2\sigma^2\sum\limits_{g\in\overline{S}} w_g^2\cdot h_{\max}(\overline{S})}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}\cdot\left(\frac{d_{\max}\log 5}{n}+\frac{\log m}{n}+\delta\right)\\ &\leqslant \frac{144c^2\sigma^2}{\kappa^2}\cdot\frac{\sum\limits_{g\in\overline{S}} w_g^2\cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}\cdot\left(\frac{d_{\max}\log 5}{n}+\frac{\log m}{n}+\delta\right)\end{aligned}$$

□

A.3.6 Lemmas for the proof of Theorem 11

In these lemmas, we abbreviate $\hat{\beta}^G$ by $\hat{\beta}$.

Lemma 14. *Under the Assumption 7 and (2.2), taking*

$$\lambda_n=\frac{8c\sigma}{\sqrt{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}}\sqrt{\frac{d_{\max}\log 5}{n}+\frac{\log m}{n}+\delta}\quad \textit{for some } \delta\in[0,1],$$

then $P\left(\lambda_n\geqslant 2\phi^*(\frac{X^\top\varepsilon}{n})\right)\geqslant 1-e^{-2n\delta}$.

Proof of Lemma 14. Let $V_{i\cdot g}=-\varepsilon_i\left(\frac{X_{ig_1}}{h_{g_1}w_g},\frac{X_{ig_2}}{h_{g_2}w_g},\ldots,\frac{X_{ig_{d_g}}}{h_{g_{d_g}}w_g}\right)\in\mathbb{R}^{d_g}$. According to the variational form of ℓ_2 norm, we have $\frac{1}{n}\left\|\sum_{i=1}^n V_{i\cdot g}\right\|_2=\sup\limits_{u\in S^{d_g-1}}\left\langle u,\frac{1}{n}\sum_{i=1}^n V_{i\cdot g}\right\rangle$, where S^{d_g-1} is the Euclidean sphere in $\mathbb{R}^{d_g}$. Also, for any vector $u\in S^{d_g-1}$ and $t\in\mathbb{R}$, we have

$$\begin{aligned}\frac{1}{n}\log\mathbb{E}\left(e^{t\left\langle u,\sum\limits_{i=1}^n V_{i\cdot g}\right\rangle}\right)&=\frac{1}{n}\log\mathbb{E}\left(e^{t\sum\limits_{j=1}^{d_g}u_j\sum\limits_{i=1}^n V_{i\cdot g_j}}\right)=\frac{1}{n}\log\mathbb{E}\left(e^{t\sum\limits_{i=1}^n\left(\sum\limits_{j=1}^{d_g}u_j V_{i\cdot g_j}\right)}\right)\\ &=\frac{1}{n}\log\mathbb{E}\left(e^{-t\sum\limits_{i=1}^n\left(\sum\limits_{g=1}^{d_g}\frac{u_j X_{ig_j}\varepsilon_i}{h_{g_j}w_g}\right)}\right)=\frac{1}{n}\log\mathbb{E}\left(e^{-t\sum\limits_{i=1}^n\varepsilon_i\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{ig_j}}{h_{g_j}w_g}\right)}\right).\end{aligned}$$

Since $\{\epsilon_i\}_{i=1}^n$ are i.i.d zero mean sub-Gaussian random variables with parameter σ, let $u = (u_1, \cdots, u_{d_g})^T \in \mathbb{R}^{d_g \times 1}$, $X_{i,g} = (X_{ig_1}, \cdots X_{ig_{d_g}})^T \in \mathbb{R}^{d_g \times 1}$, then we have

$$\begin{aligned}
\frac{1}{n}\log\mathbb{E}\left(e^{-t\sum\limits_{i=1}^{n}\varepsilon_i\left(\sum\limits_{j=1}^{d_g}\frac{u_j x_{ig_j}}{h_{g_j}w_g}\right)}\right) &= \frac{1}{n}\log\mathbb{E}\left(e^{-t\varepsilon_1\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{1g_j}}{h_{g_j}w_g}\right)}\right) + \cdots + \frac{1}{n}\log\mathbb{E}\left(e^{-t\varepsilon_n\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{ng_j}}{h_{g_j}w_g}\right)}\right) \\
&\leqslant \frac{t^2\sigma^2}{2n}\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{d_g}\frac{u_j X_{ig_j}}{w_g h_{g_j}}\right)^2\right) \leqslant \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{d_g}u_j X_{ig_j}\right)^2\right) \\
&= \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(\sum_{i=1}^{n}\langle u,\ X_{i,g}\rangle^2\right) = \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(\sum_{i=1}^{n}(u^T X_{i,g}X_{i,g}^T u)\right) \\
&= \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(u^T\left(\frac{1}{n}\sum_{i=1}^{n}X_{i,g}X_{i,g}^T\right)u\right) \\
&= \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(u^T\frac{X_{G_g}^T X_{G_g}}{n}u\right) \\
&\leqslant \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\left(\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n})\right)
\end{aligned}$$

By Assumption 7, we have $\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n}) \leqslant c^2$. Combining this with the previous proof, we have $\frac{1}{n}\log\mathbb{E}\left(e^{t\left\langle u,\sum\limits_{i=1}^{n}V_{i\cdot g}\right\rangle}\right) \leqslant \frac{c^2t^2\sigma^2}{2w_g^2\left(h_{\min}^g\right)}$. Therefore, the random variable $\left\langle u, \sum_{i=1}^{n} V_{i\cdot g}\right\rangle$ is the sub-Gaussian with the parameter at most $\sqrt{\frac{c^2\sigma^2}{w_g^2\left(h_{\min}^g\right)}}$, and by properties of sub-Gaussian variables, we have

$$\log\mathbb{P}\left(\left\langle u, \sum_{i=1}^{n}V_{i\cdot g}\right\rangle \geqslant \frac{\lambda_n}{4}\right) \leqslant -\frac{\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2}.$$

We can find a $\frac{1}{2}$ covering of S^{d_g-1} in Euclidean norm:$\{u^1, u^2, \ldots, u^N\}$ with $N \leq 5^{d_g}$, recall that $\frac{1}{n}\left\|\sum_{i=1}^{n}V_{i\cdot g}\right\|_2 = \frac{1}{n}\sup\limits_{u\in S^{d_g-1}}\left\langle u, \sum_{i=1}^{n}V_{i\cdot g}\right\rangle$, so that for any $u \in S^{d_g-1}$,

we can find a $u^{q(u)} \in \{u^1, \ldots, u^N\}$, such that $\|u^{q(u)} - u\|_2 \leqslant \frac{1}{2}$, and

$$\begin{aligned}\frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle &= \frac{1}{n}\sup_{u\in S^{d_g-1}}\Big(\Big\langle u-u^{q(u)}, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle + \Big\langle u^{q(u)}, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle\Big)\\ &\leqslant \frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u-u^{q(u)}, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle + \frac{1}{n}\max_{q\in[N]}\Big\langle u^q, V_{i\cdot g}\Big\rangle\end{aligned}$$

By applying the Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u-u^{q(u)}, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle \leqslant \frac{\|u-u^{q(u)}\|_2}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \leqslant \frac{1}{2n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2.$$

Hence, we obtain $\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \leqslant \frac{1}{2n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 + \frac{1}{n}\max_{q\in[N]}\Big\langle u^q, \sum_{i=1}^{n} V_{i\cdot g}\Big\rangle$, which indicates that

$$\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \leqslant 2\max_{q\in[N]}\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle.$$

Consequently, we can express the probability as

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big) &\leqslant \mathbb{P}\Big(\max_{q\in[N]}\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle \geqslant \frac{\lambda_n}{4}\Big)\\ &\leqslant \sum_{q=1}^{N}\mathbb{P}\Big(\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle \geqslant \frac{\lambda_n}{4}\Big)\\ &\leqslant N\exp\Big(-\frac{n\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2}\Big) \leqslant \exp\Big(-\frac{n\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2} + d_g\log 5\Big),\end{aligned}$$

and by setting $\lambda_n = \frac{8C\sigma}{\sqrt{\min_{g\in[m]}(w_g^2 h_{\min}^g)}}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, we get

$$\begin{aligned}
\mathbb{P}\Big(\max_{g\in[m]}\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big) &\leqslant \sum_{g=1}^{m}\mathbb{P}\Big(\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big) \\
&\leqslant \exp\Big(-\frac{n\lambda_n^2}{32C^2\sigma^2}\min_{g\in[m]}(w_g^2 h_{\min}^g) + d_{\max}\log 5 + \log m\Big) \\
&\leqslant \exp\{-2n\delta\}.
\end{aligned}$$

From Proposition 1, we have

$$\phi^*\Big(\frac{X^\top\varepsilon}{n}\Big) \leqslant \max_{g\in[m]}\frac{1}{w_g}\Big\|\Big(\frac{HX^\top\varepsilon}{n}\Big)_{G_g}\Big\|_2 = \max_{g\in[m]}\frac{1}{w_g}\Big\|\frac{1}{n}\sum_{i=1}^{n} -\varepsilon_i\Big(\frac{X_{ig_1}}{h_{g_1}},\cdots,\frac{X_{ig_{d_g}}}{h_{g_{d_g}}}\Big)\Big\|_2 = \max_{g\in[m]}\Big\|\frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2.$$

Therefore, $P\left(\lambda_n \geqslant 2\phi^*(\frac{X^\top\varepsilon}{n})\right) \geqslant 1 - e^{-2n\delta}$.

□

Lemma 15. *The group lasso regularizer (2.1) is decomposable with respect to the pair* $\{M(S), M^\perp(\overline{S})\}$. *That is,* $\phi(a+b) = \phi(a)+\phi(b)$, *for all* $a \in M(S)$ *and for all* $b \in M^\perp(\overline{S})$.

Proof of Lemma 15.

$$\begin{aligned}
\phi(a+b) &= \sum_{g=1}^{m} w_g\left\|(a+b)_{G_g}\right\|_2 = \sum_{g\in M(\overline{S})} w_g\left\|(a+b)_{G_g}\right\|_2 + \sum_{g\notin M(\overline{S})} w_g\left\|(a+b)_{G_g}\right\|_2 \\
&= \sum_{g\in M(\overline{S})} w_g\left\|a_{G_g}\right\|_2 + \sum_{g\in M^\perp(\overline{S})} w_g\left\|b_{G_g}\right\|_2 = \sum_{g\in M(S)} w_g\left\|a_{G_g}\right\|_2 + \sum_{g\in M^\perp(\overline{S})} w_g\left\|b_{G_g}\right\|_2 \\
&= \phi(a) + \phi(b)
\end{aligned}$$

□

Lemma 16. *If* $\lambda_n \geqslant 2\phi^*\left(\frac{X^T\varepsilon}{n}\right)$, *then* $\phi\left((\hat{\beta}-\beta^*)_{M^\perp(\overline{S})}\right) \leqslant 3\phi\left((\hat{\beta}-\beta^*)_{M(\overline{S})}\right)$.

Proof of Lemma 16 (also see proposition 9.13 in Wainwright (2019). From equation (A.9), we have

$$\phi(\hat{\beta}) - \phi\left(\beta^*\right) \geqslant \phi\left((\hat{\beta} - \beta^*)_{M^{\perp}(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right),$$

On the other hand, by the convexity of the cost function, we have

$$\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 \geqslant \left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle \geqslant -\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle.$$

By applying Holder's inequality with the regularizer ϕ and its dual norm ϕ^*, we have

$$\left|\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle\right| \leqslant \phi^*\left(\bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}\right)\phi\left(\hat{\beta} - \beta^*\right).$$

Therefore,

$$\begin{aligned}\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 &\geqslant -\left\langle \bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle \geqslant -\phi^*\left(\bigtriangledown \frac{\left\|Y - X\beta^*\right\|_2^2}{n}\right)\phi\left(\hat{\beta} - \beta^*\right) \\ &\geqslant -\frac{\lambda_n}{2}\phi\left(\hat{\beta} - \beta^*\right) \geqslant -\frac{\lambda_n}{2}\left(\phi(\hat{\beta} - \beta^*)_{M(\overline{S})} + \phi(\hat{\beta} - \beta^*)_{M^{\perp}(\overline{S})}\right),\end{aligned}$$

and

$$\begin{aligned}0 &= \frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\ &\geq \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^{\perp}(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) - 2\phi(\beta^*_{M^{\perp}(S)})\right) - \frac{\lambda_n}{2}\left(\phi(\hat{\beta} - \beta^*)_{M(\overline{S})} + \phi(\hat{\beta} - \beta^*)_{M^{\perp}(\overline{S})}\right) \\ &= \frac{\lambda_n}{2}\left(\phi\left((\hat{\beta} - \beta^*)_{M^{\perp}(\overline{S})}\right) - 3\phi\left(\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right),\end{aligned}$$

from which the claim follows. □

A.3.7 Proof of Theorem 12

Two lemmas are used in this proof:

Lemma 17 (Theorem 6.5 in (Wainwright, 2019)). *Let $|||.|||_2$ be the spectral norm of a matrix. There are universal constants c_2, c_3, c_4, c_5 such that, for any matrix $A \in \mathbb{R}^{n\times p}$, if all rows are drawn i.i.d from $N(0,\Theta)$, then the sample covariance matrix $\hat{\Theta}$ satisfies the bound*

$$\mathbb{E}\left(e^{t|||\hat{\Theta}-\Theta|||_2}\right) \leqslant e^{c_3\frac{t^2\theta^2}{n}+4p} \quad \textit{for all } |t| < \frac{n}{64e^2|||\Theta|||_2},$$

and hence for all $\delta \in [0,1]$

$$\mathbb{P}\Big(\frac{|||\hat{\Theta}-\Theta|||_2}{|||\Theta|||_2} \leqslant c_5(\sqrt{\frac{p}{n}}+\frac{p}{n})+\delta\Big) > 1-c_4e^{-c_2n\delta^2} \tag{A.10}$$

Lemma 18. *Under Assumptions 1,2, and 3, and use $\rho(\Theta)$ to denote the maximum diagonal of a covariance matrix Θ. For any vector $\beta \in \mathbb{R}^p$ and a given group structure with m groups, we have*

$$\frac{\|X\beta\|_2}{\sqrt{n}} \geq \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta\right\|_2 - 8\rho(\Theta)\left(\max_{g\in[m]}\frac{1}{w_g\sqrt{h^g_{\min}}}\right)\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\phi(\beta), \tag{A.11}$$

with probability at least $1-\frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{64}}}$.

Proof. We first prove the first part of Theorem 12. By Lemma 17, we have

$$\mathbb{P}\Big(\frac{|||\frac{X_{G_g}^TX_{G_g}}{n}-\Theta_{G_g,G_g}|||_2}{|||\Theta_{G_g,G_g}|||_2} \leqslant c_5(\sqrt{\frac{d_g}{n}}+\frac{d_g}{n})+\delta\Big) > 1-c_4e^{-c_2n\delta^2}$$

By the triangle inequality, since $X_{G_g}^T X_{G_g}$ is a positive semi-definite, we have

$$\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n}) = |||\frac{X_{G_g}^T X_{G_g}}{n}|||_2 = |||\frac{X_{G_g}^T X_{G_g}}{n} - \Theta_{G_g,G_g}|||_2 + |||\Theta_{G_g,G_g}|||_2$$
$$\leqslant (1 + c_5(\sqrt{\frac{d_g}{n}} + \frac{d_g}{n}) + \delta)|||\Theta_{G_g,G_g}|||_2,$$

with probability at least $1 - c_4 e^{-c_2 n\delta^2}$. Because $|||\Theta_{G_g,G_g}|||_2 \leq |||\Theta|||_2 \leqslant c_1$ for some constant c_1 and $d_g \leqslant n$, we have $\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n}) \leqslant c + \delta$ for some constant c, with probability at least $1 - e^{-c_2 n\delta^2}$. Taking the union probability for all m groups, we have

$$\max_{g\in[m]} \gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n}) \leq c + \delta$$

with probability at least $1 - \exp(-c' 2n\delta^2)$ for some constant $c' > 0$ as long as

$$\log m \ll n\delta^2.$$

For simplicity, we take δ as a constant.

Now we proceed to prove the second part. First note that we must have $\rho(\Theta) \leq \gamma_{\max}(\Theta) \leq c_1$ by Assumptions 1,2, and 3. By applying Minkowski inequality, we have

$$\phi(\beta) = \sum_{g=1}^{m} w_g \left\|\beta_{G_g}\right\|_2 \leqslant \sqrt{m}\sqrt{\sum_{g=1}^{m} w_g^2 \left\|\beta_{G_g}\right\|_2^2} \leqslant \sqrt{m}\sqrt{\max_{g\in[m]} w_g^2 h_{\max}^g \left\|\beta\right\|_2^2}$$

Let $\beta = \beta^* - \bar{\beta}$, we now want to prove that $\phi\left(\beta_{M^\perp(\bar{S})}\right) \leqslant 3\phi\left(\beta_{M(\bar{S})}\right)$ implies $\frac{\|X\beta\|_2^2}{n} \geqslant \frac{\gamma_{\min}}{64}\left\|\beta\right\|_2^2$.

Since $\phi\left(\beta_{M^{\perp}(\bar{S})}\right) \leqslant 3\phi\left(\beta_{M(\bar{S})}\right)$, combining with triangle inequality, we have

$$\begin{aligned}\phi(\beta) = \phi\left(\beta_{M(\bar{S})}\right) + \phi\left(\beta_{M^{\perp}(\bar{S})}\right) &\leqslant 4\phi\left(\beta_{M(\bar{S})}\right) \leqslant 4\sqrt{\bar{s}_g}\sqrt{\max_{g\in\bar{S}} w_g^2 h_{\max}^g}\left\|\beta_{M(\bar{S})}\right\|_2 \\ &\leqslant 4\sqrt{\bar{s}_g}\sqrt{\max_{g\in\bar{S}} w_g^2 h_{\max}^g}\left\|\beta\right\|_2\end{aligned}$$

From Lemma 18, we have

$$\begin{aligned}\frac{\|X\beta\|_2}{\sqrt{n}} \geqslant& \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta\right\|_2 - 8\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\phi(\beta) \\ \geqslant& \frac{1}{4\sqrt{c_1}}\|\beta\|_2 - 32\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\sqrt{\bar{s}_g}\sqrt{\max_{g\in\bar{S}} w_g^2 h_{\max}^g}\|\beta\|_2 \\ \geqslant& \frac{1}{64\sqrt{c_1}}\|\beta\|_2,\end{aligned}$$

where the last step is valid due to Assumption 2 and 3. □

Lemmas for the Proof of Theorem 12

Proof. **of Lemma 18** To begin with, for a vector $\beta \in \mathbb{R}^p$ with a fixed group structure, we define the set:

$$S^{p-1}(\Theta) = \left\{\beta \in \mathbb{R}^p \middle| \left\|\Theta^{\frac{1}{2}}\beta\right\|_2 = 1\right\},$$

the function:

$$g(t) = 4\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\cdot t$$

and the event:

$$\mathscr{E}\left(S^{p-1}(\Theta)\right) = \left\{X \in \mathbb{R}^{n\times p}\middle| \inf_{\beta\in S^{p-1}(\Theta)}\frac{\|X\beta\|_2}{\sqrt{n}} + 2g(\phi(\beta)) \leqslant \frac{1}{4}\right\},$$

where $\phi(.)$ is the overlapping group lasso regularizer. In addition, given $0 \leqslant r_\ell \leqslant r_u$, we define the set

$$\mathbb{K}(r_\ell, r_u) = \left\{\beta \in S^{p-1}(\Theta) \middle| g(\phi(\beta)) \in [r_\ell, r_u]\right\},$$

and the event:

$$\mathscr{A}(r_\ell, r_u) = \left\{X \in \mathbb{R}^{n\times p} \middle| \inf_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{2} - r_u\right\}.$$

Now we introduce two additional lemmas:

Lemma 19. *For $\upsilon = \frac{1}{4}$, we have $\mathscr{E} \subseteq \mathscr{A}(0,\upsilon) \cup \left(\bigcup_{\ell=1}^{\infty} \mathscr{A}\left(2^{\ell-1}\upsilon, 2^\ell\upsilon\right)\right)$.*

Lemma 20. *For any pair (r_ℓ, r_u), where $0 \leqslant r_\ell \leqslant r_u$, we have $\mathbb{P}(\mathscr{A}(r_\ell, r_u)) \leqslant e^{-\frac{n}{32}} e^{-\frac{n}{8}r_u^2}$.*

Based on Lemma 19 and Lemma 20, we have

$$\mathbb{P}(X \in \mathscr{E}) \leqslant \mathbb{P}(\mathscr{A}(0,\upsilon)) + \sum_{\ell=1}^{\infty} \mathbb{P}\left(\mathscr{A}(2^{\ell-1}\upsilon, 2^\ell\upsilon)\right) \leqslant e^{-\frac{n}{32}} \left\{\sum_{t=0}^{\infty} e^{-\frac{n}{8}2^{2\ell}\upsilon^2}\right\}.$$

Since $\upsilon = \frac{1}{4}$ and $2^{2\ell} \geqslant 2\ell$, we have

$$\mathbb{P}(X \in \mathscr{E}) \leqslant e^{-\frac{n}{32}} \sum_{\ell=0}^{\infty} e^{-\frac{n}{8}2^{2\ell}\upsilon^2} \leqslant e^{-\frac{n}{32}} \sum_{\ell=0}^{\infty} e^{-n\frac{\ell}{4}\upsilon^2} \leqslant \frac{e^{-\frac{n}{32}}}{1 - e^{-\frac{n}{64}}}.$$

We now get the upper bound of $\mathbb{P}(X \in \mathscr{E})$. We next show that the bound in (A.11) always hold on the complementary set $\mathscr{E}^c$. If $X \notin \mathscr{E}$, based on the definition of $\mathscr{E}$, we have $\inf_{\beta\in S^{p-1}(\Theta)} \frac{|X\beta\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g(\phi(\beta))$. That is $\forall\beta \in S^{p-1}(\Theta)$.

$\frac{\|X\beta\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g\left(\phi(\beta)\right)$. Therefore, for any $\beta' \in \{\beta' \in \mathbb{R} | \frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2} \in S^{p-1}(\Theta)\}$, we have

$$\frac{\left\|X\frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2}\right\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g\left(\phi\left(\frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2}\right)\right)$$
$$\frac{\left\|X\beta'\right\|_2}{\sqrt{n}} \geqslant \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2 - 2g\Big(\phi(\beta')\Big),$$

where we finish the proof by substituting the definition of $g(\phi(\beta))$. □

Proof. **of Lemma 19** By definition, $\mathbb{K}(0,\upsilon) \cup \left(\bigcup_{\ell=1}^{\infty} \mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)\right)$ is a cover of $S^{p-1}(\Theta)$. Therefore, for any β, it either belongs to $\mathbb{K}(0,\upsilon)$ or $\mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$, which leads to the following two cases:

Case 1 If $\beta \in \mathbb{K}(0,\upsilon)$, by definition, we have $g\left(\phi(\beta)\right) \in [0,\upsilon]$ and

$$\frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{4} - 2g\left(\phi(\beta)\right) \leqslant \frac{1}{4} = \frac{1}{2} - \upsilon.$$

Therefore, the event $\mathscr{A}(0,\upsilon)$ must happen in this case.

Case 2: If $\beta \notin \mathbb{K}(0,\upsilon)$, we must have $\beta \in \mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$ for some $\ell = 1, 2, \cdots$, and moreover

$$\frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{4} - 2g\left(\phi(\beta)\right) \leqslant \frac{1}{4} - 2 \cdot \left(2^{\ell-1}\upsilon\right) \leqslant \frac{1}{2} - \left(2 \cdot 2^{\ell-1}\right)\upsilon \leqslant \frac{1}{2} - 2^{\ell}\upsilon.$$

So that the event $\mathscr{A}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$ must happen. Therefore, $\mathscr{E} \subseteq \mathscr{A}(0,\upsilon) \cup \left(\bigcup_{\ell=1}^{\infty} \mathscr{A}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)\right)$.

□

Proof. **of Lemma 20** To prove Lemma 20, we first introduce the following lemmas:

Lemma 21 (Gordon's Inequality)**.** *Let* $\{Z_{u,v}\}_{u\in U, v\in V}$ *and* $\{Y_{u,v}\}_{u\in U, v\in V}$ *be zero-mean Gaussian process indexed by a non-empty index set* $I = U \times V$*. If*

1. $\mathbb{E}\left((Z_{u,v} - Z_{u'v'})^2\right) \leq \mathbb{E}\left((Y_{u,v} - Y_{u',v'})^2\right)$ *for all pairs* (u, v) and $(u'\, v') \in I$.

2. $\mathbb{E}\left((Z_{u,v} - Z_{u'v})^2\right) = \mathbb{E}\left((Y_{u,v} - Y_{u',v})^2\right)$,

then we have $\mathbb{E}(\max_{v\in V} \min_{u\in U} Z_{u,v}) \leq \mathbb{E}(\max_{v\in V} \min_{u\in U} Y_{u,v})$.

Lemma 22. *Suppose that* $\alpha = (\alpha_1, ..., \alpha_d)$, *where each* $\alpha_i, i \in [d]$ *is a zero-mean sub-Gaussian random variable with parameter at most* σ^2, *then for any* $t \in \mathbb{R}$, *we have* $\mathbb{E}\left(\exp\left(t \left\|\alpha\right\|_2\right)\right) \leqslant 5^d \exp\left(2t^2\sigma^2\right)$.

Lemma 23. *Suppose that* $\alpha = (\alpha_1, ..., \alpha_d)$, *where each* $\alpha_i, i \in [d]$ *is a zero-mean sub-Gaussian random variable with parameter at most* σ^2, *and for a given group structure* G, *let* $\left\|\alpha_{G_g}\right\|$ *be the corresponding group norm,* m *be the number of groups and* d_{max} *be the maximum group size, then*

$$\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant 2\sqrt{2\sigma^2\left(\log m + d_{max}\log 5\right)}.$$

Lemma 24 (Theorem 2.26 in (Wainwright, 2019))**.** *Let* $x = (x_1, \cdots, x_n)$ *be a vector of i.i.d standard Gaussian variable, and* $f : \mathbb{R}^n \to \mathbb{R}$ *be a L-Lipschitz, with respect to the Euclidean norm, then* $f(x) - \mathbb{E}f(x)$ *is sub-Gaussian with parameter at most* L, *and hence* $\mathbb{P}\left(\left(f(x) - \mathbb{E}\left[f(x)\right)\right) \geqslant t\right] \leqslant e^{-\frac{t^2}{2L^2}}$, $\forall t \geqslant 0$.

We now start to prove. First, we define and bound the random variable $T\left(r_\ell, r_u\right) = -\inf_{\beta\in\mathbb{K}(r_\ell, r_u)} \frac{\|X\beta\|_2}{\sqrt{n}}$. Let S^{n-1} be a unit ball on $\mathbb{R}^n$, by the variational representation of the ℓ_2-norm, we have

$$T\left(r_\ell, r_u\right) = -\inf_{\beta\in\mathbb{K}(r_\ell, r_u)} \frac{\|X\beta\|_2}{\sqrt{n}} = -\inf_{\beta\in\mathbb{K}(r_\ell, r_u)} \sup_{u\in S^{n-1}} \frac{\langle u, X\beta\rangle}{\sqrt{n}} = \sup_{\beta\in\mathbb{K}(r_\ell, r_u)} \inf_{u\in S^{n-1}} \frac{\langle u, X\beta\rangle}{\sqrt{n}}.$$

Let $X = W\Theta^{\frac{1}{2}}$, where $W \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix, and define the

transformed vector $v = \Theta^{\frac{1}{2}}\beta$, then

$$T\left(r_\ell, r_u\right) = \sup_{\beta \in \mathbb{K}(r_\ell, r_u)} \inf_{u \in S^{n-1}} \frac{\langle u, X\beta \rangle}{\sqrt{n}} = \sup_{v \in \bar{\mathbb{K}}(r_\ell, r_u)} \inf_{u \in S^{n-1}} \frac{\langle u, Wv \rangle}{\sqrt{n}},$$

where $\bar{\mathbb{K}}\left(r_\ell, r_u\right) = \left\{ v \in \mathbb{R}^p \middle| \|v\|_2 = 1, g\left(\phi(\Theta^{-\frac{1}{2}}v)\right) \in [r_\ell, r_u] \right\}$.

Define $Z_{u,v} = \frac{\langle u, Wv \rangle}{\sqrt{n}}$, since (u, v) range over a subset of $S^{n-1} \times S^{p-1}$, each variable $Z_{u,v}$ is zero-mean Gaussian with variance n^{-1}. We compare the Gaussian process $Z_{u,v}$ to the zero-mean Gaussian process $Y_{u,v}$ which defined as:

$$Y_{u,v} = \frac{\langle \zeta, u \rangle}{\sqrt{n}} + \frac{\langle \xi, v \rangle}{\sqrt{n}} \qquad \text{where } \zeta \in \mathbb{R}^n, \xi \in \mathbb{R}^p, \text{have i.i.d } N(0,1) \text{ entries.}$$

Next, we show that the $Y_{u,v}$ and $Z_{u,v}$ defined above satisfy conditions in Gordon's inequality. By definition, we have

$$\begin{aligned}
\mathbb{E}\left(Z_{u,v} - Z_{u',v'}\right)^2 &= \mathbb{E}\left(\frac{\langle u, Wv \rangle}{\sqrt{n}} - \frac{\langle u', Wv' \rangle}{\sqrt{n}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(u_i v_j - u_i' v_j'\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(u_i v_j - u_i' v_j + u_i' v_j - u_i' v_j'\right)^2 \\
&= \frac{1}{n}\left(\|v\|_2^2 \|u - u'\|_2^2 + \|u'\|_2^2 \|v - v'\|_2^2 + 2\left(\|v\|_2^2 - \langle v, v' \rangle\right)\left(\langle u, u' \rangle - \|u\|_2^2\right)\right),
\end{aligned} \tag{A.12}$$

Since $\|v\|_2^2 \leqslant 1$, $\|u'\|_2^2 \leqslant 1$, we have $\mathbb{E}\left(Z_{u,v} - Z_{u',v'}\right)^2 \leqslant \frac{1}{n}\left(\|u - u'\|_2^2 + \|v - v'\|_2^2\right)$. On the other hand, we have

$$\begin{aligned}
\mathbb{E}\left(Y_{u,v} - Y_{u',v'}\right)^2 &= \mathbb{E}\left(\frac{\langle \zeta, u - u' \rangle}{\sqrt{n}} + \frac{\langle \xi, v - v' \rangle}{\sqrt{n}}\right)^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{p}(u - u')^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}(v - v')^2\right) = \frac{1}{n}\left(\|u - u'\|_2^2 + \|v - v'\|_2^2\right).
\end{aligned} \tag{A.13}$$

Taking equation (A.12) and (A.13) together, we have

$$\mathbb{E}\left(Z_{u,v}-Z_{u',v'}\right)^2 \leqslant \frac{1}{n}\left(\|u-u'\|_2^2+\|v-v'\|_2^2\right)=\mathbb{E}\left(Y_{u,v}-Y_{u',v'}\right)^2.$$

If $V=V'$, then $n\mathbb{E}\left(\left(Z_{u,v}-Z_{u',v'}\right)^2\right)=\|u-u'\|_2=n\mathbb{E}\left(\left(Y_{u,v}-Y_{u',v'}\right)^2\right)$. By applying Lemma 21, we have

$$\mathbb{E}\left(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}} Z_{u,v}\right)\leqslant\mathbb{E}\left(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}} Y_{u,v}\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left(T\left(r_\ell,r_u\right)\right)&=\mathbb{E}\left(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}}\frac{\langle u,Wv\rangle}{\sqrt{n}}\right)\leqslant\mathbb{E}\left(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}}\left(\frac{\langle\xi,v\rangle}{\sqrt{n}}+\frac{\langle\zeta,u\rangle}{\sqrt{n}}\right)\right)\\&=\mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{\left\langle\Sigma^{\frac{1}{2}}\xi,\beta\right\rangle}{\sqrt{n}}\right)-\mathbb{E}\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right)\end{aligned}$$

Next, we bound these two terms. For the second term, we have $\mathbb{E}\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right)=\mathbb{E}\left(\sqrt{\frac{\xi_1^2+\ldots+\xi_n^2}{n}}\right)\geqslant\mathbb{E}\left(\frac{|\xi_1|+\ldots+|\xi_n|}{n}\right)=\sqrt{\frac{2}{\pi}}$. For the first term, we have $\mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{\left\langle\Theta^{\frac{1}{2}}\xi,\beta\right\rangle}{\sqrt{n}}\right)\leqslant\mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{\phi(\beta)\phi^*(\Theta^{\frac{1}{2}}\xi)}{\sqrt{n}}\right)$, where $\phi^*(\Theta^{\frac{1}{2}}\xi)$ is the the dual norm defined before. Since $\beta\in\mathbb{K}\left(r_\ell,r_u\right)$, $g\left(\phi(\beta)\right)\leqslant r_u$, by the definition of $g(t)$, we have

$$\phi(\beta)\leqslant\frac{r_u}{\left(4\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}. \tag{A.14}$$

Let $\eta_{G_g}=(\Theta^{\frac{1}{2}}\xi)_{G_g}$, to bound $\mathbb{E}\left(\max_g\left\|(\Theta^{\frac{1}{2}}\xi)_{G_g}\right\|_2\right)=\mathbb{E}\left(\max_g\left\|\eta_{G_g}\right\|_2\right)$. Since $\Theta^{\frac{1}{2}}\xi\sim N(0,\Theta)$, by the properties of normal distribution, its corresponding marginal distribution of jth variable $(\Theta^{\frac{1}{2}}\xi)_j$ also follows zero mean normal distribution with covariance matrix Θ_{jj}, which is the jth diagonal elements of Θ. Therefore, any

subset of $\Theta^{\frac{1}{2}}\xi$ is a zero-mean sub-Gaussian random sequence with parameters at most $\rho(\Theta)$. By (A.14) and Lemma 23, we have

$$\begin{aligned}
\mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{\phi(\beta)\phi^*\Theta^{\frac{1}{2}}\xi}{\sqrt{n}}\right) &\leqslant \mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{r_u}{\left(4\rho(\Theta)\left(\max\limits_{g\in[m]}\frac{1}{w_g h_{\min}^g}\right)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\frac{\phi^*\left(\Theta^{\frac{1}{2}}\xi\right)}{\sqrt{n}}\right)\\
&= \frac{r_u}{\left(4\rho(\Theta)\left(\max\limits_{g\in[m]}\frac{1}{w_g h_{\min}^g}\right)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\mathbb{E}\left(\frac{\phi^*\left(\Theta^{\frac{1}{2}}\xi\right)}{\sqrt{n}}\right)\\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\left(\max\limits_{g\in[m]}\frac{1}{w_g h_{\min}^g}\right)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\mathbb{E}\left(\max_{g\in[m]}\frac{1}{\sqrt{n}w_g}\left\|H\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right)\\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\left(\max\limits_{g\in[m]}\frac{1}{w_g h_{\min}^g}\right)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\mathbb{E}\left(\max_{g\in[m]}\frac{1}{\sqrt{n}w_g h_{\min}^g}\left\|\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right)\\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\mathbb{E}\left(\left\|\max_{g\in[m]}\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right)\\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)}\left(2\rho(\Theta)\sqrt{(\log m+d_{\max}\log 5)\,2\sigma^2}\right) \leqslant \frac{r_u}{2}
\end{aligned}$$

Therefore, $\mathbb{E}\left[T\left(r_\ell, r_u\right)\right] \leqslant -\sqrt{\frac{2}{\pi}}+\frac{r_u}{2}$. Next we want to bound $\mathbb{P}\left(T\left(r_\ell, r_u\right) \geqslant -\frac{1}{2}+r_u\right)$ based on the bound of this expectation. To apply Lemma 24, we first show that, the $f = T(r_l, r_u)$, a function of the random variable W is a $\frac{1}{\sqrt{n}}$-Lipschitz function and without making confusion, we denote the corresponding function as $T(W)$. For

any standard Gaussian matrix W_1 and W_2, we have

$$\begin{aligned}\Big|T(W_1)-T(W_2)\Big| &= \Big|\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}}\frac{\Big\langle u, W_1v\Big\rangle}{\sqrt{n}}-\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}}\frac{\Big\langle u, W_2v\Big\rangle}{\sqrt{n}}\Big|\\ &= \Big|\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\Big(-\frac{\big\|W_1v\big\|_2}{\sqrt{n}}\Big)-\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\Big(-\frac{\big\|W_2v\big\|_2}{\sqrt{n}}\Big)\Big|\\ &= \Big|\Big(-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\big\|W_1v\big\|_2}{\sqrt{n}}\Big)-\Big(-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\big\|W_2v\big\|_2}{\sqrt{n}}\Big)\Big|\\ &= \Big|\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\big\|W_2v\big\|_2}{\sqrt{n}}-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\big\|W_1v\big\|_2}{\sqrt{n}}\Big|.\end{aligned}$$

Suppose that $\frac{\|W_1v_1\|_2}{\sqrt{n}}=\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_1v\|_2}{\sqrt{n}}$ and $\frac{\|W_2v_2\|_2}{\sqrt{n}}=\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_2v\|_2}{\sqrt{n}}$.
Case 1 If $\|W_1v_1\|_2>\|W_2v_2\|_2$, then we have

$$\begin{aligned}|T(W_1)-T(W_2)| &= \Big|\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_2v\|_2}{\sqrt{n}}-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_1v\|_2}{\sqrt{n}}\Big|\\ &= \frac{\|W_1v_1\|_2-\|W_2v_2\|_2}{\sqrt{n}}\leqslant\frac{\|W_1v_2\|_2-\|W_2v_2\|_2}{\sqrt{n}}\\ &\leqslant\frac{\|(W_1-W_2)v_2\|_2}{\sqrt{n}}\leqslant\frac{\|W_1-W_2\|_F}{\sqrt{n}}\end{aligned}$$

.

Case 2 If $\|W_1v_1\|_2\leqslant\|W_2v_2\|_2$, then we have

$$\begin{aligned}|T(W_1)-T(W_2)| &= \Big|\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_2v\|_2}{\sqrt{n}}-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\frac{\|W_1v\|_2}{\sqrt{n}}\Big|\\ &= \frac{\|W_2v_2\|_2-\|W_1v_1\|_2}{\sqrt{n}}\leqslant\frac{\|W_2v_1\|_2-\|W_1v_1\|_2}{\sqrt{n}}\\ &\leqslant\frac{\|(W_1-W_2)v_1\|_2}{\sqrt{n}}\leqslant\frac{\|W_1-W_2\|_F}{\sqrt{n}}\end{aligned}$$

.

where $\|.\|_F$ represent the Frobenious norm of a matrix. Thus under the Euclidean norm, $T(W)$ is a $\frac{1}{\sqrt{n}}$-Lipschitz function. Therefore, by Lemma 23, we have

$$\mathbb{P}(T(r_l, r_u) - \mathbb{E}(T(r_l, r_u)) \geqslant t) \leqslant e^{-nt^2/2}, \forall t \geqslant 0$$

Set t $= \sqrt{\frac{2}{\pi}} - \frac{1}{2} + \frac{r_u}{2} \geqslant \frac{1}{4} + \frac{r_u}{2}$, we have, $\mathbb{E}(T(r_l, r_u)) + t \leqslant -\frac{1}{2} + r_u$ and $\mathbb{P}\left[T\left(r_\ell, r_u\right) \geqslant -\frac{1}{2} + r_u\right] \leqslant e^{-\frac{n}{32}} e^{-\frac{n}{8} r_u^2}$, which is actually the Lemma 20. □

Proof. **of Lemma 22** We can find a $\frac{1}{2}$ - cover of S^{d-1}, and for any $u \in S^{d-1}$ in the Euclidean norm with cardinally at most $N \leqslant 5^d$. Suppose that there exists $u^{q(u)} \in \left\{u^1, \ldots, u^N\right\}$, such that $\left\|u^{q(u)} - u\right\|_2 \leqslant \frac{1}{2}$. By the variational representation of the ℓ_2 norm, we have

$$\|\alpha\|_2 = \max_{u \in S^{d-1}} \langle u, \alpha \rangle \leqslant \max_{q(u) \in [N]} \left\langle u^{q(u)}, \alpha \right\rangle + \frac{1}{2} \|\alpha\|_2 .$$

Therefore, $\|\alpha\|_2 \leqslant 2 \max_{q(u) \in [N]} \left\langle u^{q(u)}, \alpha \right\rangle$. Consequently,

$$\begin{aligned} \mathbb{E}\left(\exp\left(t \|\alpha\|_2\right)\right) &\leqslant \mathbb{E}\left(\exp\left(2t \max_{q \in [N]} \langle u^q, \alpha \rangle\right)\right) = \mathbb{E}\left(\max_{q \in [N]} \exp\left(2t \langle u^q, \alpha \rangle\right)\right) \\ &\leqslant \sum_{q=1}^{N} \mathbb{E}\left(\exp\left(2t \langle u^q, \alpha \rangle\right)\right) \leqslant 5^d \exp\left(\frac{4t^2\sigma^2}{2}\right) \leqslant 5^d \exp\left(2t^2\sigma^2\right). \end{aligned}$$

□

Proof. **of Lemma 23** For any $t > 0$, by Jensen's inequality, we have

$$\begin{aligned} \exp\left(t\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right)\right) &\leqslant \mathbb{E}\left(\exp\left(t \max_g \left\|\alpha_{G_g}\right\|_2\right)\right) = \mathbb{E}\left(\max_j \exp\left(t \left\|\alpha_{G_g}\right\|_2\right)\right) \\ &\leqslant \sum_{j=1}^{m} \mathbb{E}\left(\exp\left(t \left\|\alpha_{G_g}\right\|_2\right)\right) \leqslant \sum_{j=1}^{m} 5^{d_g} \exp\left(2t^2\sigma^2\right) \leqslant m \cdot 5^{d_{\max}} \cdot \exp(2t^2\sigma^2). \end{aligned}$$

By taking log at both sides, we have $t\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant \log m + d_{\max}\log 5 + 2t^2\sigma^2$. Consequently, let $t = \sqrt{\frac{\log m + d_{\max}\log 5}{2\sigma^2}}$, we have $\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant 2\sqrt{(\log m + d_{\max}\log 5)\, 2\sigma^2}$.

□

A.3.8 Proof of Theorem 4

The two lemmas below are integral to the proof:

Lemma 25 (Packing Number for Binary Sets)**.** *Consider a set A defined for real numbers m, s_g as*

$$A = \left\{ a \in \{0,1\}^m \mid \sum_{j=1}^{m} a_j \leq s_g \right\}.$$

Then the $\sqrt{\frac{s_g}{2}}$-packing number of set $A \geqslant \frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}\cdot 2^{\frac{s_g}{2}}}$, and

$$\log\left(\frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}\cdot 2^{\frac{s_g}{2}}}\right) \asymp s_g \log(\frac{m}{s_g}).$$

Lemma 26 (Packing Number for Sparse Group Vectors)**.** *For the set $\Omega(G, s_g)$, the $\sqrt{\frac{2ds_g}{5}}$-packing number $\gtrsim \frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}\cdot 2^{\frac{s_g}{2}}} \cdot (\sqrt{2})^{ds_g}$, and*

$$\log\left(\frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}\cdot 2^{\frac{s_g}{2}}} \cdot (\sqrt{2})^{ds_g}\right) \asymp s_g(d + \log(\frac{m}{s_g})).$$

Proof. **of Theorem 4** First, select N points $\omega^{(1)}, \ldots, \omega^{(N)}$ from $\Omega(G, s_g)$ such that $\left\|\omega^{(i)} - \omega^{(j)}\right\| > \sqrt{\frac{2ds_g}{5}}$ for all distinct i, j. Clearly, $\left\|\omega^{(i)} - \omega^{(j)}\right\| \leqslant \sqrt{4s_g d}$. Define $\beta^{(i)} = r\omega^{(i)}$ for each i. This results in

$$\frac{2ks_g r^2}{5} \leq \left\|\beta^{(i)} - \beta^{(j)}\right\|_2^2 \leqslant 4s_g d r^2.$$

Next, let $y^{(i)} = X\beta^{(i)} + \varepsilon$ for $1 \leqslant i \leqslant N$. Consider the Kullback-Leibler divergence between different distribution pairs:

$$D_{KL}\left((y^{(i)}, X), (y^{(j)}, X)\right) = \mathbb{E}_{(y^{(j)},X)}\left[\log\left(\frac{p\left(y^{(i)}, X\right)}{p\left(y^{(j)}, X\right)}\right)\right].$$

where $p\left(y^{(i)}, X\right)$ is the probability density of $\left(y^{(i)}, X\right)$. Conditioning on X, we have

$$\mathbb{E}_{(y^{(j)},X)}\left[\log\left(\frac{p\left(y^{(i)}, X\right)}{p\left(y^{(j)}, X\right)}\right) \mid X\right] = \frac{\|X(\beta^{(i)} - \beta^{(j)})\|_2^2}{2\sigma^2}.$$

Thus, for $1 \leq i \neq j \leq N$,

$$D_{KL}\left(\left(y^{(i)}, X\right), \left(y^{(j)}, X\right)\right) = \mathbb{E}_X \frac{\left\|X\left(\beta^{(i)} - \beta^{(j)}\right)\right\|_2^2}{2\sigma^2} = \frac{n(\beta^{(i)} - \beta^{(j)})^\top \Sigma(\beta^{(i)} - \beta^{(j)})}{2\sigma^2}$$
$$\leq \frac{3c_1\left\|\beta^{(i)} - \beta^{(j)}\right\|_2^2}{2\sigma^2} \leq \frac{2c_1 ndr^2 s_g}{\sigma^2}.$$

From Lemma 26, $\log N \asymp s_g\left(d + \log\frac{m}{s_g}\right)$. Setting $\frac{\frac{ndr^2 s_g}{\sigma^2} + \log 2}{\log N} = \frac{1}{2}$, we obtain

$$r \gtrsim \sqrt{\frac{\left(d + \log\frac{m}{s_g}\right)\sigma^2}{3nd}}.$$

By generalized Fano's Lemma, $\inf_{\hat{\beta}} \sup_{\beta} \mathbb{E}\|\hat{\beta} - \beta\|_2 \geqslant \sqrt{\frac{2r^2 k s_g}{5}}\left(1 - \frac{\frac{ndr^2 s_g}{\sigma^2} + \log 2}{\log N}\right)$. Consequently,

$$\inf \sup \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \geq \left(\inf \sup \mathbb{E}\|\hat{\beta} - \beta\|_2\right)^2 \gtrsim \frac{\sigma^2\left(s_g(d + \log(\frac{m}{s_g}))\right)}{n}.$$

□

Proof. **of Lemma 25** Notice that the cardinality of A is $\binom{m}{s_g}$. Denote the hamming distance between any two points $x, y \in A$ by

$$h(a,b) = |\{j : a_j \neq b_j\}|.$$

Then, for a fixed point a $\in A$,

$$\left| \left\{ b \in A, h(a,b) \leq \frac{s_g}{2} \right\} = \binom{m}{\lfloor \frac{s_g}{2} \rfloor} \cdot 2^{\lfloor \frac{s_g}{2} \rfloor} \right|.$$

In fact, all elements $b \in A$ with $h(a,b) \leq \frac{s_g}{2}$ can be obtained as follows. First, take any subset $J \subset [m]$ of cardinality $\lfloor \frac{s_g}{2} \rfloor$, then set $a_j = b_j$ for $j \notin J$ and choose $b_j \in \{0,1\}$ for $j \in J$.

Now let A_s be any subset of A with cardinality at most $T = \frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}\cdot 2^{\frac{s_g}{2}}}$, then we have

$$|\{b \in A \mid \text{there exist } a \in A_s \text{ with } h(a,b) \leq \frac{s_g}{2}\} \leq (|A_s|) \cdot \binom{m}{\lfloor \frac{s_g}{2} \rfloor} \cdot 2^{\frac{s_g}{2}} | < |A|.$$

It implies that one can find an element $b \in A$ with $h(a,b) > \frac{s_g}{2}$ for all $a \in A_s$. Therefore one can construct a subset A_s with $|A_s| \geq T$ and the property $h(a,b) > \frac{s_g}{2}$ for any two distinct elements $a, b \in A_s$.

On the other hand, $h(a,b) > \frac{s_g}{2}$ implies $\|a-b\| > \sqrt{\frac{s_g}{2}}$. Therefore, there exist at least T points in A such that the distance between any two points is greater than $\sqrt{\frac{s_g}{2}}$.

Moreover, since $\frac{\binom{m}{s_g}}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor}} = \frac{\lfloor \frac{s_g}{2} \rfloor ! (m - \lfloor \frac{s_g}{2} \rfloor)!}{s_g!(m-s_g)!} = \frac{(m-s_g+1)\cdots m - \lfloor \frac{s_g}{2} \rfloor}{(\lfloor \frac{s_g}{2} \rfloor + 1)\cdots s_g} = \prod_{j=1}^{\lceil \frac{s_g}{2} \rceil} \frac{m-s_g+j}{\lfloor \frac{s_g}{2} \rfloor + j}$, we have

$$\left(\frac{m - \lfloor \frac{s_g}{2} \rfloor}{2s_g} \right)^{\lfloor \frac{s_g}{2} \rfloor} \leqslant \frac{\binom{m}{s_g}}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor} 2^{\frac{s_g}{2}}} \leqslant \left(\frac{m - s_g + 1}{\lceil s_g \rceil} \right)^{\lceil \frac{s_g}{2} \rceil},$$

and therefore we can find C_1, C_2, such that $C_1 s_g \log(\frac{m}{s_g}) \leqslant \log T \leqslant C_2 s_g \log(\frac{m}{s_g})$, so that

$$\log\left(\frac{\binom{m}{s_g} - 2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor} \cdot 2^{\frac{s_g}{2}}}\right) \asymp s_g \log(\frac{m}{s_g}).$$

□

Proof. **of Lemma 26** Given a group support $a \in A$, define $k_a = \left|\left\{i \mid i \in \left(\bigcup_{\{g|a_g=0\}} G_g\right)^c\right\}\right|$, and the set

$$\Omega^{(a)} = \left\{\omega \in \mathbb{R}^p \mid \omega_i = 0 \text{ if } i \in \bigcup_{\{g|a_g=0\}} G_g, \omega_i \in \{-1, 1\} \text{ if } i \in \left(\bigcup_{\{g|a_g=0\}} G_g\right)^c\right\}.$$

Notice that $\Omega^{(a)} \subseteq \Omega(G, s_g)$, and $|\Omega^{(a)}| = 2^{k_a}$. Also denote the hamming distance between $x, y \in \Omega^{(a)}$ by

$$h(x, y) = |\{j : x_j \neq y_j\}|.$$

Then for any fixed $x \in \Omega_G^{(a)}$, we have

$$\left|\{y \in \Omega^{(a)}, h(x, y) \leq \frac{k_a}{10}\}\right| = \sum_{j=0}^{\lfloor \frac{k_a}{10} \rfloor} \binom{k_a}{j}$$

Let $\Omega_s^{(a)}$ be any subset of $\Omega^{(a)}$ with cardinality at most $N^{(a)} = \frac{2^{k_a}-2}{\sum_{j=0}^{\lfloor \frac{k_a}{10} \rfloor} \binom{k_a}{j}}$. Then,

$$\left|\{y \in \Omega^{(a)} \mid \exists x \in \Omega_s^{(a)} \text{ with } h(x, y) \leq \frac{k_a}{10}\}\right| < |\Omega^{(a)}|.$$

On the other hand, $h(x, y) > \frac{k_a}{10}$ implies $\|x - y\| \geq \sqrt{\frac{2k_a}{5}}$. Thus, there are at least $N^{(a)}$ points in $\Omega^{(a)}$ with pairwise distances greater than $\sqrt{\frac{2k_a}{5}}$. From the results in

Graham et al. (1994, Chap. 9),

$$\sum_{j\le\lfloor\frac{k_a}{10}\rfloor}\binom{k_a}{j} < \frac{9}{8}\binom{k_a}{\lfloor\frac{k_a}{10}\rfloor} \le \frac{9}{8}(10e)^{\frac{k_a}{10}} \le \frac{9}{8}2^{\frac{k_a}{2}}.$$

Consequently, we have $N^{(a)} > \frac{8}{9}2^{\frac{k_a}{2}} \gtrsim (\sqrt{2})^{k_a}$.

The value of k_a depends on the predefined groups and group support a and spans a range from 0 to $s_g d$. Lemma 26 seeks a lower bound for all conceivable overlapping patterns, necessitating an analysis of the maximum value of k_a.

Furthermore, according to Lemma 25, we can identify at least T points in A where the distance between any two points exceeds $\sqrt{\frac{s_g}{2}}$. For $\{a_1, \cdots, a_T\}$ group supports, if there is a group structure such that we could find at least $\frac{8}{9}(\sqrt{2})^{s_g d}$ on each group support, and the distance between every pair of these points is greater than $\sqrt{\frac{2s_g d}{5}}$, then Lemma 26 is proved.

Considering m non-overlapping groups, $k_a = s_g d$ for each group support a. In addition, given any two group support a, b with $\|a-b\| > \sqrt{\frac{s_g}{2}}$, $\|x-y\| > \sqrt{\frac{ds_g}{2}} > \sqrt{\frac{2ds_g}{5}}$ for any $x \in \Omega^{(a)}$ and $y \in \Omega^{(b)}$. Thus, considering all possible overlapping patterns, we can find at least $\frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}} \cdot \frac{8}{9}(\sqrt{2})^{ds_g}$ point in $\Omega(G, s_g)$, such that the distance between every pair of points is greater than $\sqrt{\frac{2ds_g}{5}}$.

□

A.3.9 Proof of Theorem 5

This proof consists of parts: Parts I-IV dedicated to Theorem 5.1, and Part V is for Theorem 5.2. To be more specific, Part I provides some additional concepts, Part II introduces the reduced problem, Part III shows the successful selection of the correct pattern under favorable conditions, and Part IV establishes that certain

conditions are satisfied with high probability.

Part I

Recall that $\mathbf{S} = supp(\beta^*)$. With $\mathbf{S}$, we define the norm $\phi_{\mathbf{S}}$ for any $\beta \in \mathbb{R}^p$ as

$$\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) = \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S} \cap G_g}\|_2,$$

along with its dual norm $(\phi_{\mathbf{S}})^*[u] = \sup_{\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leq 1} \beta_{\mathbf{S}}^\top u$. Similarly, for $\mathbf{S}^c = [p] \setminus \mathbf{S}$, we define the norm $\phi_{\mathbf{S}}^c$ for any $\beta \in \mathbb{R}^p$ as

$$\phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) = \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2,$$

accompanied by its corresponding dual norm $(\phi_{\mathbf{S}}^c)^*[u] = \sup_{\phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) \leq 1} \beta_{\mathbf{S}^c}^\top u$.

We also introduce equivalence parameters $a_{\mathbf{S}}, A_{\mathbf{S}}, a_{\mathbf{S}^c}, A_{\mathbf{S}^c}$ as follows:

$$\forall \beta \in \mathbb{R}^p,\ a_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant \phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant A_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1, \tag{A.15}$$

$$\forall \beta \in \mathbb{R}^p,\ a_{\mathbf{S}^c} \|\beta_{\mathbf{S}^c}\|_1 \leqslant \phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) \leqslant A_{\mathbf{S}^c} \|\beta_{\mathbf{S}^c}\|_1. \tag{A.16}$$

We now study the equivalence parameters from two aspects. First, since

$$\sup_{a_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant 1} \beta_{\mathbf{S}}^\top u \geqslant \sup_{\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant 1} \beta_{\mathbf{S}}^\top u \geqslant \sup_{A_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant 1} \beta_{\mathbf{S}}^\top u,$$

by the definition of dual norm, we have

$$\forall u \in \mathbb{R}^{|\mathbf{S}|},\ A_{\mathbf{S}}^{-1} \|u\|_\infty \leqslant (\phi_{\mathbf{S}})^*[u] \leqslant a_{\mathbf{S}}^{-1} \|u\|_\infty. \tag{A.17}$$

Similarly, by order-reversing,

$$\forall u \in \mathbb{R}^{|\mathbf{S}^c|}, A_{\mathbf{S}^c}^{-1}\|u\|_\infty \leqslant (\phi_{\mathbf{S}}^c)^*[u] \leqslant a_{\mathbf{S}^c}^{-1}\|u\|_\infty. \tag{A.18}$$

Second, by the Cauchy-Schwarz inequality, for any $\beta \in \mathbb{R}^p$ and $g \in \mathsf{G}_{\mathbf{S}}$,

$$\frac{w_g}{\sqrt{d_g}}\|\beta_{\mathbf{S}\cap G_g}\|_1 \leqslant w_g\|\beta_{\mathbf{S}\cap G_g}\|_2 \leqslant \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g\|\beta_{\mathbf{S}\cap G_g}\|_1.$$

Consequently, we have

$$\min_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g}{\sqrt{d_g}}\|\beta_{\mathbf{S}}\|_1 \leqslant \phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g\|\beta_{\mathbf{S}}\|_1,$$

Therefore, we can set $a_{\mathbf{S}} = \min_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g}{\sqrt{d_g}}$ and $A_{\mathbf{S}} = h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g$. With an trivial extension, we can set $a_{\mathbf{S}^c} = \min_{g\in\mathsf{G}_{\mathbf{S}^c}} w_g/\sqrt{d_g}$.

Part II

From the full problem to the reduced problem

Recall that the group lasso estimator in (2.14) is defined as

$$\hat{\beta}^G = \arg\min_{\beta\in\mathbb{R}^p} \quad \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n\phi^G(\beta). \tag{A.19}$$

Now we write $\phi^G(\beta) = \phi(\beta)$ and $L(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2$ for ease of notation. Following Jenatton et al. (2011a); Wainwright (2009), we consider the following restricted problem

$$\begin{aligned}\hat{\beta}^R &= \underset{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0}{\arg\min}\ L(\beta)+\lambda_n\phi(\beta) = \underset{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0}{\arg\min}\ L(\beta)+\lambda_n\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\left\|\beta_{\mathbf{S}\cap G_g}\right\|_2 \\ &:= \underset{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0}{\arg\min}\ L(\beta)+\lambda_n\phi_{\mathbf{S}}(\beta_{\mathbf{S}}).\end{aligned} \tag{A.20}$$

Let $L_{\mathbf{S}}(\beta_{\mathbf{S}}) = \frac{1}{2n}\|Y - X_{\mathbf{S}}\beta_{\mathbf{S}}\|_2^2$. Due to the restriction of $\hat{\beta}^R$, we can obtain $\hat{\beta}^R$ by first solving the following reduced problem

$$\begin{aligned}\hat{\beta}_{\mathbf{S}} &= \underset{\beta_{\mathbf{S}}\in\mathbb{R}^{|\mathbf{S}|}}{\arg\min}\quad \frac{1}{2n}\|Y - X_{\mathbf{S}}\beta_{\mathbf{S}}\|_2^2 + \lambda_n\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\left\|\beta_{\mathbf{S}\cap G_g}\right\|_2 \\ &= \underset{\beta_{\mathbf{S}}\in\mathbb{R}^{|\mathbf{S}|}}{\arg\min}\, L_{\mathbf{S}}(\beta_{\mathbf{S}}) + \lambda_n\phi_{\mathbf{S}}(\beta_{\mathbf{S}})\end{aligned} \tag{A.21}$$

and then padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$. In addition,

$$\begin{aligned}L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) &= \frac{1}{2n}\|Y - X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\|_2^2 \\ &= \frac{1}{2n}\left(Y^\top Y - 2Y^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right) \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X\beta^* + \epsilon)^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right) \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X_{\mathbf{S}}\beta^*_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} - 2\epsilon^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right),\end{aligned}$$

and consequently,

$$\begin{aligned}\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) &= \frac{1}{n}X_{\mathbf{S}}^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} - \frac{1}{n}X_{\mathbf{S}}^\top X_{\mathbf{S}}\beta^*_{\mathbf{S}} - \frac{1}{n}\epsilon^\top X_{\mathbf{S}} \\ &:= Q_{\mathbf{SS}}(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}) - q_{\mathbf{S}},\end{aligned} \tag{A.22}$$

where $Q = \frac{1}{n}X^\top X$, $q = \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i$.

Part III

Part III mostly follows the proof in Theorem 7 of Jenatton et al. (2011a). Here we aim to show that $supp(\hat{\beta}^G) = \mathbf{S}$ under certain conditions.

To begin with, Given $\beta \in \mathbb{R}^p$, we define $J^G(\beta)$ as:

$$J^G(\beta) = [p] \setminus \Big\{ \bigcup_{G_g \cap supp(\beta) = \emptyset} G_g \Big\}.$$

$J^G(\beta)$ is called the adapted hull of the support of β in Jenatton et al. (2011a). For simplicity, we write $J^G(\beta) = J(\beta)$. Notice that by assumption we have

$$J(\beta^*) = [p] \setminus \Big\{ \bigcup_{G_g \cap supp(\beta^*) = \emptyset} G_g \Big\} = \mathbf{S}.$$

Now we consider the reduced problem (A.21), and we want to show that for all $g \in \mathsf{G}_{\mathbf{S}}$, $\left\| \hat{\beta}_{\mathbf{S} \cap G_g} \right\|_\infty > 0$. That is, no active group is missing.

Lemma 27. *(Lemma 14 of Jenatton et al. (2011a))*

For the loss $L(\beta)$ and norm ϕ in (A.19), $\hat{\beta} \in \mathbb{R}^p$ is a solution of

$$\min_{\beta \in \mathbb{R}^p} L(\beta) + \lambda_n \phi(\beta) \tag{A.23}$$

if and only if

$$\begin{cases} \nabla L(\hat{\beta})_{J(\hat{\beta})} + \lambda_n r(\hat{\beta})_{J(\hat{\beta})} = \mathbf{0} \\ (\phi^c_{J(\hat{\beta})})^* \left[\nabla L(\hat{\beta})_{J(\hat{\beta})^c} \right] \leqslant \lambda_n. \end{cases} \tag{A.24}$$

In addition, the solution $\hat{\beta}$ satisfies

$$\phi^*[\nabla L(\hat{\beta})] \leqslant \lambda_n. \tag{A.25}$$

As $\hat{\beta}_{\mathbf{S}}$ is the solution of (A.21), Equation (A.25) in Lemma 27 implies that

$$(\phi_{\mathbf{S}})^* \left[\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})\right] \overset{(A.22)}{=} (\phi_{\mathbf{S}})^* \left[Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right] \leqslant \lambda_n. \tag{A.26}$$

By the property of the equivalent parameters, we have

$$A_{\mathbf{S}}^{-1} \left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right\|_\infty \overset{(A.17)}{\leqslant} (\phi_{\mathbf{S}})^* \left[Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right] \overset{(A.26)}{\leqslant} \lambda_n. \tag{A.27}$$

If

$$\lambda_n \leqslant \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3|\mathbf{S}|^{\frac{1}{2}} A_{\mathbf{S}}}, \tag{A.28}$$

and

$$\|q_{\mathbf{S}}\|_\infty \leqslant \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3|\mathbf{S}|^{\frac{1}{2}}}, \tag{A.29}$$

then we have

$$\begin{aligned}
\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty &= \left\|Q_{\mathbf{SS}}^{-1} Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant \left\|Q_{\mathbf{SS}}^{-1}\right\|_{\infty,\infty} \left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}} \gamma_{\max}\left(Q_{\mathbf{SS}}^{-1}\right) \left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}} \gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right) \left(\left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right\|_\infty + \|q_{\mathbf{S}}\|_\infty\right) \\
&\overset{(A.27)}{\leqslant} |\mathbf{S}|^{\frac{1}{2}} \gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\left(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty\right) \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}} \gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\lambda_n A_{\mathbf{S}} + |\mathbf{S}|^{\frac{1}{2}} \gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\|q_{\mathbf{S}}\|_\infty \\
&\leqslant \frac{2}{3}\beta^*_{\min}.
\end{aligned} \tag{A.30}$$

If there exist a group $g \in \mathsf{G}_{\mathbf{S}}$ such that $\left\|\hat{\beta}_{\mathbf{S}\cap G g}\right\|_\infty < \frac{\beta^*_{\min}}{3}$, then

$$\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty > \beta^*_{\min} - \frac{\beta^*_{\min}}{3} = \frac{2\beta^*_{\min}}{3}.$$

Thus, Equation (A.30) implies that for all $g \in \mathsf{G}_{\mathbf{S}}$,

$$\left\|\hat{\beta}_{\mathbf{S}\cap G_g}\right\|_\infty > \frac{\beta^*_{\min}}{3} > 0. \tag{A.31}$$

Secondly, we want to show that $\hat{\beta}^R$ solves problem (A.19). As $\hat{\beta}^R$ is obtained by padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$,

$$\begin{aligned} J(\hat{\beta}^R) &= [p] \setminus \left\{ \bigcup_{G_g \cap supp(\hat{\beta}^R)=\emptyset} G_g \right\} = [p] \setminus \left\{ \bigcup_{G_g \cap supp(\hat{\beta}_{\mathbf{S}})=\emptyset} G_g \right\} \\ &\overset{(A.30)}{=} [p] \setminus \left\{ \bigcup_{G_g \cap \mathbf{S}=\emptyset} G_g \right\} = \mathbf{S}. \end{aligned}$$

From Lemma 27 we know that $\hat{\beta}^R$ is the optimal for problem (A.19) if and only if

$$\nabla L(\hat{\beta}^R)_{\mathbf{S}} + \lambda_n r(\hat{\beta}^R)_{\mathbf{S}} = \mathbf{0}, \tag{A.32}$$

and

$$(\phi^c_{\mathbf{S}})^* \left[\nabla L(\hat{\beta}^R)_{\mathbf{S}^c}\right] \leqslant \lambda_n. \tag{A.33}$$

We now verify the condition in Equation (A.32). Since

$$\begin{aligned} L(\hat{\beta}^R) &= \frac{1}{2n}\|Y - X\hat{\beta}^R\|_2^2 \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X\beta^*)^\top X\hat{\beta}^R - 2\epsilon^\top X\hat{\beta}^R + (X\hat{\beta}^R)^\top X\hat{\beta}^R\right), \end{aligned}$$

we have

$$\begin{aligned}\nabla L(\hat{\beta}^R)_{\mathbf{S}} &= \left[\frac{1}{n}X^\top X\left(\hat{\beta}^R-\beta^*\right)-\frac{1}{n}\epsilon^\top X\right]_{\mathbf{S}} \\ &= \left[Q\left(\hat{\beta}^R-\beta^*\right)\right]_{\mathbf{S}}-q_{\mathbf{S}}=Q_{\mathbf{SS}}\left(\hat{\beta}^R-\beta^*\right)_{\mathbf{S}}-q_{\mathbf{S}} \\ &= Q_{\mathbf{SS}}\left(\hat{\beta}^R_{\mathbf{S}}-\beta^*_{\mathbf{S}}\right)-q_{\mathbf{S}}=Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}}-\beta^*_{\mathbf{S}}\right)-q_{\mathbf{S}} \\ &= \nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}).\end{aligned} \tag{A.34}$$

On the other hand, as $\hat{\beta}^R$ is obtained by padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$, we have

$$\lambda_n r(\hat{\beta}^R)_{\mathbf{S}} = \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}).$$

Because $\hat{\beta}_{\mathbf{S}}$ is the optimal for problem (A.21), Equation (A.24) in Lemma 27 implies that

$$\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})+\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) \overset{(A.22)}{=} Q_{\mathbf{SS}}(\hat{\beta}_{\mathbf{S}}-\beta^*_{\mathbf{S}})-q_{\mathbf{S}}+\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})=\mathbf{0}. \tag{A.35}$$

Thus, Equation (A.32) holds as

$$\nabla L(\hat{\beta}^R)_{\mathbf{S}}+\lambda_n r_{\mathbf{S}}(\hat{\beta}^R)=\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})+\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) \overset{(A.35)}{=} \mathbf{0}. \tag{A.36}$$

Now we continue to show Equation (A.33). Notice that

$$\left(\hat{\beta}^R-\beta^*\right)_{\mathbf{S}} \overset{(A.34)}{=} \left(\hat{\beta}_{\mathbf{S}}-\beta^*_{\mathbf{S}}\right) \overset{(A.35)}{=} Q^{-1}_{\mathbf{SS}}(q_{\mathbf{S}}-\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})). \tag{A.37}$$

Let $q_{\mathbf{S}^c|\mathbf{S}}=q_{\mathbf{S}^c}-Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}q_{\mathbf{S}}$, we have

$$\begin{aligned}\nabla L(\hat{\beta}^R)_{\mathbf{S}^c} &\overset{(A.34)}{=} \left(Q(\hat{\beta}^R - \beta^*)\right)_{\mathbf{S}^c} - q_{\mathbf{S}^c} = Q_{\mathbf{S}^c\mathbf{S}}(\hat{\beta}^R - \beta^*)_{\mathbf{S}} - q_{\mathbf{S}^c} \\ &\overset{(A.37)}{=} Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(q_{\mathbf{S}} - \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})\right) - q_{\mathbf{S}^c} \\ &= -Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) + Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}} - q_{\mathbf{S}^c} \\ &= -\lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right) - \lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}r_{\mathbf{S}}(\beta^*_{\mathbf{S}}) - q_{\mathbf{S}^c|\mathbf{S}}.\end{aligned} \quad . \tag{A.38}$$

The previous expression leads us to study the difference of $r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})$. We now introduce the following lemma.

Lemma 28. *(Lemma 12 of Jenatton et al. (2011a))*

For any $J \subset [p]$, let u_J and v_J be two nonzero vectors in $\mathbb{R}^{|J|}$, and define the mapping $r_J : \mathbb{R}^{|J|} \mapsto \mathbb{R}^{|J|}$ such that

$$r_J(\beta_J)_j = \beta_j \sum_{g \in \mathsf{G}_J, G_g \cap j \neq \phi} \frac{\omega_g}{\left\|\beta_{J \cap G_g}\right\|_2}.$$

Then there exists $\xi_J = t_0 u_J + (1 - t_0)v_J$ for some $t_0 \in (0, 1)$, such that

$$\left\|r_J(u_J) - r_J(v_J)\right\|_1 \leqslant \left\|u_J - v_J\right\|_\infty \left(\sum_{j \in J}\sum_{g \in \mathsf{G}_J} \frac{w_g \mathbb{1}_{\{j \in G_g\}}}{\left\|\xi_{J \cap G_g}\right\|_2} + \sum_{j \in J}\left(\sum_{k \in J}\sum_{g \in \mathsf{G}_J} \frac{|\xi_j||\xi_k| w_g^4 \mathbb{1}_{\{j,k \in G_g\}}}{\left\|\xi_{J \cap G_g}\right\|_2^3}\right)\right).$$

Lemma 28 implies that

$$\left\|r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_1 \leqslant \left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \left(\sum_{j \in \mathbf{S}}\sum_{g \in \mathsf{G}_{\mathbf{S}}} \frac{w_g \mathbb{1}_{\{j \in G_g\}}}{\left\|\tilde{\beta}_{\mathbf{S} \cap G_g}\right\|_2} + \sum_{j \in \mathbf{S}}\sum_{k \in \mathbf{S}}\sum_{g \in \mathsf{G}_{\mathbf{S}}} \frac{(w_g)^4 \mathbb{1}_{\{j,k \in G_g\}}|\tilde{\beta}_j||\tilde{\beta}_k|}{w_g^3\left\|\tilde{\beta}_{\mathbf{S} \cap G_g}\right\|_2^3}\right), \tag{A.39}$$

where $\tilde{\beta} = t_0\hat{\beta}_{\mathbf{S}} + (1 - t_0)\,\beta^*_{\mathbf{S}}$.

To find an upper bound of the right-hand side. Recall that Equation (A.30) implies that $\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \leqslant \frac{2}{3}\beta^*_{\min}$, so we have

$$\begin{aligned}\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2 &\geqslant \sqrt{|\mathbf{S}\cap G_g|}\min\{|\tilde{\beta}|_j \mid \tilde{\beta}_j \neq 0\}\\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}(\beta^*_{\min} - t_0\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty)\\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}(\beta^*_{\min} - \left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty)\\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}\frac{\beta^*_{\min}}{3}.\end{aligned}$$

Consequently, the first term could be upper bounded by

$$\sum_{j\in\mathbf{S}}\sum_{g\in\mathsf{G}_\mathbf{S}}\frac{w_g\mathbb{1}_{\{j\in G_g\}}}{\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2} = \sum_{g\in\mathsf{G}_\mathbf{S}}\frac{w_g|\mathbf{S}\cap G_g|}{\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2} \leqslant \frac{3}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_\mathbf{S}} w_g\sqrt{|\mathbf{S}\cap G_g|}$$

On the other hand, the Cauchy-Schwarz inequality gives

$$\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_1^2 \leqslant |\mathbf{S}\cap G_g|\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2^2.$$

Thus, the second term could also be upper bounded by

$$\begin{aligned}\sum_{j\in\mathbf{S}}\sum_{k\in\mathbf{S}}\sum_{g\in\mathsf{G}_\mathbf{S}}\frac{(w_g)^4\mathbb{1}_{\{j,k\in G_g\}}|\tilde{\beta}_j||\tilde{\beta}_k|}{w_g^3\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2^3} &= \sum_{g\in\mathsf{G}_\mathbf{S}}\frac{w_g^4\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_1^2}{w_g^3\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2^3}\\ &\leqslant \sum_{g\in\mathsf{G}_\mathbf{S}}\frac{w_g|\mathbf{S}\cap G_g|}{\left\|\tilde{\beta}_{\mathbf{S}\cap G_g}\right\|_2}\\ &\leqslant \frac{3}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_\mathbf{S}} w_g\sqrt{|\mathbf{S}\cap G_g|}.\end{aligned}$$

Let $c_2 = \frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_\mathbf{S}} w_g\sqrt{|\mathbf{S}\cap G_g|}$, then Equation (A.39) implies

$$\left\|r_\mathbf{S}(\hat{\beta}_\mathbf{S}) - r_\mathbf{S}(\beta^*_\mathbf{S})\right\|_1 \leqslant c_2\left\|\hat{\beta}_\mathbf{S} - \beta^*_\mathbf{S}\right\|_\infty.$$

If

$$\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} \leqslant 3, \tag{A.40}$$

then we have

$$\begin{aligned}
\left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty &= \left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty \\
&\leqslant \left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\right\|_{\infty,2}\left\|Q_{\mathbf{SS}}^{-\frac{1}{2}}\right\|_2\left\|r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_2 \\
&\leqslant 3\gamma_{\max}(Q_{\mathbf{SS}}^{-\frac{1}{2}})\left\|r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_\infty \\
&\leqslant 3\gamma_{\min}^{-\frac{1}{2}}(Q_{\mathbf{SS}})c_2\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \\
&\overset{(A.30)}{\leqslant} 3c_2\gamma_{\min}^{-\frac{1}{2}}\left(Q_{\mathbf{SS}}\right)|\mathbf{S}|^{\frac{1}{2}}\gamma_{\min}^{-1}\left(Q_{\mathbf{SS}}\right)\left(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty\right) \\
&= 3\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}\left(Q_{\mathbf{SS}}\right)|\mathbf{S}|^{\frac{1}{2}}\left(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty\right).
\end{aligned}$$

If the following conditions are satisfied:

$$a_{\mathbf{S}^c}^{-1}\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}\left(Q_{\mathbf{SS}}\right)|\mathbf{S}|^{\frac{1}{2}}\lambda_n A_{\mathbf{S}} \leqslant \frac{\tau}{12}, \tag{A.41}$$

$$a_{\mathbf{S}^c}^{-1}\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}\left(Q_{\mathbf{SS}}\right)|\mathbf{S}|^{\frac{1}{2}}\|q_{\mathbf{S}}\|_\infty \leqslant \frac{\tau}{12}, \tag{A.42}$$

$$(\phi^c_{\mathbf{S}})^*[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\mathbf{r}_{\mathbf{S}}] \leqslant 1-\tau, \tag{A.43}$$

$$(\phi^c_{\mathbf{S}})^*[q_{\mathbf{S}^c|\mathbf{S}}] \leqslant \frac{\lambda_n\tau}{2}, \tag{A.44}$$

then we have

$$
\begin{aligned}
(\phi^c_{\mathbf{S}})^* \left[\nabla L(\hat{\beta}^R)_{\mathbf{S}^c}\right] &\overset{(A.38)}{=} (\phi^c_{\mathbf{S}})^* \left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} \left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right) + \lambda_n Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} r_{\mathbf{S}}(\beta^*_{\mathbf{S}}) - q_{\mathbf{S}^c|\mathbf{S}}\right] \\
&\leqslant (\phi^c_{\mathbf{S}})^* \left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} \left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right] + (\phi^c_{\mathbf{S}})^* \left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right] + (\phi^c_{\mathbf{S}})^* \left[-q_{\mathbf{S}^c|\mathbf{S}}\right] \\
&\leqslant \lambda_n (\phi^c_{\mathbf{S}})^* \left[Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} \left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right] + \lambda_n(1-\tau) + \frac{\lambda_n \tau}{2} \\
&\overset{(A.18)}{\leqslant} \lambda_n a (\mathbf{S}^c)^{-1} \left\| Q_{\mathbf{S}^c\mathbf{S}} Q^{-1}_{\mathbf{SS}} \left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty + \lambda_n - \frac{\lambda_n \tau}{2} \\
&\leqslant \frac{\lambda_n \tau}{4} + \frac{\lambda_n \tau}{4} + \lambda_n - \frac{\lambda_n \tau}{2} \leqslant \lambda_n,
\end{aligned}
$$

which is Equation (A.33).

Because Equation (A.32) and Equation (A.33) are satisfied, Lemma 27 implies that $\hat{\beta}^R$ is the optimal. Thus,

$$supp(\hat{\beta}^G) = supp(\hat{\beta}^R) = \mathbf{S}.$$

Part IV

The results in Part III depend on conditions (A.28), (A.29), (A.40), (A.41), (A.42), (A.43), and (A.44), which are summarized as follows:

$$\|Q_{\mathbf{S}^c\mathbf{S}} Q^{-\frac{1}{2}}_{\mathbf{SS}}\|_{2,\infty} \leqslant 3, \tag{A.45}$$

$$\lambda_n |\mathbf{S}|^{\frac{1}{2}} \leqslant \min\left\{ \frac{\gamma_{\min}(Q_{\mathbf{SS}})\, \beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau \gamma^{\frac{3}{2}}_{\min}(Q_{\mathbf{SS}}) a_{\mathbf{S}^c} \beta^*_{\min}}{72 A_{\mathbf{S}} \sum\limits_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \right\}, \tag{A.46}$$

$$(\phi^c_{\mathbf{S}})^* [Q_{\mathbf{S}^c\mathbf{S}} Q_{\mathbf{SS}} \mathbf{r}_{\mathbf{S}}] \leqslant 1 - \tau, \tag{A.47}$$

$$(\phi^{c}_{\mathbf{S}})^{*}[q_{\mathbf{S}^c}|\mathbf{S}] \leqslant \frac{\lambda_n \tau}{2}, \tag{A.48}$$

$$\|q_{\mathbf{S}}\|_{\infty} \leqslant \min\left\{ \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^{*}_{\min}}{3A_{\mathbf{S}}}, \frac{\tau\gamma^{\frac{3}{2}}_{\min}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta^{*}_{\min}}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}} \right\}. \tag{A.49}$$

In Part IV, we want to make sure that these conditions hold with high probability.

Condition (A.45)

To begin with, for any matrix $A \in \mathbb{R}^{m\times n}$, the Cauchy-Schwarz inequality implies that

$$\begin{aligned}
\|A\|_{2,\infty} &= \sup_{\|u\|_2\leqslant 1}\|Au\|_{\infty} = \sup_{\|u\|_2\leqslant 1}\max_{i\in[m]}\Big(\sqrt{\sum_{j\in[n]}A_{ij}u_j}\Big) \\
&\leqslant \sup_{\|u\|_2\leqslant 1}\max_{i\in[m]}\Big(\sqrt{\sum_{j\in[n]}A^2_{ij}}\sqrt{\sum_{j\in[n]}u^2_j}\Big) \\
&\leqslant \max_{i\in[m]}\Big(\sqrt{\sum_{j\in[n]}A^2_{ij}}\Big) \leqslant \max_{i\in[m]}\Big\{\sqrt{\operatorname{diag}(AA^{\top})}\Big\}.
\end{aligned}$$

Recall that $Q = \frac{1}{n}X^{\top}X$. Let $A = Q_{\mathbf{S}^c\mathbf{S}}Q^{-\frac{1}{2}}_{\mathbf{SS}}$, we have

$$\|Q_{\mathbf{S}^c\mathbf{S}}Q^{-\frac{1}{2}}_{\mathbf{SS}}\|_{2,\infty} \leqslant \max\{\sqrt{\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}Q_{\mathbf{S}\mathbf{S}^c})}\}.$$

Using the Schur complement of Q on the block matrices $Q_{\mathbf{SS}}$ and $Q_{\mathbf{S}^c\mathbf{S}^c}$, the positiveness of Q implies the positiveness of $Q_{\mathbf{S}^c\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}Q_{\mathbf{S}\mathbf{S}^c}$. Thus,

$$\max\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}Q_{\mathbf{S}\mathbf{S}^c}) \leqslant \max\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}^c}) \leqslant \max_{j\in\mathbf{S}^c} Q_{jj}.$$

Lemma 29. *(Lemma 1 of Laurent and Massart (2000))*

Suppose that the random variable U follows χ^2 distribution with d degrees of

freedom, then for any positive x,

$$\mathbb{P}(U - d \geq 2\sqrt{dx} + 2x) \leqslant \exp(-x),$$
$$\mathbb{P}(d - U \geq 2\sqrt{dx}) \leqslant \exp(-x).$$

As X follows multivariate normal, $\tilde{Q}_{jj} = \frac{nQ_{jj}}{\Theta^2_{jj}} \sim \chi^2_n$. Then by Lemma 29, we have

$$\begin{aligned}
\mathbb{P}(\max_{j\in\mathbf{S}^c} \sqrt{Q_{jj}} > 3) &\leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c} Q_{jj} > 5) \leqslant \mathbb{P}(\bigcup_{j\in\mathbf{S}^c} Q_{jj} > 5) \leqslant \sum_{j\in\mathbf{S}^c} \mathbb{P}(Q_{jj} > 5) \\
&\leqslant \sum_{j\in\mathbf{S}^c} \mathbb{P}(Q_{jj} > 5\Theta^2_{jj}) = \sum_{j\in\mathbf{S}^c} \mathbb{P}(n\frac{Q_{jj}}{\Theta^2_{jj}} > 5n) \\
&\leqslant \sum_{j\in\mathbf{S}^c} \mathbb{P}(\tilde{Q}_{jj} > n + 2n + 2n) \leqslant (p - |\mathbf{S}|)\exp(-n) \\
&= \exp(-n + \log(p - |\mathbf{S}|)) \\
&\leqslant \exp(-\frac{n}{2}),
\end{aligned} \tag{A.50}$$

where the last inequality holds as $n > 2\log(p - |\mathbf{S}|)$. Thus,

$$\mathbb{P}(\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} > 3) \leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c} \sqrt{Q_{jj}} > 3) \leqslant \exp(-\frac{n}{2}).$$

Similarly, let $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = Q_{\mathbf{S}^c\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}$. The diagonal terms of $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}$ is less than the diagonal terms of $Q_{\mathbf{S}^c\mathbf{S}^c}$, which implies

$$\mathbb{P}(\|Q^{1/2}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_{2,\infty} > 3) \leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c} \sqrt{Q_{jj}} > 3) \leqslant \exp(-\frac{n}{2}).$$

Condition (A.46)

Lemma 30. *(Lemma 9 of Wainwright (2009))*

Suppose that $d \leqslant n$ *and* $X \in \mathbb{R}^{n\times d}$ *have i.i.d rows* $X_i \sim N(0,\Theta)$*, then*

$$\mathbb{P}\left(\gamma_{\max}\left(\frac{1}{n}X^\top X\right) \geqslant 9\gamma_{\max}(\Theta)\right) \leqslant 2\exp(-\frac{n}{2}),$$

$$\mathbb{P}\left(\gamma_{\max}\left((\frac{1}{n}X^\top X)^{-1}\right) \geqslant \frac{9}{\gamma_{\min}(\Theta)}\right) \leqslant 2\exp(-\frac{n}{2}).$$

As we assume that $|\mathbf{S}| \leqslant n$ and $X_{\mathbf{SS}} \sim N\left(0,\Theta_{\mathbf{SS}}\right)$, then Lemma 30 implies

$$\mathbb{P}\left(\gamma_{\max}(Q_{\mathbf{SS}}) \geqslant 9\gamma_{\max}(\Theta_{\mathbf{SS}})\right) \leqslant 2\exp(-\frac{n}{2}),$$

and also

$$\mathbb{P}\left(\gamma_{\min}(\Theta_{\mathbf{SS}}) \geqslant 9\gamma_{\min}(Q_{\mathbf{SS}})\right) \leqslant 2\exp(-\frac{n}{2}).$$

Thus, by assuming that

$$\lambda_n|\mathbf{S}|^{\frac{1}{2}} \leqslant \min\left\{\frac{3\gamma_{\min}(\Theta)\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\tau\gamma^{\frac{3}{2}}_{\min}(\Theta)a_{\mathbf{S}^c}\beta^*_{\min}}{8A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\},$$

we have

$$\lambda_n|\mathbf{S}|^{\frac{1}{2}} \leqslant \min\left\{\frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau\gamma^{\frac{3}{2}}_{\min}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta^*_{\min}}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\}$$

holds with high probability.

Condition (A.47)

For any $j \in \mathbf{S}^c$, $X_j \in \mathbb{R}^n$ is zero-mean Gaussian. Following the decomposition in Wainwright (2009), we have

$$X_j^\top = \Theta_{j\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top + E_j^\top, \tag{A.51}$$

where E_j are i.i.d from $N\left(0, \left[\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right]_{jj}\right)$ with $\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = \Theta_{\mathbf{S}^c\mathbf{S}^c} - \Theta_{\mathbf{S}^c\mathbf{S}}\left(\Theta_{\mathbf{SS}}\right)^{-1}\Theta_{\mathbf{SS}^c}$. Let $E_{\mathbf{S}^c}$ be an $|S^c| \times n$ matrix, with each row representing E_j for an element $j \in \mathbf{S}^c$, then we have

$$\begin{aligned} Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\mathbf{r}_{\mathbf{S}} &= X_{\mathbf{S}^c}^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}} \\ &\overset{(A.51)}{=} \left(\Theta_{\mathbf{S}^c\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top + E_{\mathbf{S}^c}^\top\right)X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}} \\ &= \Theta_{\mathbf{S}^c\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\mathbf{r}_{\mathbf{S}} + E_{\mathbf{S}^c}^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}} \\ &:= \Theta_{\mathbf{S}^c\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\mathbf{r}_{\mathbf{S}} + \eta. \end{aligned} \tag{A.52}$$

The preceding expression prompts us to establish an upper bound for the dual norm of η. To achieve this, we begin by examining the scenario in which $X_{\mathbf{S}}$ is fixed. Our objective now is to derive the covariance matrix of η. For any $j \in \mathbf{S}^c$, we have

$$\mathbb{E}[\eta_j] = \mathbb{E}\left[E_j^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\right] = 0.$$

For any pair of $j, k \in \mathbf{S}^c$, we have

$$\begin{aligned} \mathbb{E}[\eta_j\eta_k] &= \mathbb{E}\left[E_j^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}E_k^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\right] \\ &= \mathbb{E}\left[\mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top E_jE_k^\top X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\right] \\ &= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top\mathbb{E}\left[E_jE_k^\top\right]X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}, \end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}\left[E_j E_k^\top\right] \overset{(A.51)}{=} \mathbb{E}\left[\left(X_j - X_{\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\Theta_{j\mathbf{S}}^\top\right)\left(X_k^\top - \Theta_{k\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top\right)\right] \\
&= \mathbb{E}\left[X_j X_k^\top\right] - \mathbb{E}\left[X_{\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\Theta_{j\mathbf{S}}X_k^\top\right] - \mathbb{E}\left[X_j\Theta_{k\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top \mid X_{\mathbf{S}}\right] + \mathbb{E}\left[X_{\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\Theta_{j\mathbf{S}}^\top\Theta_{k\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top\right] \\
&= \mathbb{E}\left[X_j X_k^\top\right] - X_{\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\Theta_{j\mathbf{S}}\mathbb{E}\left[X_k^\top\right] - \mathbb{E}\left[X_j\right]\Theta_{k\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top + X_{\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}\Theta_{j\mathbf{S}}\Theta_{k\mathbf{S}}\Theta_{\mathbf{SS}}^{-1}X_{\mathbf{S}}^\top \\
&= \mathbb{E}\left[X_j X_k^\top\right] - \mathbb{E}\left[X_j\right]\mathbb{E}\left[X_k^\top\right] = \mathrm{Cov}\left[X_j, X_k^\top\right] = \left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk}\mathbf{I}_{n\times n}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}[\eta_j\eta_k] &= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top\mathbb{E}\left[E_j E_k^\top\right]X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}} \\
&= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk}\mathbf{I}_{n\times n}X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}} \\
&= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk} = \frac{\mathbf{r}_{\mathbf{S}}^\top(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk}.
\end{aligned}$$

And we have $\mathrm{Cov}(\eta) = \frac{\mathbf{r}_{\mathbf{S}}^\top(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right) := \Xi$.

Lemma 31. *(Theorem 2.26 in Wainwright (2019))*

Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a Lipschitz function with respect to the Euclidean norm and Lipschitz constant L. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geqslant t] \leqslant 2\exp(-\frac{t^2}{2L^2}) \quad \textit{for all } t \geqslant 0.$$

To apply the concentration bound in Lemma 31, we define function $\Psi(u) = (\phi_{\mathbf{S}^c}^*)\left[\Xi^{\frac{1}{2}}u\right]$. As $\eta = \Xi^{\frac{1}{2}}W$ where $W \sim N(0, I_{|\mathbf{S}^c|\times|\mathbf{S}^c|})$, $(\phi_{\mathbf{S}}^c)^*(\eta)$ has the same distribution as $\Psi(W)$. We continue to show that Ψ is a Lipschitz function given fixed $X_{\mathbf{S}}$.

$$
\begin{aligned}
|\Psi(u)-\Psi(v)| \leqslant \Psi(u-v) &= (\phi_{\mathbf{S}}^{c})^{*}\left[\Xi^{\frac{1}{2}}(u-v)\right] \\
&\leqslant a_{\mathbf{S}}^{-1}\left\|\Xi^{\frac{1}{2}}(u-v)\right\|_{\infty} \\
&= a_{\mathbf{S}}^{-1}\left\|\left[\frac{\mathbf{r}_{\mathbf{S}}^{\top}(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^{c}\mathbf{S}^{c}|\mathbf{S}}\right)\right]^{\frac{1}{2}}(u-v)\right\|_{\infty} \\
&\leqslant a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_{2} n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^{c}\mathbf{S}^{c}|\mathbf{S}}\right)\|u-v\|_{2}.
\end{aligned}
$$

Thus, the corresponding Lipstichiz constant is

$$
L_{\eta}=a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_{2} n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^{c}\mathbf{S}^{c}|\mathbf{S}}\right).
$$

On the other hand, suppose that $\mathbb{E}\left[(\phi_{\mathbf{S}}^{c})^{*}(\eta)\right] \leqslant \frac{\tau}{4}$, since Ψ is a Lipschitiz function, by applying $t=\frac{\tau}{4}$ in concentration Lemma 31 on Lipschitz functions of multivariate standard random variables, we have

$$
\begin{aligned}
\mathbb{P}\left((\phi_{\mathbf{S}}^{c})^{*}[\eta]>\frac{\tau}{2}\right) &= \mathbb{P}\left(\Psi(W)>\frac{\tau}{2}\right)=\mathbb{P}\left(\Psi(W)-\frac{\tau}{4}>\frac{\tau}{4}\right) \\
&\leqslant \mathbb{P}\left(\Psi(W)-E\left[(\phi_{\mathbf{S}}^{c})^{*}(\eta)\right]>\frac{\tau}{4}\right) \\
&= \mathbb{P}\left(\Psi(W)-E\left[\Psi(W)\right]>\frac{\tau}{4}\right)\leqslant \exp\left(-\frac{\tau^{2}}{4L_{\eta}^{2}}\right).
\end{aligned}
$$

Now we further assume that $\{\gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\}$. Under this condition, we have

$$
L_{\eta}=a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_{2} n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^{c}\mathbf{S}^{c}|\mathbf{S}}\right)\leqslant\frac{3a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_{2}\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^{c}\mathbf{S}^{c}|\mathbf{S}}\right)}{\left(n\gamma_{\min}(\Theta_{\mathbf{SS}})\right)^{\frac{1}{2}}}. \quad \text{(A.53)}
$$

Lemma 32. *(Sudakov inequality, Theorem 5.27 in Wainwright (2019))*

If X and Y are a.s. bounded, centered Gaussian processes on T such that

$$
\mathbb{E}\left(X_{t}-X_{s}\right)^{2}\leq\mathbb{E}\left(Y_{t}-Y_{s}\right)^{2}
$$

then

$$\mathbb{E}\sup_T X_t \leq \mathbb{E}\sup_T Y_t.$$

Lemma 33. *(Exercise 2.12 in Wainwright (2019)) Let $X_1, \ldots, X_n$ be independent σ^2-subgaussian random variables. Then*

$$\mathbb{E}[\max_{1\leq i\leq n} |X_i|] \leq 2\sqrt{\sigma^2 \log n}.$$

On the other hand, for any u_t, u_s, we have

$$\begin{aligned}&\mathbb{E}(u_t^\top \eta - u_s^\top \eta)^2 = \mathbb{E}(u_t^\top \Xi^{\frac{1}{2}} W - u_s^\top \Xi^{\frac{1}{2}} W)^2 = (u_t - u_s)^\top \Xi (u_t - u_s)\\ \leqslant &||u_t - u_s||_2^2 \gamma_{\max}(\Xi) = \mathbb{E}(\gamma_{\max}^{\frac{1}{2}}(\Xi)\, u_t^\top W - \gamma_{\max}^{\frac{1}{2}}(\Xi)\, u_s^\top W)^2\end{aligned}$$

By using Sudakov-Fernique inequality in Lemma 32, we have

$$\mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \Xi^{\frac{1}{2}} W\Big] \leqslant \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} \gamma_{\max}^{\frac{1}{2}}(\Xi)\, u^\top W\Big]$$

Consequently,

$$\begin{aligned}\mathbb{E}\Big[(\phi_{\mathbf{S}}^c)^*(\eta)\Big] &= \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \eta\Big] = \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \Xi^{\frac{1}{2}} W\Big]\\ &\leqslant \gamma_{\max}^{\frac{1}{2}}(\Xi)\, \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top W\Big] = \gamma_{\max}(\Xi)^{\frac{1}{2}}\, \mathbb{E}\Big[(\phi_{\mathbf{S}}^c)^*(W)\Big].\end{aligned} \tag{A.54}$$

Notice that

$$
\begin{aligned}
\|\mathbf{r}_{\mathbf{S}}\|_2^2 &\leqslant |\mathbf{S}| \max_{j\in\mathbf{S}} \mathbf{r}_j^2 = |\mathbf{S}|\Big(\max_{j\in\mathbf{S}}\{\beta_j^* \cdot \sum_{g\in \mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\|\beta^*_{G_g\cap\mathbf{S}}\|_2}\}\Big)^2 \\
&\leqslant |\mathbf{S}|\Big(\max_{j\in\mathbf{S}}\{|\beta_j^*|\} \cdot \max\{ \sum_{g\in \mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\|\beta^*_{G_g\cap\mathbf{S}}\|_2}\}\Big)^2 \\
&\leqslant |\mathbf{S}|\Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot \max\{ \sum_{g\in \mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\sqrt{|G_g\cap\mathbf{S}}}\}\Big)^2 \\
&\leqslant |\mathbf{S}|\Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot \max\{ \sum_{g\in \mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} w_g\}\Big)^2 \\
&\leqslant |\mathbf{S}|\Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}^G} w_g\}\Big)^2 \\
&\leqslant \Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}}\Big)^2 |\mathbf{S}| A_{\mathbf{S}}^2 = \max_{j\in\mathbf{S}}\{(\beta_j^*)^2\}|\mathbf{S}|\Big(\frac{A_{\mathbf{S}}}{\beta^*_{\min}}\Big)^2 \\
&\lesssim \frac{\max_{j\in\mathbf{S}}\{(\beta_j^*)^2\}}{\lambda_n^2}
\end{aligned}
\tag{A.55}
$$

Thus, if $X_{\mathbf{S}}$ satisfies $\gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}$, we have

$$
\begin{aligned}
\mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(\eta)\right] &\overset{(A.54)}{\leqslant} \gamma_{\max}\left(\Xi\right)^{\frac{1}{2}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2 \, \gamma_{\min}^{-\frac{1}{2}}\left(Q_{\mathbf{SS}}\right) \gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{n^{\frac{1}{2}}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2 \, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\overset{(A.18)}{\leqslant} \frac{\|\mathbf{r}_{\mathbf{S}}\|_2 \, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[a_{\mathbf{S}^c}^{-1}\|W\|_\infty\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2 \, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[\|W\|_\infty\right] \\
&\overset{\text{Lemma 33}}{\leqslant} \frac{6\,\|\mathbf{r}_{\mathbf{S}}\|_2 \, \gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \sqrt{\log(p-|\mathbf{S}|)} \leqslant \frac{\tau}{4},
\end{aligned}
\tag{A.56}
$$

where the last inequality holds as Assumption 6 implies that

$$n \gtrsim \frac{\max_{j\in\mathbf{S}}\{(\beta_j^*)^2\}\log(p-|\mathbf{S}|)}{a_{\mathbf{S}^c}^2\lambda_n^2} \overset{(A.55)}{\gtrsim} \frac{\|\mathbf{r}_{\mathbf{S}}\|_2^2\log(p-|\mathbf{S}|)}{a_{\mathbf{S}^c}^2} \geqslant \frac{576\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\log(p-|\mathbf{S}|)\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}^2\gamma_{\min}(\Theta_{\mathbf{SS}})\tau^2}.$$

Consequently, Equation (A.53) and (A.56) together implies

$$\begin{aligned}&\mathbb{P}\Big(\,(\phi_{\mathbf{S}}^c)^*\,[\eta] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big)\\ &\leqslant \exp\Big(-\frac{\tau^2}{4L_\eta^2}\Big) \leqslant \exp\Big(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{12\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\Big).\end{aligned} \tag{A.57}$$

Thus, let $\mathscr{A}$ be the event $\{X_{\mathbf{S}} \mid \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\}$. We have

$$\begin{aligned}\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\,[\eta] > \frac{\tau}{2} \mid X_{\mathbf{S}}\right) &= \mathbb{P}\Big(\,(\phi_{\mathbf{S}}^c)^*\,[\eta] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big)\\ &\quad+\mathbb{P}\Big(\,(\phi_{\mathbf{S}}^c)^*\,[\eta] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) > \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big)\\ &\leqslant \exp\left(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right) + \mathbb{P}\left(\mathscr{A}^c\right)\\ &\leqslant \exp\left(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right) + 2\exp(-\frac{n}{2}).\end{aligned}$$

Condition (A.48)

Now we are going to study condition (A.48). Recall that $q_{\mathbf{S}^c|\mathbf{S}} = q_{\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}$ and $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = Q_{\mathbf{S}^c\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}$. Given X, $q_{\mathbf{S}^c|\mathbf{S}}$ is a centered Gaussian random vector with covariance matrix

$$\begin{aligned}
\mathbb{E}\left[q_{\mathbf{S}^c|\mathbf{S}}q_{\mathbf{S}^c|\mathbf{S}}^\top\right] &= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top - q_{\mathbf{S}^c}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}^c}^\top + Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right] \\
&= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right] \\
&= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top\right] - \mathbb{E}\left[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right] \\
&= \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}^c} - \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c} := \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}.
\end{aligned}$$

Next, we define $\psi(u) = (\phi_{\mathbf{S}}^c)^*\left(\sigma n^{-1/2}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}u\right)$ so that $(\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right]$ has the same distribution as $\psi(W)$. Now we want to show that ψ is a Lipschitz function

$$\begin{aligned}
|\psi(u) - \psi(v)| \leqslant \psi(u-v) &= (\phi_{\mathbf{S}}^c)^*\left(\sigma n^{-1/2}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}(u-v)\right) \\
&\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}(u-v)\right\|_\infty \\
&\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}\right\|_{2,\infty}\|(u-v)\|_\infty \\
&\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}\right\|_{2,\infty}\|(u-v)\|_2
\end{aligned}$$

Suppose that $\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} \leqslant 3$, then ψ is a Lipschitz function with Lipschitz constant $3\sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}$. In addition, if $\mathbb{E}[(\phi_{\mathbf{S}}^c)^*(q_{\mathbf{S}^c|\mathbf{S}})] \leqslant \frac{\lambda_n\tau}{4}$, then by Lemma 31 , we have for $t = \frac{\lambda_n\tau}{4}$,

$$\mathbb{P}\left((\phi^c_{\mathbf{S}})^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2}\right) = \mathbb{P}\left(\psi(W) > \frac{\lambda_n\tau}{2}\right) = \mathbb{P}\left(\psi(W) - \frac{\lambda_n\tau}{4} > \frac{\lambda_n\tau}{4}\right)$$
$$\leqslant \mathbb{P}\left(\psi(W) - \mathbb{E}[(\phi^c_{\mathbf{S}})^*(q_{\mathbf{S}^c|\mathbf{S}})] > \frac{\lambda_n\tau}{4}\right)$$
$$= \mathbb{P}\left(\psi(W) - \mathbb{E}\left[\psi(W)\right] > \frac{\lambda_n\tau}{4}\right) \leqslant \exp\left(-\frac{\tau^2\lambda_n^2 n a^2_{\mathbf{S}^c}}{144\sigma^2}\right).$$

Now, we consider random X. For any u_t, u_s, we have

$$\mathbb{E}\big[(u_t-u_s)^\top q_{\mathbf{S}^c|\mathbf{S}}\big]^2 = \frac{\sigma^2}{n}(u_t-u_s)^\top Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}(u_t-u_s) \leqslant \frac{\sigma^2}{n}\big\|Q^{\frac{1}{2}}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\big\|_2^2\big\|(u_t-u_s)\big\|_2^2$$
$$= \mathbb{E}\big[\sigma n^{-\frac{1}{2}}\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}}(u_t-u_s)^\top W\big]^2$$

By using Sudakov-Fernique inequality, if $\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2 \leqslant 9$, we get

$$\begin{aligned}
\mathbb{E}[(\phi^c_{\mathbf{S}})^*(q_{\mathbf{S}^c|\mathbf{S}})] &= \mathbb{E}\sup_{\phi^c_{\mathbf{S}}(u)\leq 1} u^\top q_{\mathbf{S}^c|\mathbf{S}} \\
&\leqslant \sigma n^{-1/2}\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}}\mathbb{E}\sup_{\phi^c_{\mathbf{S}}(u)\leq 1} u^\top W \\
&\leqslant \sigma n^{-\frac{1}{2}}\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}}\mathbb{E}\left[(\phi^c_{\mathbf{S}})^*(W)\right] \\
&\leqslant 3\sigma n^{-\frac{1}{2}}\mathbb{E}\left[(\phi^c_{\mathbf{S}})^*(W)\right] \\
&\leqslant \frac{\lambda_n\tau}{4}.
\end{aligned} \tag{A.58}$$

On the other hand, Assumption 1' and 6 imply that

$$\frac{9\sigma^2\mathbb{E}^2\left[(\phi^c_{\mathbf{S}})^*(W)\right]}{n} \leqslant \frac{9\sigma^2\log(p-|\mathbf{S}|)}{a^2_{\mathbf{S}^c}n} \leqslant \frac{\lambda_n^2\tau^2}{16}.$$

Therefore, we have

$$\mathbb{P}\left((\phi^c_{\mathbf{S}})^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q^{1/2}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right\|_{2,\infty} \leqslant 3\right) \leqslant \exp\left(-\frac{\tau^2 n\lambda_n^2 a^2_{\mathbf{S}^c}}{144\sigma^2}\right).$$

Let $\mathscr{B}$ be the event $\{X \mid \left\|Q^{1/2}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right\|_{2,\infty} \leqslant 3\}$. We have

$$\begin{aligned}\mathbb{P}\left((\phi^c_{\mathbf{S}})^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X\right) &= \mathbb{P}\left((\phi^c_{\mathbf{S}})^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q^{1/2}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right\|_{2,\infty} \leqslant 3\right) \\ &\quad + \mathbb{P}\left((\phi^c_{\mathbf{S}})^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q^{1/2}_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right\|_{2,\infty} > 3\right) \\ &\leqslant \exp\left(-\frac{\tau^2 n\lambda_n^2 a^2_{\mathbf{S}^c}}{144\sigma^2}\right) + \mathbb{P}\left(\mathscr{B}^c\right) \\ &\overset{(A.45)}{\leqslant} \exp\left(-\frac{\tau^2 n\lambda_n^2 a^2_{\mathbf{S}^c}}{144\sigma^2}\right) + \exp(-\frac{n}{2}).\end{aligned}$$

Condition (A.49)

The last condition (A.49) lead us to control the term $\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G)\right)$, with

$$c'(\mathbf{S}, G) = \min\left\{\frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau\gamma^{\frac{3}{2}}_{\min}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta^*_{\min}}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g \cap \mathbf{S}|}}\right\}.$$

For any given X, Jenatton et al. (2011a) showed that for any $\delta > 0$,

$$\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant \delta\right) \leqslant 2|\mathbf{S}|\exp\left(-\frac{n\delta^2}{2\sigma^2}\right).$$

Recall under the event $\mathscr{A}$, we have

$$\frac{\gamma_{\min}(\Theta_{\mathbf{SS}})}{9} \leqslant \gamma_{\min}(Q_{\mathbf{SS}}).$$

Which implies that

$$c'(\mathbf{S}, G) \geqslant \min \left\{ \frac{\gamma_{\min}\left(\Theta_{\mathbf{SS}}\right) \beta^*_{\min}}{27 A_{\mathbf{S}}}, \frac{\tau \gamma_{\min}(\Theta_{\mathbf{SS}})^{\frac{3}{2}} a_{\mathbf{S}^c} \beta^*_{\min}}{648 A_{\mathbf{S}} \sum\limits_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \right\}$$
$$\geqslant \min \left\{ \frac{\beta^*_{\min}}{27 c_1 A_{\mathbf{S}}}, \frac{\tau a_{\mathbf{S}^c} \beta^*_{\min}}{648 c_1^{\frac{3}{2}} A_{\mathbf{S}} \sum\limits_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \right\} := c(\mathbf{S}, G).$$

Thus, consider random X, we have

$$\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right) \leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c(\mathbf{S}, G) \mid \mathscr{A}\right) \leqslant 2|\mathbf{S}| \exp\left(-\frac{n c^2(\mathbf{S}, G)}{2\sigma^2}\right)$$

Thus,

$$\begin{aligned}
\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G)\right) &= \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}\right) + \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}^c\right) \\
&\leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&= \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right) \mathbb{P}\left(\mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&\leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&\leqslant 2|\mathbf{S}| \exp\left(-\frac{n c^2(\mathbf{S}, G)}{2\sigma^2}\right) + 2\exp(-n/2).
\end{aligned}$$

In summary, the probability of one of the conditions being violated is upper bound by

$$8\exp(-\frac{n}{2}) + \exp\left(-\frac{n a_{\mathbf{S}}^2 \tau^2 \gamma_{\max}(\Theta_{\mathbf{SS}})}{4\left\|\mathbf{r}_{\mathbf{S}}\right\|_2^2 \gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right) + \exp\left(-\frac{n\lambda_n^2 \tau^2 a_{\mathbf{S}^c}^2}{32\sigma^2 c_2^4}\right) + 2|\mathbf{S}| \exp\left(-\frac{n c^2(\mathbf{S}, G)}{2\sigma^2}\right).$$

Part V

First, given the original group structure G and its induced counterpart $\mathscr{G}$, along with their respective weights w and $\mathscr{w}$, we consider the scenario where $\mathbf{J} = \mathbf{S}$. For all $\beta \in \mathbb{R}^p$, we have

$$\begin{aligned}
\phi_{\mathbf{S}}^{G}(\beta_{\mathbf{S}}) &= \sum_{g \in \mathsf{G}_{\mathbf{S}}^{G}} w_g \|\beta_{\mathbf{S} \cap G_g}\|_2 \leqslant \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \Big(\sum_{\mathscr{g}: \mathscr{g} \in F^{-1}(g), \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}} \|\beta_{\mathbf{S} \cap \mathscr{G}_{\mathscr{g}}}\|_2 \Big) \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}} \Big(\sum_{g: g \in F(\mathscr{g}), g \in \mathsf{G}_{\mathbf{S}}} w_g \Big) \|\beta_{\mathbf{S} \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}} \Big(\sum_{g: g \in F(\mathscr{g})} w_g \Big) \|\beta_{\mathbf{S} \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g} \in \mathsf{G}_{\mathbf{S}}^{\mathscr{G}}} w_{\mathscr{g}} \|\beta_{\mathbf{S} \cap \mathscr{G}_{\mathscr{g}}}\|_2 = \phi_{\mathbf{S}}^{\mathscr{G}}(\beta).
\end{aligned} \tag{A.59}$$

Since $\phi_{\mathbf{S}}^{G}(\beta) \leqslant \phi_{\mathbf{S}}^{\mathscr{G}}(\beta)$, we can set $a_{\mathbf{S}}^{\mathscr{G}} = a_{\mathbf{S}}^{G} = \min_{g \in \mathsf{G}_{\mathbf{S}}^{G}} \frac{w_g}{\sqrt{d_g}}$. Since

$$\max_{\mathscr{g} \in \mathsf{G}_{\mathbf{S}}^{\mathscr{G}}} w_{\mathscr{g}} = \max_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \cap \mathbf{S} \neq \emptyset} \sum_{g \in F(\mathscr{g})} w_g \leqslant h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g \in \mathsf{G}_{\mathbf{S}}^{G}} w_g,$$

we can set $A_{\mathbf{S}}^{\mathscr{G}} = A_{\mathbf{S}}^{G}$. On the other hand, for all $\beta \in \mathbb{R}^p$, we have

$$\begin{aligned}
(\phi_{\mathbf{S}}^{G})^c(\beta_{\mathbf{S}}^c) &= \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}^{G}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2 \leqslant \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}} w_g \Big(\sum_{\mathscr{g}: \mathscr{g} \in F^{-1}(g), \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \Big) \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \Big(\sum_{g: g \in F(\mathscr{g}), g \in [m] \setminus \mathsf{G}_{\mathbf{S}}^{G}} w_g \Big) \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \Big(\sum_{g: g \in F(\mathscr{g})} w_g \Big) \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g} \in [m] \setminus \mathsf{G}_{\mathbf{S}}^{\mathscr{G}}} w_{\mathscr{g}} \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 = (\phi_{\mathbf{S}}^{\mathscr{G}})^c(\beta).
\end{aligned} \tag{A.60}$$

Consequently, with an trivial extension, we can set $a^{\mathscr{G}}_{\mathbf{S}^c} = a^{G}_{\mathbf{S}^c} \leqslant \min_{g \in \mathsf{G}^{G}_{\mathbf{S}^c}} w_g/\sqrt{d_g}$. Based on the result of Theorem 5.1, Equation (2.28) holds if

$$\lambda_n |\mathbf{S}|^{\frac{1}{2}} \lesssim \min\Big\{ \frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S}^c}}{A_{\mathbf{S}} \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}} \Big\}.$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|} &\leqslant \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sum_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \\ &= \sum_{\mathscr{g} \in F^{-1}(g), g \in \mathsf{G}_{\mathbf{S}}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \big(\sum_{g \in F(\mathscr{g})} w_g \big) \\ &= \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \end{aligned}$$

If $F^{-1}(g) = O(1)$ for every $g \in \mathsf{G}_{\mathbf{S}}$, we have

$$|G_g \cap \mathbf{S}| = \sum_{\mathscr{g} \in F^{-1}(g)} |\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}| \asymp \big(\sum_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \big)^2.$$

Consequent, we have $\sqrt{|G_g \cap \mathbf{S}|} \asymp \sum_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}$,

$$\sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|} \asymp \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|},$$

and

$$\min\Big\{ \frac{\beta^*_{\min}}{A^G_{\mathbf{S}}}, \frac{\beta^*_{\min} a^G_{\mathbf{S}^c}}{A^G_{\mathbf{S}} \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \Big\} \asymp \min\Big\{ \frac{\beta^*_{\min}}{A^{\mathscr{G}}_{\mathbf{S}}}, \frac{\beta^*_{\min} a^{\mathscr{G}}_{\mathbf{S}^c}}{A^{\mathscr{G}}_{\mathbf{S}} \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}} \Big\}.$$

C Appendix for Chapter III

B.1 Closed-form solution for (3.1)

First, we have

$$\Omega^G(U) = \min \sum_{g=1}^{m} w_g \left\|U_g\right\|_2, \text{ s.t } \sum_{g=1}^{m} U_g = U.$$

Consequently, we have

$$\begin{aligned}
&\underset{U\in\mathbb{R}^{p\times\hat{k}}}{\operatorname{argmin}} \frac{1}{2}\|U_0 - U\|_F^2 + \lambda\Omega^G(U) \\
=&\operatorname{argmin} \frac{1}{2}\left\|U_0 - \sum_{g=1}^{m} U_g\right\|_F^2 + \lambda\Omega^G(U).
\end{aligned}$$

Now we consider optimizing with one grup g_0 . While fix other $m-1$ group.

$$\begin{aligned}
&\underset{U_{g_0}\in R^{d_{g_0}\times\hat{k}}}{\operatorname{argmin}} \frac{1}{2}\left\|U_0 - \sum_{g=1}^{m} U_g\right\|_F^2 + \lambda\Omega^G(U) \\
=&\operatorname{argmin} \frac{1}{2}\left\|U_0 - \sum_{g\neq g_0} U_g - U_{g_0}\right\|_F^2 + \lambda\omega_{g_0}\left\|U_{g_0}\right\|_2 \\
=&\operatorname{argmin} F\left(U_{g_0}\right)
\end{aligned}$$

By taking derivation with respect to U_{g_0} , we have

$$\nabla F\left(U_{g_0}\right) = U_{g_0} - \left(U_0 - \sum_{g\neq g_0} U_g\right) + \lambda W_{g_0} S_{g_0}.$$

where

$$\begin{cases}
S_{g_0} = \frac{U_{g_0}}{\left\|U_{g_0}\right\|_2} & \text{if } \left\|U_{g_0}\right\|_2 \neq 0 \\
\left\|S_{g_0}\right\|_2 \leq 1 & \text{if } \left\|U_{g_0}\right\|_2 = 0
\end{cases}$$

If $\|U_{g_0}\|_2 \neq 0$, by K.K.T condition:

$$\begin{aligned}
U_{g_0} - U_0 + \sum_{g \neq g_0} U_g + \lambda w_{g_0} \frac{U_{g_0}}{\|U_{g_0}\|_2} &= 0 \\
\Longrightarrow \quad \left(1 + \frac{\lambda w_{g_0}}{\|U_{g_0}\|_2}\right) U_{g_0} &= U_0 - \sum_{g \neq g_0} U_g \\
\Longrightarrow \quad U_{g_0} &= \left(\frac{\|U_{g_0}\|_2 + \lambda W_{g_0}}{\|U_{g_0}\|_2}\right)^{-1} \left(U_0 - \sum_{g \neq g_0} U_g\right) \\
\Longrightarrow \quad \|U_{g_0}\|_2 &= \|U_{g_0}\|_2 \frac{\|U_0 - \sum_{g \neq g_0} U_g\}\|_2}{\|U_{g_0}\|_2 + \lambda w_{g_0}} \\
\Longrightarrow \quad ||(U_{g_0})|_2 + \lambda W_{g_0} &= -||U_0 - \sum_{g \neq g_0} U_g||_2 \\
||U_{g_0}||_2 &= ||U_0 - \sum_{g \neq g_0} U_g||_2 - \lambda w_{g_0}.
\end{aligned}$$

Thus

$$\begin{aligned}
U_{g_0} &= -\left(1 + \frac{\lambda W g_0}{||U_{g_0}||_2}\right)^{-1} \left(U_0 - \sum_{g \neq g_0} U_g\right) \\
&= \left(1 + \frac{\lambda W_{g_0}}{||U_0 + \sum_{g \neq g_0} U_g||_2 - \lambda W g_0}\right)^{-1} \left(U_0 - \sum_{g \neq g_0} U_g\right) \\
&= \left(1 - \frac{\lambda W_{g_0}}{||U_0 - \sum_{g \neq g_0} U_g||_2}\right) \left(U_0 - \sum_{g \neq g_0} U_g\right).
\end{aligned}$$

B.2 Proofs

Lemma 34 (Theorem 6.1 in Wainwright (2019)). *Let $X \in \mathbb{R}^{n\times d}$ be drawn according to the Σ-Gaussian ensemble. Then for all $\delta > 0$, the maximum singular value $\sigma_{\max}(X)$ satisfies the upper deviation inequality*

$$\mathbb{P}\left[\frac{\sigma_{\max}(X)}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\Sigma})(1+\delta) + \sqrt{\frac{tr(\Sigma)}{n}}\right] \leq e^{-n\delta^2/2}.$$

Moreover, for $n \geq d$, the minimum singular value $\sigma_{\min}(X)$ satisfies the analogous lower deviation inequality

$$\mathbb{P}\left[\frac{\sigma_{\min}(X)}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\Sigma})(1-\delta) - \sqrt{\frac{tr(\Sigma)}{n}}\right] \leq e^{-n\delta^2/2}.$$

Proof of Proposition 6. To begin with, we have

$$\hat{D} = \hat{\Sigma}_1 - \hat{\Sigma}_2 = \frac{X_1^T X_1}{n_1} - \frac{X_2^T X_2}{n_2}.$$

Consequently,

$$\begin{aligned}
\gamma_{\max}(\hat{D}) &\leq \gamma_{\max}\left(\frac{X_1^T X_1}{n_1}\right) + \gamma_{\max}\left(\frac{X_2^T X_2}{n_2}\right) \\
&= \frac{\gamma_{\max}^2(X_1)}{n_1} + \frac{\gamma_{\max}^2(X_2)}{n_2} - \gamma_{\max}(\Sigma_1) - \gamma_{\max}(\Sigma_2) \\
&\quad + \gamma_{\max}(\Sigma_1) + \gamma_{\max}(\Sigma_2).
\end{aligned}$$

Let $A \subset [p]$ be any set, we have

$$\begin{aligned}\gamma_{\max}(\hat{D}_{A,A}) \leq & \frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) \\ & + \frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) \\ & + \gamma_{\max}(\Sigma_{1,A,A}) + \gamma_{\max}(\Sigma_{2,A,A}).\end{aligned}$$

Suppose that $\gamma_{\max}(\Sigma_1) \leq c_1$ and $\gamma_{\max}(\Sigma_2) \leq c_2$. Then we have

$$\gamma_{\max}(\hat{D}_{A,A}) \leq \frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) + \frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) + c_1 + c_2.$$

By lemma 34, for all $\delta > 0$:

$$\begin{aligned}Pr\left\{\frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) \leq \frac{tr(\Sigma_{1,A,A})}{n_1} + \gamma_{\max}(\Sigma_{1,A,A})\delta\right\} \geq 1 - \exp(-n\delta^2/2). \\ Pr\left\{\frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) \leq \frac{tr(\Sigma_{2,A,A})}{n_2} + \gamma_{\max}(\Sigma_{2,A,A})\delta\right\} \geq 1 - \exp(-n\delta^2/2).\end{aligned}$$

Thus, let $n = \min\{n_1, n_2\}/2$,

$$\begin{aligned}&\Pr\left\{\gamma_{\max}(\hat{D}_{A,A}) \leq \frac{|A|}{n} + (c_1 + c_2)(1+\delta)\right\} \\ &\geq \Pr\left\{\frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) + \frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) \leq \frac{|A|}{n} + (c_1 + c_2)\delta\right\} \\ &\geq \Pr\left\{\frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) \leq \frac{|A|}{2n} + c_1\delta\right\} \times \Pr\left\{\frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) \leq \frac{|A|}{2n} + c_2\delta\right\} \\ &\geq \Pr\left\{\frac{\gamma^2_{\max}(X_{1,A,A})}{n_1} - \gamma_{\max}(\Sigma_{1,A,A}) \leq \frac{tr(\Sigma_{1,A,A})}{n_1} + c_1\delta\right\} \\ &\quad \cdot \Pr\left\{\frac{\gamma^2_{\max}(X_{2,A,A})}{n_2} - \gamma_{\max}(\Sigma_{2,A,A}) \leq \frac{tr(\Sigma_{2,A,A})}{n_2} + c_2\delta\right\} \\ &= \left(1 - \exp\left(-n\delta^2/2\right)\right)^2 = 1 - 2\exp\left(-n\delta^2/2\right) + \exp\left(-n\delta^2\right) \geq 1 - 2\exp\left(-n\delta^2/2\right).\end{aligned}$$

□

Proof of Proposition 7. Recall that

$$\begin{aligned}\theta(\hat{S}_t) - \theta(\hat{S}_{t+1}) &= \frac{\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t})}{|\hat{S}_t|} - \frac{\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})}{|\hat{S}_{t+1}|} \\ &= \frac{\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t})}{|\hat{S}_t|} - \frac{\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})}{|\hat{S}_t|}\frac{|\hat{S}_t|}{|\hat{S}_{t+1}|}\end{aligned}$$

Suppose that $\hat{S}_{t+1} = \hat{S}_t \setminus \Delta_t$. By Taylor's expansion we have

$$\frac{|\hat{S}_t|}{|\hat{S}_{t+1}|} = \frac{|\hat{S}_t|}{|\hat{S}_t \setminus \Delta_t|} = \frac{1}{1 - \frac{|\Delta_t|}{|\hat{S}_t|}} \approx 1 + \frac{|\Delta_t|}{|\hat{S}_t|}.$$

Thus,

$$\begin{aligned}\theta(\hat{S}_t) - \theta(\hat{S}_{t+1}) &= \frac{\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t})}{|\hat{S}_t|} - \frac{\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})}{|\hat{S}_t|}\left(1 + \frac{|\Delta_t|}{|\hat{S}_t|}\right) \\ &= \frac{\gamma_{\max}(\hat{D}_{\hat{S}_t,\hat{S}_t}) - \gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}})}{|\hat{S}_t|} + \frac{\gamma_{\max}(\hat{D}_{\hat{S}_{t+1},\hat{S}_{t+1}}) \cdot |\Delta_t|}{|\hat{S}_t|^2} + o(|\Delta_t|).\end{aligned}$$

□

B Appendix for Chapter IV

Abstract

The supplementary material is organized as follows: Section C.1 presents additional theoretical results. Section C.2 presents two examples to assist readers in understanding Definition C.63. Section C.2 provides the proofs for the results outlined in Section C.1. Section C.2 contains the proofs for the results discussed in Section C.1. Section C.2 offers the proofs for the results presented in Section C.1. Section C.2 elucidates the proofs for the results shown in Section C.1. The consistency results is shown in Section C.2 and Section C.2. Finally, the additional simulation results are shown in Section C.9.

We first introduce a few notations we used in the supplement material. Denote the set $\{1, \cdots, n\}$ for any positive integer n by $[n]$. Given a set $\mathcal{S}$, $|\mathcal{S}|$ is the cardinality of the set. Given a matrix A, $A_{\mathcal{S},\mathcal{S}}$ denotes the submatrix formed by selecting the rows and columns indexed by $\mathcal{S}$. Let G be an undirected unweighted graph whose node set is $V(G) = \{v_1, \ldots, v_n\}$ and edge set is $\mathscr{E}(G) = \{(v_i, v_j), v_i, v_j \in V(G)\}$. Let $A(G)$ be the $n \times n$ adjacency matrix of G, such that $A(G)_{ij} = 1$ if and only if $(v_i, v_j) \in \mathscr{E}(G)$. When it is clear in context, we may suppress G and write the adjacency matrix as A.

A graph S is a subgraph of G, written as $S \subset G$, if $V(S) \subset V(G)$ and $\mathscr{E}(S) \subset \mathscr{E}(G)$. In particular, a subgraph $S \subset G$ is called an induced subgraph of G, denoted by $S \Subset G$, if for any $v_i, v_j \in V(S)$, $(v_i, v_j) \in \mathscr{E}(S)$ whenever $(v_i, v_j) \in \mathscr{E}(G)$. Lastly, two graphs S and G are isomorphic, denoted by $S \cong G$, if there exists a bijective function ϕ: $V(S) \to V(G)$ such that $(v_i, v_j) \in \mathscr{E}(S)$ if and only if $[\phi(v_i), \phi(v_j)] \in \mathscr{E}(G)$.

For subsampling framework, we use $\mathbb{G}_b^{(*G)}$ to represent the random induced subgraph of G from node subsampling (conditioning on $\mathbb{G}_n = G$). When discussing

distributional quantities of $\mathbb{G}_b^{(*G)}$, like the expectation and variance, we employ $(*G)$ in our notation. For example, $\text{var}(*G)$ is the variance of $\mathbb{G}_b^{(*G)}$ given $\mathbb{G}_n = G$. We write $(*G)$ as $*$ when the context clearly identifies G. When we study the asymptotic properties, we are considering a sequence of random networks $\{\mathbb{G}_n\}$, with $n \to \infty$.

C.1 Supporting propositions, lemmas, and additional theoretical results

For notational simplicity, we define

$$h_n(u,v) = \rho_n w(u,v) \mathbb{1}_{\{\rho_n w(u,v) \leq 1\}}. \tag{C.61}$$

If a network $\mathbb{G}_n \sim h_n$, we also denote it by $\mathbb{G}_n^{h_n}$ for abbreviation.

The network moment $U_R(\mathbb{G}_b^*)$ is essentially a function of $\mathbb{G}_b^*$, and $\mathbb{G}_b^*$ can be conceptualized treated as a conditional random variable. Since any network G is mathematically equivalent to its adjacent matrix $A(G)$, $\mathbb{G}_b^*$ can be represented as

$$\mathbb{G}_b^* = A(\mathbb{G}_b^*) = A \mid G,$$

where A represents a random matrix. However, discussing the marginal distribution of A without reference to G is conceptually challenging, as a graph is necessary for the discussion of its subgraphs. Therefore, we turn to focus on

$$\mathbb{G}_b^{(*\mathbb{G}_n)} = A[\mathbb{G}_b^{(*\mathbb{G}_n)}] = A \mid \mathbb{G}_n.$$

Under model (C.61), $A(\mathbb{G}_n)_{ij}$ are identical and independent distributed. On the other hand, the Algorithm 5 implies that $A[\mathbb{G}_b^{(*\mathbb{G}_n)}]$ is essentially a subset of $A(\mathbb{G}_n)$.

Thus, $A[\mathbb{G}_b^{(*\mathbb{G}_n)}]_{ij}$ are also identical and independent distributed, with $A[\mathbb{G}_b^{(*\mathbb{G}_n)}] \sim A(\mathbb{G}_b)$.

Based on this point of view, we studied the statistical properties of $U_R(\mathbb{G}_b^*)$ in Lemma 42, which is left in Section C.1. Our results resembles the law of iterated expectations and the law of total variance, and generalized the results in Bhattacharyya and Bickel (2015b).

One can also formulated $U_R(\mathbb{G}_b^*)$ as a finite population U-statistic (Zhang and Xia, 2022). In this way, the network G is conceptually treated as a finite population: $G = \{\mathfrak{v}_1, \cdots, \mathfrak{v}_n\}$, where each unit $\mathfrak{v}_i$ represent the adjacency information between the ith vertex and other vertices. The finite population U-statistic has been studied in Zhao and Chen (1990); Bloznelis and Götze (2001, 2002), which is defined as follows.

Definition 35 (Finite population U-statistic)**.** *Let $\mathscr{V} = (\mathfrak{v}_1, \cdots, \mathfrak{v}_n)$ be a finite population consisting of n units. Let $T = t(\mathbb{V}_1, \cdots, \mathbb{V}_b)$ denote a statistic based on simple random sample $\mathbb{V}_1, \cdots, \mathbb{V}_b$ drawn without replacement from $\mathscr{V}$. If the kernel function t is invariate under permutations of its arguments, then T is called a finite population U-statistic.*

C.1.1 Properties of motif counts

In this section, we introduce two useful features of motif counts. The first feature is the relationship between motif counts and graph injective homomorphisms:

Lemma 36 (Proposition 1 of Amini et al. (2012))**.** *For any motif R and graph G,*

$$X_R(G) = \mathrm{inj}(R, G)/|\mathrm{Aut}(R)|,$$

where $\mathrm{inj}(R, G)$ denotes the number of injective graph homomorphisms (Lovász

and Szegedy, 2006), and $\mathrm{Aut}(R)$ denotes the set of all automorphisms of R. To be specific, A mapping ϕ: $V(R) \to V(G)$ is a graph homomorphism if $(v_i, v_j) \in \mathscr{E}(R)$ implies $[\phi(v_i), \phi(v_j)] \in \mathscr{E}(G)$. Furthermore, ϕ is a injective graph homomorphism if $\phi(v_i) = \phi(v_j)$ implies $v_i = v_j$. On the other hand, $\mathrm{Aut}(R)$ is the set of all permutations ψ of the vertex set $V(R)$ such that $(x, y) \in \mathscr{E}(R)$ if and only if $[\psi(x), \psi(y)] \in \mathscr{E}(R)$. More explanations and details of $\mathrm{Aut}(R)$ are provided in Rodriguez (2014). We now introduce the second feature, which is termed as the linearity of motif count:

Lemma 37 (Lemma 1 in Maugis et al. (2020))**.** *For any two motifs R and R'.*

$$X_R(G)X_{R'}(G) = \sum_{S \in \mathcal{S}_{R,R'}} c_S X_S(G) = \sum_{q=0}^{\min\{r,r'\}} \sum_{S \in \mathcal{S}^{(q)}_{R,R'}} c_S X_S(G), \tag{C.62}$$

where $\mathcal{S}^{(q)}_{R,R'}$ is special collection of patterns defined as

Definition 38 (Definition 5 in Maugis et al. (2020))**.**

Let $\mathcal{S}_{R,R'}$ denote the set of all unlabeled graphs that can be formed from R and R'. $\mathcal{S}_{R,R'}$ is defined as

$$\mathcal{S}_{R,R'} = \Big\{ S \subset K_{r+r'} : V(S) = V(R_1) \cup V(R_2), \mathscr{E}(S) = \mathscr{E}(R_1) \cup \mathscr{E}(R_2), R_1 \cong R, R_2 \cong R' \Big\}, \tag{C.63}$$

where K_n denotes the complete graph of size n. Furthermore, the set $\mathcal{S}_{R,R'}$ can be partitioned into disjoint sets $\mathcal{S}^{(q)}_{R,R'}$ based on the number of merged vertices q, where $\mathcal{S}^{(q)}_{R,R'}$ is defined as

$$\mathcal{S}^{(q)}_{R,R'} = \Big\{ S : S \subset \mathcal{S}_{R,R'}, |V(S)| = r + r' - q \Big\}.$$

Lastly, for each $S \subset \mathcal{S}_{R,R'}$*, we define a constant* c_S *as*

$$c_S = \Big| \{(R_1, R_2) \subset S : V(S) = V(R_1) \cup V(R_2), \mathscr{E}(S) = \mathscr{E}(R_1) \cup \mathscr{E}(R_2), R_1 \cong R, R_2 \cong R'\} \Big|. \tag{C.64}$$

Section C.2 provides two examples to aid readers in understanding Definition C.63. The linearity property in Lemma C.62 is crucial to the proofs in Bhattacharyya and Bickel (2015b); Maugis et al. (2020).

C.1.2 Statistical properties of network moments of graphs from graphon model

Similar to Bickel et al. (2011), we define the following quantities:

$$\begin{aligned} P_{h_n}(R) &= \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} h_n(\xi_i, \xi_j) \prod_{v_i \in V(R)} d\xi_i, \\ P_w(R) &= \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} w(\xi_i, \xi_j) \prod_{v_i \in V(R)} d\xi_i. \end{aligned} \tag{C.65}$$

Lemma 39 below documents some fundamental properties of network moment $U_R(\mathbb{G}_n)$.

Lemma 39. *For any motif* R*,*

$$E\big[U_R(\mathbb{G}_n)\big] = \frac{r!}{|\mathrm{Aut}(R)|} P_{h_n}(R). \tag{C.66}$$

Moreover, consider motifs R and R',

$$\operatorname{cov}\big[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)\big] = \binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}^{(q)}_{R,R'}} c_S E\big[X_S(\mathbb{G}_n)\big] - \Big[\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big]\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{S\in\mathcal{S}^{(0)}_{R,R'}} c_S E\big[X_S(\mathbb{G}_n)\big]. \tag{C.67}$$

Following Bickel et al. (2011), we focus on the statistical properties of $\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n)$ rather than $U_R(\mathbb{G}_n)$ because both the expectation and variance of $U_R(\mathbb{G}_n)$ shrink to zero when ρ_n converges to zero. We now state the following results:

Proposition 40. *For any motif R,*

$$E\big[\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n)\big] = \frac{r!}{|\mathrm{Aut}(R)|}P_w(R). \tag{C.68}$$

Furthermore, consider motifs R and R' with sizes $r \leqslant r'$. Assume that $n\rho_n^{r/2} \to \infty$, then

$$\lim_{n\to\infty}\operatorname{cov}\big[\sqrt{n}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n), \sqrt{n}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_n)\big] = \sum_{S\in\mathcal{S}^{(1)}_{R,R'}}\frac{c_S r!r'!}{|\mathrm{Aut}(S)|}P_w(S) - \sum_{S\in\mathcal{S}^{(0)}_{R,R'}}\frac{c_S r!r'!rr'}{|\mathrm{Aut}(S)|}P_w(S). \tag{C.69}$$

The right-hand side in (C.69) describes the limit of covariance. When the limit of variance is non-zero, $\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n)$ is called non-degenerate.

C.1.3 Statistical properties of network moments of subsampled graphs: Part I

We introduce some additional notations before presenting Lemma 41. Let $\mathcal{S}(\mathbb{G}_b^*)$ denote the collection of all possible $\mathbb{G}_b^*$. For a fixed vertex $v \in V(G)$, we use $\mathbb{G}_b^{v*}$

to denote a random induced subgraph of G based on the vertex v and other $b-1$ random vertices drawn without replacement from $V(G)\setminus v$. Similarly, we use $\mathcal{S}(\mathbb{G}_b^{v*})$ to denote the sample space of $\mathbb{G}_b^{v*}$. Let $G_b^{v*} \in \mathcal{S}(\mathbb{G}_b^{v*})$ be a realization. We use $\mathbb{G}_{b,r}^{v**}$ to denote a random induced subgraph of G_b^{v*} based on the vertex v and other $r-1$ random vertices drawn without replacement from $V(\mathbb{G}_b^*)\setminus v$, and use $\mathcal{S}(\mathbb{G}_{b,r}^{v**})$ to denote the set contains all possible $\mathbb{G}_{b,r}^{v**}$. The following lemma provides useful identities.

Lemma 41.

$$\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^*)} X_R(\mathscr{G}) = \binom{n-r}{b-r} X_R(G). \tag{C.70}$$

$$\Big|\{S : S\subset G_b^{v*}, v\in V(S), S\cong R\}\Big| = \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{b,r}^{v**})} X_R(\mathscr{G}). \tag{C.71}$$

$$\Big|\{S : S\subset G, v\in V(S), S\cong R\}\Big| = \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v*})} X_R(\mathscr{G}). \tag{C.72}$$

$$\sum_{G_b^*\in\mathcal{S}(\mathbb{G}_b^{v*})} \Big(\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{b,r}^{v**})} X_R(\mathscr{G})\Big) = \binom{n-r}{b-r} \Big|\{S : S\subset G, v\in V(S), S\cong R\}\Big|. \tag{C.73}$$

$$\sum_{i=1}^{n} \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G}) = \sum_{i=1}^{n} \Big|\{S : S\subset G, v_i\in V(S), S\cong R\}\Big| = rX_R(G). \tag{C.74}$$

Based on Lemma 41, we derived the following lemma which generlize the results in Bhattacharyya and Bickel (2015b).

Lemma 42. *Given the network G, for any motif R,*

$$E_*\left[U_R(\mathbb{G}_b^*)\right] = U_R(G). \tag{C.75}$$

And for any two motifs R and R' with $r + r' < b$,

$$\begin{aligned}\mathrm{cov}_*\left[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\right] =& \binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} C_S\binom{b}{s}\binom{n}{s}^{-1}X_S(G) \\ &- U_R(G)U_{R'}(G),\end{aligned} \tag{C.76}$$

where $s = |V(S)| = r + r' - q$. Moreover, suppose that $G \sim \mathbb{G}_n$. Then

$$E[U_R(\mathbb{G}_b^{h_n})] = E\Big\{E_*\big[U_R(\mathbb{G}_b^{(*\mathbb{G}_n=G)})\big]\Big\} = E\Big\{E_*\big[U_R(\mathbb{G}_b^*)\big]\Big\}, \tag{C.77}$$

$$\mathrm{cov}[U_R(\mathbb{G}_b^{h_n}), U_{R'}(\mathbb{G}_b^{h_n})] = \mathrm{cov}\Big\{E_*\big[U_R(\mathbb{G}_b^*)\big], E_*\big[U_{R'}(\mathbb{G}_b^*)\big]\Big\} + E\Big\{\mathrm{cov}_*\big[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\big]\Big\}. \tag{C.78}$$

C.1.4 Statistical properties of network moments of subsampled graphs: Part II

The following proposition extends the results about finite population statistic in Bloznelis and Götze (2001, 2002) into network subsampling context.

Proposition 43.

(a) The Hoeffding's decomposition of $U_R(\mathbb{G}_b^*)$ is

$$U_R(\mathbb{G}_b^*) = E_*[U_R(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right) + \sum_{1\leqslant i<j\leqslant b} g_{2,R}\left(\mathbb{V}_i, \mathbb{V}_j\right) + \cdots, \tag{C.79}$$

where

$$\begin{aligned} g_{1,R}(\mathbb{V}_1) =& \frac{r!(n-r-1)!}{b(n-2)!} \sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}) - \frac{r(n-1)}{b(n-r)} U_R(G) \\ =& \frac{(n-1)}{b}[U_R(G) - U_R(G \setminus \mathbb{V}_1)], \end{aligned} \tag{C.80}$$

with

$$\begin{aligned} & \mathrm{cov}_{\mathbb{V}_1*}[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)] \\ =& \frac{r!(n-r-1)!}{b(n-2)!} \frac{r'!(n-r'-1)!}{b(n-2)!} \sum_{k=0}^{\min\{r,r'\}} \sum_{S\in\mathscr{S}_{R,R'}^{(q)}} c_S \frac{nq-rr'}{n^2} X_S(G). \end{aligned} \tag{C.81}$$

Furthermore, we have

$$\mathrm{cov}_*[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i), \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)] = \frac{b(n-b)}{(n-1)} \mathrm{cov}_*[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)], \tag{C.82}$$

and as $n, b \to \infty$

$$\lim_{b,n\to\infty} \mathrm{var}_*[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_1)] = 0. \tag{C.83}$$

(b) For two motifs R *and* R', $U_R(\mathbb{G}_b^*) + U_{R'}(\mathbb{G}_b^*)$ *is also a symmetric finite population statistic with the following Hoeffding's decomposition*

$$\begin{aligned} U_R(\mathbb{G}_b^*) + U_{R'}(\mathbb{G}_b^*) =& E_*[U_R(\mathbb{G}_b^*) + U_{R'}(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b} g_{1,R,R'}(\mathbb{V}_i) \\ & + \sum_{1\leqslant i<j\leqslant b} g_{2,R,R'}(\mathbb{V}_i, \mathbb{V}_j) + \cdots, \end{aligned}$$

where

$$g_{1,R,R'}(\mathbb{V}_1) = g_{1,R}(\mathbb{V}_1) + g_{1,R'}(\mathbb{V}_1). \tag{C.84}$$

Moreover, the variance of linear parts satisfies:

$$\mathrm{var}_* \sum_{1\leqslant i\leqslant b} g_{1,R,R'}\left(\mathbb{V}_i\right) = \frac{b(n-b)}{(n-1)}\mathrm{var}_*\left[g_{1,R,R'}\left(\mathbb{V}_1\right)\right], \tag{C.85}$$

$$\lim_{b,n\to\infty} \mathrm{var}_*[\sum_{1\leqslant i\leqslant b} g_{1,R,R'}\left(\mathbb{V}_i\right)] = 0. \tag{C.86}$$

C.1.5 Asymptotic distribution of network moments of subsampled graphs

Assumption 3 (Non-degenerate moment)**.** *For motif R and graphon h_n,*

$$\sigma_R^2 = \lim_{n\to\infty} \mathrm{var}[\sqrt{n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_n)] > 0.$$

Based on the result in Bloznelis and Götze (2001), we derived the following asymptotic results for the subsampling distribution.

Theorem 44. *Suppose that $\{G^{(n)}\}$ is a sequence of networks, where each $G^{(n)} \sim \mathbb{G}_n$.*

(a) The Hoeffding's decomposition of $\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*)$ is

$$\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*) = \sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R[G^{(n)}] + \sum_{1\leqslant i\leqslant b_n} \sqrt{b_n}\rho_n^{-\mathfrak{r}} g_{1,R}\left(\mathbb{V}_i\right) + \Delta[\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*)]. \tag{C.87}$$

For any network sequence, the following conditions hold with probability one.

(i) Under Assumptions 3, 2 and 1, we have

$$\lim_{n\to\infty} E_*\Delta^2[\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*)] = 0, \tag{C.88}$$

$$0 < c_3 \leqslant \mathrm{var}_*\big[\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}^*_{b_n})\big] \leqslant c_4 < \infty \text{ for some } c_3, c_4 > 0. \quad \text{(C.89)}$$

(ii) Under Assumptions 2, for every $\epsilon > 0$,

$$\lim_{n\to\infty} b_n E_*\big[b_n\rho_n^{-2\mathfrak{r}} g^2_{1,R}(\mathbb{V}_1)\mathbb{1}_{\{b_n\rho_n^{-2\mathfrak{r}} g^2_{1,R}(\mathbb{V}_1)>\epsilon\}}\big] = 0, \quad \text{(C.90)}$$

Consequently, if Assumptions 3, 2, 1 are satisfied, then with probability one:

$$\sqrt{b}\big[\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}^*_b) - \rho_n^{-\mathfrak{r}} U_R(G)\big] \to \mathcal{N}(0, \sigma^2_{*R}) \text{ in distribution}, \quad \text{(C.91)}$$

(b) Let $\{R_1, \cdots, R_m\}$ *be* m *motifs with* $\max\{r_1, \cdots, r_m\} \leq r$ *and* $\max\{\mathfrak{r}_1, \cdots, \mathfrak{r}_m\} \leq \mathfrak{r}$. *Suppose that Assumption 3 holds for every* R_i *with* $i \in [m]$, *and Assumption 2, 1 are satisfied. Then with probability one:*

$$\sqrt{b}\Big\{\big[\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}^*_b), \cdots, \rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}^*_b)\big] - \big[\rho_n^{-\mathfrak{r}_1} U_{R_1}(G), \cdots, \rho_n^{-\mathfrak{r}_m} U_{R_m}(G)\big]\Big\}$$
$$\to \mathcal{N}\big[0, \Sigma(*\boldsymbol{R})\big] \text{ in distribution}, \quad \text{(C.92)}$$

Lemma 45. *Let* R *and* R' *be two motifs with* $\max\{r, r'\} \leq r_1$ *and* $\max\{\mathfrak{r}, \mathfrak{r}'\} \leq \mathfrak{r}_1$. *Suppose that Assumption 2 holds after replacing* r *by* r_1 *and* $\mathfrak{r}$ *by* $\mathfrak{r}_1$, *and Assumption 1 holds. Then*

$$pr\Big\{\lim_{b\to\infty} \rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}_*\big[\sqrt{b}U_R(\mathbb{G}^*_b), \sqrt{b}U_{R'}(\mathbb{G}^*_b)\big]$$
$$= (1-c_2)\lim_{b\to\infty} \rho_b^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}\big[\sqrt{b}U_R(\mathbb{G}_b), \sqrt{b}U_{R'}(\mathbb{G}_b)\big]\Big\} = 1.$$

C.1.6 Consistency of empirical distribution

Similar to Lunde and Sarkar (2023), we also considered the following empirical

CDF:

$$\hat{J}_{*,n,b}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big\{\sqrt{b}\big[\widehat{\rho_G}^{\mathfrak{r}_1} U_{R_1}(G_b^{*(i)}) - \widehat{\rho_G}^{\mathfrak{r}_1} U_{R_1}(G)\big] \le t_1, \\ \cdots, \sqrt{b}\big[\widehat{\rho_G}^{\mathfrak{r}_m} U_{R_m}(G_b^{*(i)}) - \widehat{\rho_G}^{\mathfrak{r}_m} U_{R_m}(G)\big] \le t_m\Big\}. \tag{C.93}$$

The following consistency result is developed based on Theorem 1 of Lunde and Sarkar (2023):

Theorem 46. *For* $\{R_1,\cdots,R_m\}$ *with* $\max\{r_1,\cdots,r_m\} \le r$ *and* $\max\{\mathfrak{r}_1,\cdots,\mathfrak{r}_m\} \le \mathfrak{r}$. *Under Assumptions 3, 2, 1, with probability one:*

$$\sup_{[t_m]\in\mathbb{R}^m} \left|\hat{J}_{*,n,b}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m) - J_{*,n,b}^{\{R_1,\cdots,R_m\}}(t_1,\cdots,t_m)\right| \to 0.$$

C.2 Two examples for Definition C.63

Following Maugis et al. (2020), we use two examples to explain this definition. In the first example, let R be a  and R' also be a . Then the set $\mathcal{S}_{R,R'}$ can be constructed as $\{$,,,$\}$. Each element in $\mathcal{S}_{R,R'}$ can be obtained by building blocks based on R and R'. Let R_1 be a copy of R, and R_2 be a copy to R'. The pattern can be built by either put R_1 in the left side or in the right side. Thus, $c_{\text{}} = 2$. Similarly, $c_{\text{}} = 2$, $c_{\text{}} = 2$ and $c_{\text{}} = 1$. Generally speaking, c_S denotes the number of ways S can be built from copies of R and R'. Based on the number of merged vertices, we have $S_{R,R'}^{(0)} = \{$$\}$,$S_{R,R'}^{(1)} = \{$$\}$, $S_{R,R'}^{(2)} = \{$$\}$, and $S_{R,R'}^{(3)} = \{$$\}$. For the second example, let R be a and R' be a . Then $S_{R,R'} = \{S_{R,R'}^{(0)}, S_{R,R'}^{(1)}, S_{R,R'}^{(2)}, S_{R,R'}^{(3)}, S_{R,R'}^{(4)}\}$, with $S_{R,R'}^{(0)} = \{$$\}$, $S_{R,R'}^{(1)} = \{$$\}$, $S_{R,R'}^{(2)} = \{$, , $\}$, $S_{R,R'}^{(3)} = \{$, , $\}$, $S_{R,R'}^{(4)} = \{$$\}$. Correspondingly, $c_{\text{}} = 2$, $c_{\text{}} = 2$, $c_{\text{}} = 6$, $c_{\text{}} = 2$, $c_{\text{}} = 2$, $c_{\text{}} = 2$, $c_{\text{}} = 6$, $c_{\text{}} = 6$ and $c_{\text{}} = 1$.

As noted in Maugis et al. (2020), $X_R(G)X_{R'}(G)$ involves counting pairs of motifs, and could be recovered by counting the number of the all motifs that are formed by using one copy of R and one copy of R' as building blocks. This is the intuition of Lemma 37. Moreover, the equation in (C.62) provides flexibility as it does not depend on the generation mechanism of G.

C.3 Proofs for Section C.1

C.3.1 Proof of Lemma 39

Proof. Lemma 39 mostly follow the results in Bhattacharyya and Bickel (2015a); Maugis et al. (2020); Bhattacharya et al. (2022). We provide the proof here for reader's convenience. We begin by present some useful lemmas.

Lemma 47 (Useful identities)**.**

$$E\big[U_R(\mathbb{G}_n)\big] = \binom{n}{r}^{-1} E\big[X_R(\mathbb{G}_n)\big] = \binom{n}{r}^{-1} X_R(K_n)P_{h_n}(R), \tag{C.94}$$

where K_n denotes a complete graph of size n, and $P_{h_n}(R)$ is defined in (C.65) .

$$X_R(K_n) = \binom{n}{r} X_R(K_r), \tag{C.95}$$

$$X_R(K_r) = r!/|\mathrm{Aut}(R)|. \tag{C.96}$$

The (C.94) is proved in Maugis et al. (2020) (See their equation (1)), (C.95) is used in Bollobás and Riordan (2007), and (C.96) is proved in Bhattacharya et al. (2022) (see their equation (2.7)).

We now prove the results one by one.

i) First, by (C.94), $E[U_R(\mathbb{G}_n)] = \binom{n}{r}^{-1} X_R(K_n) P_{h_n}(R)$, so that

$$X_R(K_n) \overset{(C.95)}{=} \binom{n}{r} X_R(K_r) \overset{(C.96)}{=} \binom{n}{r} \frac{r!}{|\mathrm{Aut}(R)|} \tag{C.97}$$

and

$$E[U_R(\mathbb{G}_n)] = \binom{n}{r}^{-1} X_R(K_n) P_{h_n}(R) = \frac{r!}{|\mathrm{Aut}(R)|} P_{h_n}(R),$$

which gives (C.66).

ii) To show (C.67), we start with

$$\begin{aligned}
&\mathrm{cov}\big[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)\big] = E\big[U_R(\mathbb{G}_n) U_{R'}(\mathbb{G}_n)\big] - E\big[U_R(\mathbb{G}_n)\big] E\big[U_{R'}(\mathbb{G}_n)\big] \\
&\overset{(C.94)}{=} \binom{n}{r}^{-1} \binom{n}{r'}^{-1} E\big[X_R(\mathbb{G}_n) X_{R'}(\mathbb{G}_n)\big] - \Big\{ E\big[U_R(\mathbb{G}_n)\big] E\big[U_{R'}(\mathbb{G}_n)\big] \Big\} \\
&\overset{(C.62)}{=} \binom{n}{r}^{-1} \binom{n}{r'}^{-1} E\Big[\sum_{q=0}^{\min\{r,r'\}} \sum_{S \in \mathcal{S}^{(q)}_{R,R'}} c_S X_S(\mathbb{G}_n) \Big] - E\big[U_R(\mathbb{G}_n)\big] E\big[U_{R'}(\mathbb{G}_n)\big].
\end{aligned} \tag{C.98}$$

The result in Bhattacharyya and Bickel (2015b) (see $\sigma(R_1, R_2; \rho)$ in Part $\mathbf{B_1}$) implies that

$$\begin{aligned}
\mathrm{cov}\big[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)\big] =& \binom{n}{r}^{-1} \binom{n}{r'}^{-1} \sum_{q=1}^{\min\{r,r'\}} \sum_{S \in \mathcal{S}^{(q)}_{R,R'}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&- \Big(1 - \frac{\binom{n-r}{r'}}{\binom{n}{r'}}\Big) E\big[U_R(\mathbb{G}_n)\big] E\big[U_{R'}(\mathbb{G}_n)\big].
\end{aligned}$$

Combining with (C.98), we obtain

$$E\big[U_R(\mathbb{G}_n)\big] E\big[U_{R'}(\mathbb{G}_n)\big] = \frac{\binom{n}{r'}}{\binom{n-r}{r'}} \binom{n}{r}^{-1} \binom{n}{r'}^{-1} \sum_{S \in \mathcal{S}^{(0)}_{R,R'}} c_S E\big[X_S(\mathbb{G}_n)\big]. \tag{C.99}$$

Finally, we arrive at

$$\begin{aligned}
\operatorname{cov}\big[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)\big] &\overset{(C.98)}{=} \binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&\quad - E\big[U_R(\mathbb{G}_n)\big]E\big[U_{R'}(\mathbb{G}_n)\big] \\
&\overset{(C.99)}{=} \binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&\quad + \binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{S\in\mathcal{S}_{R,R'}^{(0)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&\quad - \frac{\binom{n}{r'}}{\binom{n-r}{r'}}\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{S\in\mathcal{S}_{R,R'}^{(0)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&= \binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&\quad - \Big(\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{S\in\mathcal{S}_{R,R'}^{(0)}} c_S E\big[X_S(\mathbb{G}_n)\big],
\end{aligned}$$

which gives (C.67).

□

C.3.2 Proof of Proposition 40

Proof. we prove the results one by one.

i) The equation in (C.68) is derived as follows.

$$
\begin{aligned}
E\big[\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_n)\big] = \rho_n^{-\mathfrak{r}} E\big[U_R(\mathbb{G}_n)\big] &\overset{(C.66)}{=} \frac{\rho_n^{-\mathfrak{r}} r!}{|\mathrm{Aut}(R)|} P_{h_n}(R) \\
&\overset{(C.65)}{=} \frac{\rho_n^{-\mathfrak{r}} r!}{|\mathrm{Aut}(R)|} \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} h_n\left(\xi_i,\xi_j\right) \prod_{v_i\in V(R)} d\xi_i \\
&= \frac{r!\rho_n^{-\mathfrak{r}}\rho_n^{\mathfrak{r}}}{|\mathrm{Aut}(R)|} \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} w\left(\xi_i,\xi_j\right) \mathbb{1}_{\{\rho_n w(\xi_i,\xi_j)\leqslant 1\}} \prod_{v_i\in V(R)} d\xi_i \\
&= \frac{r!}{|\mathrm{Aut}(R)|} \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} w\left(\xi_i,\xi_j\right) \prod_{v_i\in V(R)} d\xi_i \overset{(C.65)}{=} \frac{r!}{|\mathrm{Aut}(R)|} P_w(R).
\end{aligned}
$$

ii) To show (C.69), we partition the covariance into

$$
\begin{aligned}
&\mathrm{cov}\Big[\sqrt{n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_n), \sqrt{n}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_n)\Big] = n\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}\Big[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)\Big] \\
&\overset{(C.67)}{=} n\rho_n^{-(\mathfrak{r}+\mathfrak{r}')} \binom{n}{r}^{-1}\binom{n}{r'}^{-1} \sum_{q=1}^{\min\{r,r'\}} \sum_{S\in\mathscr{S}_{R,R'}^{(q)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&\quad - n\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\Big[\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big]\binom{n}{r}^{-1}\binom{n}{r'}^{-1} \sum_{S\in\mathscr{S}_{R,R'}^{(0)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&:= \mathrm{I} - \mathrm{II}.
\end{aligned}
$$

Let $|V(S)| = s$ and $|E(S)| = \mathfrak{s}$, then part I could be expressed as

$$\begin{aligned}
\mathrm{I} &= n\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E\big[X_S(\mathbb{G}_n)\big] \\
&= \sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\frac{nr!(n-r)!}{n!}\frac{r'!(n-r')!}{n!}\binom{n}{s}E\big[U_S(\mathbb{G}_n)\big] \\
&\overset{(C.68)}{=} \sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\frac{nr!(n-r)!}{n!}\frac{r'!(n-r')!}{n!}\frac{n!}{s!(n-s)!}\frac{s!}{|\mathrm{Aut}(S)|}P_w(S) \\
&= \sum_{q=1}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} \rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!}\frac{c_S r!r'!}{|\mathrm{Aut}(S)|}P_w(S).
\end{aligned}$$

The quantities r, r', c_S, $|\mathrm{Aut}(S)|$, and $P_w(S)$ are invariant of n. The quantities $\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}$ and $(n-r')!/[(n-1)\cdots(n-r+1)(n-s)!]$ change with n. Now, we consider these two quantities based on the number of merged vertices q.

- When $q = 1$, we have $\mathfrak{s} = \mathfrak{r}+\mathfrak{r}'$ and $s = r+r'-1$. The following quantity

$$\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = \frac{(n-r')(n-r'-1)\cdots(n-r-r'+2)}{(n-1)\cdots(n-r+1)}$$

  has $r-1$ items including n in both numerator and denominator. Thus,

$$\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = 1+o(1).$$

- When $q = 2$, we have $\mathfrak{s} = \mathfrak{r}+\mathfrak{r}'-1$ because one edge is merged. The following quantity

$$\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = \frac{(n-r')(n-r'-1)\cdots(n-r-r'+3)}{(n-1)\cdots(n-r+1)}$$

has $r-2$ items with n in numerator, and $r-1$ items with n in denominator. As $n\rho_n \to \infty$,

$$\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = O(\frac{1}{n\rho_n}) = o(1).$$

- When $2 < q < \min\{r, r'\}$, at most $q(q-1)/2$ edges are merged. The following quantity

$$\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = \frac{(n-r')(n-r'-1)\cdots(n-r-r'+(q+1))}{(n-1)\cdots(n-r+1)}$$

has $r-q$ items with n in numerator and $r-1$ items with n in denominator. Since $n^{(q-1)}\rho_n^{(q(q-1)/2)} = (n\rho_n^{q/2})^{(q-1)} \to \infty$

$$\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\frac{(n-r')!}{(n-1)\cdots(n-r+1)(n-s)!} = O(\frac{1}{(n\rho_n^{q/2})^{(q-1)}}) = o(1).$$

Therefore,

$$\lim_{n\to\infty} \mathrm{I} = \sum_{S\in\mathcal{S}^{(1)}_{R,R'}} \frac{c_S r!r'!}{|\mathrm{Aut}(S)|} P_w(S).$$

Now we turn to focus on part II. Since $s = r + r'$ and $\mathfrak{s} = \mathfrak{r} + \mathfrak{r}'$ when $q = 0$, we have

$$\begin{aligned}
&\Big[\frac{n\binom{n}{r'}}{\binom{n-r}{r'}} - n\Big]\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\binom{n}{s} \\
=&\frac{r!r'!}{(r+r')!}\Big\{b\Big[1 - \frac{(n-r)!(n-r')!}{n!(n-r-r')!}\Big]\Big\} \\
=&\frac{r!r'!}{(r+r')!}\Big[\frac{n(n-1)\cdots(n-r+1) - (n-r')\cdots(n-r-r'+1)}{(n-1)\cdots(n-r+1)}\Big].
\end{aligned}$$

Consequently,

$$\begin{aligned} n(n-1)\cdots(n-r+1) &= n^r - \frac{(r-1)r}{2}n^{r-1} + o(n^{r-1}), \\ (n-r')\cdots(n-r-r'+1) &= n^r - \frac{(2r'+r-1)r}{2}n^{r-1} + o(n^{r-1}), \\ (n-1)\cdots(n-r+1) &= n^{r-1} + o(n^{r-1}). \end{aligned}$$

By L'Hospital's rule,

$$\begin{aligned} &\lim_{n\to\infty} \Big(\frac{n\binom{n}{r'}}{\binom{n-r}{r'}} - n\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\binom{n}{s} \\ &= \lim_{n\to\infty} \frac{r!r'!}{(r+r')!}\Big[\frac{n(n-1)\cdots(n-r+1) - (n-r')\cdots(n-r-r'+1)}{(n-1)\cdots(n-r+1)}\Big] \\ &= \lim_{n\to\infty} \frac{r!r'!}{(r+r')!}\frac{rr'n^{r-1} + o(n^{r-1})}{n^{r-1} + o(n^{r-1})} = \frac{r!r'!}{(r+r')!}rr'. \end{aligned} \tag{C.100}$$

Consequently,

$$\begin{aligned} \lim_{n\to\infty} \mathrm{II} &= \lim_{n\to\infty} n\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\Big(\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{S\in\mathcal{S}^{(0)}_{R,R'}} c_S E\big[X_S(\mathbb{G}_n)\big] \\ &= \lim_{n\to\infty} n\rho_n^{\mathfrak{s}-(\mathfrak{r}+\mathfrak{r}')}\Big(\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\binom{n}{s}\sum_{S\in\mathcal{S}^{(0)}_{R,R'}} c_S E\big[\rho_n^{-\mathfrak{s}}U_S(\mathbb{G}_n)\big] \\ &= \lim_{n\to\infty} n\Big(\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\binom{n}{s}\sum_{S\in\mathcal{S}^{(0)}_{R,R'}} c_S E\big[\rho_n^{-\mathfrak{s}}U_S(\mathbb{G}_n)\big] \\ &\overset{(C.68)}{=} \lim_{n\to\infty} n\Big(\frac{\binom{n}{r'}}{\binom{n-r}{r'}} - 1\Big)\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\binom{n}{s}\sum_{S\in\mathcal{S}^{(0)}_{R,R'}} c_S\frac{s!}{|\mathrm{Aut}(S)|}P_w(S) \\ &\overset{(C.100)}{=} \sum_{S\in\mathcal{S}^{(0)}_{R,R'}} \frac{c_S r!r'!rr'}{|\mathrm{Aut}(S)|}P_w(S). \end{aligned}$$

Finally, we arrive at

$$\lim_{n\to\infty} \operatorname{cov}\left[\sqrt{n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_n), \sqrt{n}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_n)\right] = \lim_{n\to\infty} \mathrm{I} - \lim_{n\to\infty} \mathrm{II}$$
$$= \sum_{S\in\mathcal{S}^{(1)}_{R,R'}} \frac{c_S r! r'!}{|\mathrm{Aut}(S)|} P_w(S) - \sum_{S\in\mathcal{S}^{(0)}_{R,R'}} \frac{c_S r! r'! r r'}{|\mathrm{Aut}(S)|} P_w(S),$$

which gives (C.69).

□

C.4 Proofs for Section C.1

C.4.1 Proof of Lemma 41

Recall that $\mathbb{G}_b^*$ denotes a random subsampled graph, and $\mathcal{S}(\mathbb{G}_b^*)$ denotes the collection of all possible $\mathbb{G}_b^*$. For a fixed vertex $v \in V(G)$, $\mathbb{G}_b^{v*}$ denotes a random subsampled graph including v, and $\mathcal{S}(\mathbb{G}_b^{v*})$ denotes the sample space of $\mathbb{G}_b^{v*}$. Let $G_b^{v*} \in \mathcal{S}(\mathbb{G}_b^{v*})$ be a realization. We use $\mathbb{G}_{b,r}^{v**}$ to denote a random induced subgraph of G_b^{v*} based on the vertex v and other $r-1$ random vertices drawn without replacement from $V(\mathbb{G}_b^*) \setminus v$, and use $\mathcal{S}(\mathbb{G}_{b,r}^{v**})$ to denote the set contains all possible $\mathbb{G}_{b,r}^{v**}$.

Proof. The following equation from Maugis et al. (2020) is used in this proof.

$$\binom{n}{r} U_R(G) = X_R(G) = \sum_{R_c \in \{S: S \subset G, S \cong R\}} 1 = \big|\{S : S \subset G, S \cong R\}\big|. \tag{C.101}$$

We now prove these identities one by one.

i) For any $R_c \in \{S : S \subset G, S \cong R\}$, recall that $R_c \subset G_b^*$ if both $V(R_c) \subset V(G_b^*)$ and $\mathscr{E}(R_c) \subset \mathscr{E}(G_b^*)$. Furthermore, since $R_c \subset G$ and $G_b^* \Subset G$, $V(R_c) \subset V(G_b^*)$ implies $\mathscr{E}(R_c) \subset \mathscr{E}(G_b^*)$. Thus,

$$\mathbb{1}_{\{R_c \subset G_b^*\}} = 1 \quad \text{if and only if} \quad V(R_c) \subset V(G_b^*). \tag{C.102}$$

Now let us consider draw b vertices from $V(G)$ by first selecting all vertices in $V(R_c)$, and then randomly draw $b - r$ vertices without replacement from $V(G) \backslash V(R_c)$. There are $\binom{n-r}{b-r}$ ways to draw these b vertices. Thus,

$$\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^*)} \mathbb{1}_{\{R_c \subset \mathscr{G}\}} = \binom{n-r}{b-r}. \tag{C.103}$$

Consequently,

$$\begin{aligned}
\sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_b^*)} X_R(\mathscr{G}) &\overset{(C.101)}{=} \sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_b^*)} \sum_{R_c\in\{S:S\subset\mathscr{G},S\cong R\}} 1 \\
&= \sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_b^*)} \sum_{R_c\in\{S:S\subset G,S\cong R\}} \mathbb{1}_{\{R_c\subset\mathscr{G}\}} \\
&= \sum_{R_c\in\{S:S\subset G,S\cong R\}} \sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_b^*)} \mathbb{1}_{\{R_c\subset\mathscr{G}\}} \\
&\overset{(C.103)}{=} \sum_{R_c\in\{S:S\subset G,S\cong R\}} \binom{n-r}{b-r} \\
&\overset{(C.101)}{=} \binom{n-r}{b-r} X_R(G),
\end{aligned}$$

which gives (C.70).

ii) For any $G_{b,r}^{v**} \in \mathscr{S}(\mathbb{G}_{b,r}^{v**})$, if $S \subset G_{b,r}^{v**}$ and $|V(S)| = |V(G_{b,r}^{v**})|$, then $V(S) = V(G_{b,r}^{v**})$. In addition, as $v \in V(G_{b,r}^{v**})$, we have

$$\{S : S \subset G_{b,r}^{v**}, S \cong R\} = \{S : S \subset G_{b,r}^{v**}, v \in V(S), S \cong R\}. \tag{C.104}$$

Let $R_c \in \{S : S \subset G, S \cong R\}$. Suppose that $R_c \subset G_b^{v*}$ and $v \in V(R_c)$. Then because every $G_{b,r}^{v**}$ is a induced subgraph, we have

$$\sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_{b,r}^{v**})} \mathbb{1}_{\{R_c\subset\mathscr{G}\}} = \sum_{\mathscr{G}\in\mathscr{S}(\mathbb{G}_{b,r}^{v**})} \mathbb{1}_{\{V(R_c)=V(\mathscr{G})\}} = 1. \tag{C.105}$$

Consequently,

$$
\begin{aligned}
&\Big|\{S : S \subset G_b^{v*}, v \in V(S), S \cong R\}\Big| \\
&= \sum_{R_c \in \{S : S \subset G_b^{v*}, v \in V(S), S \cong R\}} 1 \overset{(C.105)}{=} \sum_{R_c \in \{S : S \subset G_b^{v*}, v \in V(S), S \cong R\}} \Big(\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} \nVdash_{\{R_c \subset \mathscr{G}\}} \Big) \\
&= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} \Big(\sum_{R_c \in \{S : S \subset G_b^{v*}, v \in V(S), S \cong R\}} \nVdash_{\{R_c \subset \mathscr{G}\}} \Big) \\
&= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} \Big(\sum_{R_c \in \{S : S \subset \mathscr{G}, v \in V(S), S \cong R\}} 1 \Big) \\
&\overset{(C.101)}{=} \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} \Big|\{S : S \subset \mathscr{G}, v \in V(S), S \cong R\}\Big| \\
&\overset{(C.104)}{=} \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} \Big|\{S : S \subset \mathscr{G}, S \cong R\}\Big| \overset{(C.101)}{=} \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v**})} X_R(\mathscr{G}),
\end{aligned}
$$

which gives (C.71).

iii) If $|V(S)| = r$ and $S \subset G_r^{v*}$, then $V(S) = V(G_r^{v*})$. Thus,

$$\{S : S \subset G_r^{v*}, S \cong R\} = \{S : S \subset G_r^{v*}, v \in V(S), S \cong R\}.$$

Let $R_c \in \{S : S \subset G, S \cong R\}$. Because $G_r^{v*} \Subset G$, there exist only one $G_r^{v*} \in \mathcal{S}(\mathbb{G}_r^{v*})$ such that $V(R_c) = V(G_r^{v*})$. Also, $V(R_c) = V(G_r^{v*})$ implies $E(R_c) \subset E(G_r^{v*})$. Thus,

$$\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \nVdash_{\{R_c \subset \mathscr{G}\}} = \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \nVdash_{\{V(R_c) \subset V(\mathscr{G})\}} = 1.$$

Consequently,

$$\begin{aligned}\Big|\{S : S \subset G, v \in V(S), S \cong R\}\Big| &= \sum_{R_c \in \{S : S \subset G, v \in V(S), S \cong R\}} 1 \\ &= \sum_{R_c \in \{S : S \subset G, v \in V(S), S \cong R\}} \Big(\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \nVdash_{\{R_c \subset \mathscr{G}\}} \Big) \\ &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \Big(\sum_{R_c \in \{S : S \subset G, v \in V(S), S \cong R\}} \nVdash_{\{R_c \subset \mathscr{G}\}} \Big) \\ &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \Big|\{S : S \subset \mathscr{G}, v \in V(S), S \cong R\}\Big| \\ &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} \Big|\{S : S \subset \mathscr{G}, S \cong R\}\Big| = \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{v*})} X_R(\mathscr{G}),\end{aligned}$$

which gives (C.72)

iv) Let $R_c \in \{S : S \subset G, v \in V(S), S \cong R\}$, we have

$$\begin{aligned}\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^*)} \nVdash_{\{R_c \subset \mathscr{G}\}} &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^*), v \in V(\mathscr{G})} \nVdash_{\{R_c \subset \mathscr{G}\}} + \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^*), v \notin V(\mathscr{G})} \nVdash_{\{R_c \subset \mathscr{G}\}} \\ &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^*), v \in V(\mathscr{G})} \nVdash_{\{R_c \subset \mathscr{G}\}} + 0 \\ &= \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_b^{v*})} \nVdash_{\{R_c \subset \mathscr{G}\}}.\end{aligned} \tag{C.106}$$

Consequently,

$$\sum_{G_b^*\in\mathcal{S}(\mathbb{G}_b^{v*})}\Big(\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{b,r}^{v**})}X_R(\mathscr{G})\Big)\overset{(C.71)}{=}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^{v*})}\Big|\{S:S\subset\mathscr{G},v\in V(S),S\cong R\}\Big|$$
$$=\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^{v*})}\sum_{R_c\in\{S:S\subset G,v\in V(S),S\cong R\}}\mathbb{1}_{\{R_c\subset\mathscr{G}\}}$$
$$=\sum_{R_c\in\{S:S\subset G,v\in V(S),S\cong R\}}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^{v*})}\mathbb{1}_{\{R_c\subset\mathscr{G}\}}$$
$$\overset{(C.106)}{=}\sum_{R_c\in\{S:S\subset G,v\in V(S),S\cong R\}}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^{*})}\mathbb{1}_{\{R_c\subset\mathscr{G}\}}$$
$$\overset{(C.103)}{=}\sum_{R_c\in\{S:S\subset G,v\in V(S),S\cong R\}}\binom{n-r}{b-r}$$
$$=\binom{n-r}{b-r}\Big|\{S:S\subset G,v\in V(S),S\cong R\}\Big|, \tag{C.107}$$

which gives (C.73).

v) For the last identity, we have

$$\sum_{i=1}^{n}\Big|\{S:S\subset G,v_i\in V(S),S\cong R\}\Big|\overset{(C.101)}{=}\sum_{i=1}^{n}\sum_{R_c\in\{S:S\subset G,v_i\in V(S),S\cong R\}}1$$
$$=\sum_{i=1}^{n}\sum_{R_c\in\{S:S\subset G,S\cong R\}}\mathbb{1}_{\{v_i\in V(R_c)\}}$$
$$=\sum_{R_c\in\{S:S\subset G,S\cong R\}}\sum_{i=1}^{n}\mathbb{1}_{\{v_i\in V(R_c)\}}$$
$$=\sum_{R_c\in\{S:S\subset G,S\cong R\}}r$$
$$=r\Big|\{S:S\subset G,S\cong R\}\Big|=rX_R(G),$$

which gives (C.74).

□

C.4.2 Proof of Lemma 42

Proof. We will prove the results one by one.

i) Following the results in Maugis et al. (2020), we have

$$
\begin{aligned}
&E_*\Big[U_R(\mathbb{G}_b^*)\Big] \overset{(C.101)}{=} E_*\Big[\binom{b}{r}^{-1} X_R(\mathbb{G}_b^*)\Big] = \binom{b}{r}^{-1} E_*\Big[X_R(\mathbb{G}_b^*)\Big] \\
=&\binom{b}{r}^{-1}\binom{n}{b}^{-1} \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_b^*)} X_R(\mathscr{G}) \overset{(C.70)}{=} \binom{b}{r}^{-1}\binom{n}{b}^{-1}\binom{n-r}{b-r} X_R(G) \\
=&\binom{n}{r}^{-1} X_R(G) \overset{(C.101)}{=} U_R(G),
\end{aligned}
$$

which gives (C.75)

ii) We start by showing that

$$
\begin{aligned}
E\Big[U_R(\mathbb{G}_n)\Big] &\overset{(C.66)}{=} \frac{r!}{|\mathrm{Aut}(R)|} P_{h_n}(R) \overset{(C.97)}{=} \binom{b}{r}^{-1} X_R(K_b) P_{h_n}(R) \\
&\overset{(C.94)}{=} E[U_R(\mathbb{G}_b^{h_n})].
\end{aligned}
\tag{C.108}
$$

Hence, $E\{E_*[U_R(\mathbb{G}_b^*)]\} = E\{E_*[U_R(\mathbb{G}_b^{(*\mathbb{G}_n=G)})]\} = E[U_R(\mathbb{G}_n)] = E[U_R(\mathbb{G}_b^{h_n})]$, where the second and third equality's follow (C.75) and (C.108), respectively. This gives (C.77).

iii) Following Bhattacharyya and Bickel (2015b); Maugis et al. (2020), we have

$$
\begin{aligned}
&\mathrm{cov}_*\Big[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\Big] \overset{(C.101)}{=} \mathrm{cov}_*\Big[\binom{b}{r}^{-1} X_R(\mathbb{G}_b^*), \binom{b}{r'}^{-1} X_{R'}(\mathbb{G}_b^*)\Big] \\
=&\Big\{\binom{b}{r}^{-1}\binom{b}{r'}^{-1} E_*\Big[X_R(\mathbb{G}_b^*) X_{R'}(\mathbb{G}_b^*)\Big]\Big\} - E_*\Big[U_R(\mathbb{G}_b^*)\Big] E_*\Big[U_{R'}(\mathbb{G}_b^*)\Big].
\end{aligned}
$$

For the first part, we have

$$E_*\Big[X_R(\mathbb{G}_b^*)X_{R'}(\mathbb{G}_b^*)\Big] \overset{(C.62)}{=} E_*\Big[\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S X_S(\mathbb{G}_b^*)\Big]$$

$$= \sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E_*[X_S(\mathbb{G}_b^*)] = \sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}E_*\Big[U_S(\mathbb{G}_b^*)\Big]$$

$$\overset{(C.75)}{=} \sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}U_S(G).$$

For the second part, we have

$$E_*\big[U_R(\mathbb{G}_b^*)\big]E_*\big[U_{R'}(\mathbb{G}_b^*)\big] \overset{(C.75)}{=} U_R(G)U_{R'}(G).$$

Thus,

$$\operatorname{cov}_*\Big[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\Big] = \binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}U_S(G)$$

$$- U_R(G)U_{R'}(G),$$

which gives (C.76). As a special case,

$$\begin{aligned}
\mathrm{var}_*\Big[U_R(\mathbb{G}_b^*)\Big] &= \Big\{\binom{b}{r}^{-2}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q}} c_S\binom{b}{2r-q}U_S(G)\Big\} - [U_R(G)]^2 \\
&= \binom{b}{r}^{-2}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q}} c_S\frac{\binom{b}{2r-q}}{\binom{n}{2r-q}}X_S(G) - [U_R(G)]^2 \\
&= \binom{b}{r}^{-2}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q}} c_S\frac{\binom{n-2r+q}{b-2r+q}}{\binom{n}{b}}X_S(G) - [U_R(G)]^2 \\
&= \binom{b}{r}^{-2}\Big[\sum_{S\in\mathcal{S}_{R,R}^{(0)}} c_S\frac{\binom{n-2r}{b-2r}}{\binom{n}{b}}X_S(G) + \sum_{q=1}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q}} c_S\frac{\binom{n-2r+q}{b-2r+q}}{\binom{n}{b}}X_S(G)\Big] - [U_R(G)]^2,
\end{aligned}$$

which aligns with the results in Bhattacharyya and Bickel (2015a).

iv) It remains to examine the total covariance in terms of network vertex subsampling. First, it holds that $\mathrm{cov}\{E_*[U_R(\mathbb{G}_b^*)], E_*[U_{R'}(\mathbb{G}_b^*)]\} = \mathrm{cov}[U_R(\mathbb{G}_n), U_{R'}(\mathbb{G}_n)] = E[U_R(\mathbb{G}_n)U_{R'}(\mathbb{G}_n)] - E[U_R(\mathbb{G}_n)]E[U_{R'}(\mathbb{G}_n)]$, where the first equality follows (C.75).

Second, we have

$$\begin{aligned}
&E\Big\{\mathrm{cov}_*\big[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\big]\Big\} \\
\overset{(C.76)}{=}& E\Big\{\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}U_S(\mathbb{G}_n) - U_R(\mathbb{G}_n)U_{R'}(\mathbb{G}_n)\Big\} \\
=& \binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}E[U_S(\mathbb{G}_n)] - E\big[U_R(\mathbb{G}_n)U_{R'}(\mathbb{G}_n)\big].
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\operatorname{cov}\Big\{E_*\big[U_R(\mathbb{G}_b^*)\big], E_*\big[U_{R'}(\mathbb{G}_b^*)\big]\Big\} + E\Big\{\operatorname{cov}_*\big[U_R(\mathbb{G}_b^*), U_{R'}(\mathbb{G}_b^*)\big]\Big\} \\
=&\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}E[U_S(\mathbb{G}_n)] - E\Big[U_R(\mathbb{G}_n)\Big]E\Big[U_{R'}(\mathbb{G}_n)\Big] \\
\overset{(C.108)}{=}&\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\binom{b}{s}E[U_S(\mathbb{G}_b^{h_n})] - E\Big[U_R(\mathbb{G}_b^{h_n})\Big]E\Big[U_{R'}(\mathbb{G}_b^{h_n})\Big] \\
\overset{(C.101)}{=}&\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S E[X_S(\mathbb{G}_b^{h_n})] - E\Big[U_R(\mathbb{G}_b^{h_n})\Big]E\Big[U_{R'}(\mathbb{G}_b^{h_n})\Big] \\
=&E\Big[\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S X_S(\mathbb{G}_b^{h_n})\Big] - E\Big[U_R(\mathbb{G}_b^{h_n})\Big]E\Big[U_{R'}(\mathbb{G}_b^{h_n})\Big] \\
\overset{(C.62)}{=}&E\Big[\binom{b}{r}^{-1}\binom{b}{r'}^{-1} X_R(\mathbb{G}_b^{h_n})X_{R'}(\mathbb{G}_b^{h_n})\Big] - E\Big[U_R(\mathbb{G}_b^{h_n})\Big]E\Big[U_{R'}(\mathbb{G}_b^{h_n})\Big] \\
=&\operatorname{cov}\Big[U_R(\mathbb{G}_b^{h_n}), U_{R'}(\mathbb{G}_b^{h_n})\Big],
\end{aligned}
$$

which gives (C.78).

□

C.5 Proofs for Section C.1

C.5.1 Proof of Proposition 43

Let T denote a general finite population U-statistic. The following Hoeffding's decomposition represents T as the sum of mutually uncorrelated U-statistics of increasing order:

$$
T = E_*(T) + \sum_{1\leqslant i\leqslant b} g_1(\mathbb{V}_i) + \sum_{1\leqslant i<j\leqslant b} g_2(\mathbb{V}_i, \mathbb{V}_j) + \cdots . \tag{C.109}
$$

Bloznelis and Götze (2001, 2002) showed that such decomposition is unique and orthogonal, which implies that $\{g_i\}_{i=1}^b$ are centered and satisfy

$$E_*[g_i(\mathbb{V}_1, \cdots, \mathbb{V}_i) \mid \mathbb{V}_1, \cdots, \mathbb{V}_{i-1}] = 0. \tag{C.110}$$

In addition, Equation (2.3) in Bloznelis and Götze (2001) also showed that

$$g_1(\mathbb{V}_1) = \frac{n-1}{n-b} h_1(\mathbb{V}_1), \tag{C.111}$$

where $h_1(\mathbb{V}_1) = E_*[T - E_*(T) \mid \mathbb{V}_1]$.

We first discuss why $U_R(\mathbb{G}_b^*)$ is a finite population U-statistic. Since the network G can be treated as a population $G = \{\mathfrak{v}_1, \cdots, \mathfrak{v}_n\}$, the subsampled network $\mathbb{G}_b^*$ is uniquely determined by a random sample $\{\mathbb{V}_1, \cdots, \mathbb{V}_b\}$. Consequently, $U_R(\mathbb{G}_b^*)$ is a statistic based on $\mathbb{G}_b^*$, and is invariant of its permutation. Thus, $U_R(\mathbb{G}_b^*)$ is a finite population U-statistic by definition 35.

We next present the following technical auxiliary lemma, whose proof is in Section C.2.

Lemma 48. *For any motifs R and R',*

$$\mathrm{cov}_{\mathbb{V}_1 *}\Big[\sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1 *})} X_R(\mathscr{G}), \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1 *})} X_{R'}(\mathscr{G})\Big] = \sum_{q=0}^{\min\{r,r'\}} \sum_{S \in \mathcal{S}_{R,R'}^{(q)}} c_S \frac{nq - rr'}{n^2} X_S(G) \tag{C.112}$$

Now we start to prove Proposition 43.

of Proposition 43. We start by proving the results in part (a).

(a).i We first show (C.80). As discussed before, $U_R(\mathbb{G}_b^*)$ is a finite population U-

statistic. Thus, from (C.111) we have

$$
\begin{aligned}
&h_{1,R}\left(\mathbb{V}_1\right) = E_*\big[U_R(\mathbb{G}_b^*) - E_*[U_R(\mathbb{G}_b^*)] \mid \mathbb{V}_1\big] = E_*\big[U_R(\mathbb{G}_b^*) \mid \mathbb{V}_1\big] - E_*\big[U_R(\mathbb{G}_b^*)\big], \\
&g_{1,R}\left(\mathbb{V}_1\right) = \frac{n-1}{n-b} h_{1,R}\left(\mathbb{V}_1\right).
\end{aligned}
\tag{C.113}
$$

We now study $h_{1,R}\left(\mathbb{V}_1\right)$. Recall that $\mathbb{G}_b^{v*}$ denotes a random induced graph of G with vertex v and other $b-1$ random vertices drawn without replacement from $V(G) \setminus v$. Thus,

$$
\begin{aligned}
U_R(\mathbb{G}_b^*) \mid (\mathbb{V}_1 = \mathfrak{v}_1) =& U_R(\mathbb{G}_b^{v_1*}) = \binom{b}{r}^{-1} X_R(\mathbb{G}_b^{v_1*}) \\
\overset{\text{(C.101)}}{=}& \binom{b}{r}^{-1} \Big|\{S : S \subset \mathbb{G}_b^{v_1*}, S \cong R\}\Big|.
\end{aligned}
\tag{C.114}
$$

Next, we break the $\{S : S \subset \mathbb{G}_b^{v_1*}, S \cong R\}$ into two disjoint sets as

$$
\begin{aligned}
&\{S : S \subset \mathbb{G}_b^{v_1*}, S \cong R\} \\
=&\{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \notin V(S), S \cong R\} \cup \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \in V(S), S \cong R\}.
\end{aligned}
$$

since

$$
\{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \notin V(S), S \cong R\} \cap \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \in V(S), S \cong R\} = \phi,
$$

we can partition $U_R(\mathbb{G}_b^*) \mid (\mathbb{V}_1 = \mathfrak{v}_1)$ into two parts as follows

$$
\begin{aligned}
U_R(\mathbb{G}_b^*) \mid \mathfrak{v}_1 =& \binom{b}{r}^{-1} \Big| \{S : S \subset \mathbb{G}_b^{v_1*}, S \cong R\} \Big| \\
=& \binom{b}{r}^{-1} \Big\{ \Big| \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \notin V(S), , S \cong R\} \Big| + \Big| \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \in V(S), S \cong R\} \Big| \Big\} \\
=& \binom{b}{r}^{-1} \Big\{ \Big| \{S : S \subset \mathbb{G}_b^{v_1*} \setminus v_1, S \cong R\} \Big| + \Big| \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \in V(S), S \cong R\} \Big| \Big\} \\
\overset{(C.101)}{=}& \binom{b}{r}^{-1} X_R(\mathbb{G}_b^{v_1*} \setminus v_1) + \binom{b}{r}^{-1} \Big| \{S : S \subset \mathbb{G}_b^{v_1*}, v_1 \in V(S), S \cong R\} \Big| \\
\overset{(C.71)}{=}& \binom{b}{r}^{-1} X_R(\mathbb{G}_b^{v_1*} \setminus v_1) + \binom{b}{r}^{-1} \sum_{\mathscr{G} \in \mathcal{S}(\mathbb{G}_{b,r}^{v_1**})} X_R(\mathscr{G}) \\
:=& \, \mathrm{I} + \mathrm{II},
\end{aligned}
\tag{C.115}
$$

where $\mathbb{G}_b^{v_1*} \setminus v_1$ is a a random induced subsampled graph based on $b-1$ vertices that are randomly drawn without replacement from $V(G) \setminus v_1$. Let $G' = G \setminus v_1$ be the network after removing node v_1 and all edges that connected with v_1 from G. Then $\mathbb{G}_b^{v_1*} \setminus v_1$ is essentially a random induced graph $\mathbb{G}'^*_{b-1}$. In addition, we use $E_{*\setminus v_1}$ to indicate probability calculations with respect to other $b-1$ random vertices without v_1.

For part I, we have

$$\begin{aligned}
E_{*\setminus v_1}\big[\mathrm{I}\big] =& E_{*\setminus v_1}\Big[\binom{b}{r}^{-1} X_R(\mathbb{G}_b^{v_1*}\setminus v_1)\Big] = \frac{b-r}{b} E_{*\setminus v_1}\Big[\binom{b-1}{r}^{-1} X_R(\mathbb{G}'^{*}_{b-1})\Big] \\
=& \frac{b-r}{b} E_{*\setminus v_1}\Big[U_R(\mathbb{G}'^{*}_{b-1})\Big] \overset{(C.75)}{=} \frac{b-r}{b} U_R(G') \\
=& \frac{b-r}{b} U_R(G\setminus v_1) \overset{(C.101)}{=} \frac{b-r}{b}\binom{n-1}{r}^{-1} X_R(G\setminus v_1) \\
\overset{(C.101)}{=}& \frac{b-r}{b}\binom{n-1}{r}^{-1}\Big(\Big|\{S: S\subset G, S\cong R\}\Big| - \Big|\{S: S\subset G, v_1\in V(S), S\cong R\}\Big|\Big) \\
=& \frac{b-r}{b}\binom{n-1}{r}^{-1}\Big(X_R(G) - \Big|\{S: S\subset G, v_1\in V(S), S\cong R\}\Big|\Big) \\
\overset{(C.72)}{=}& \frac{b-r}{b}\binom{n-1}{r}^{-1}\Big(X_R(G) - \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G})\Big).
\end{aligned}$$

For part II, we have

$$\begin{aligned}
E_{*\setminus v_1}\big(\mathrm{II}\big) =& E_{*\setminus v_1}\Big[\binom{b}{r}^{-1} \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{b,r}^{v_1**})} X_R(\mathscr{G})\Big] \\
=& \binom{b}{r}^{-1}\binom{n-1}{b-1}^{-1} \sum_{G_b^*\in\mathcal{S}(\mathbb{G}_b^{v_1*})}\Big(\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{b,r}^{v_1**})} X_R(\mathscr{G})\Big) \\
\overset{(C.73)}{=}& \binom{b}{r}^{-1}\binom{n-1}{b-1}^{-1}\binom{n-r}{b-r}\Big|\{S: S\subset G, v_1\in V(S), S\cong R\}\Big| \\
\overset{(C.72)}{=}& \frac{r!(n-r)!}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G}).
\end{aligned}$$

Taking these two parts together, we have

$$
\begin{aligned}
&E_*\big[U_R(\mathbb{G}_b^*) \mid \mathfrak{v}_1\big] = E_{*\setminus v_1}\big[U_R(\mathbb{G}_b^*) \mid \mathfrak{v}_1\big] = E_{*\setminus v_1}\big[\mathrm{I} + \mathrm{II}\big] \\
&= \frac{b-r}{b}\binom{n-1}{r}^{-1}\Big[X_R(G) - \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G})\Big] + \frac{r!(n-r!)}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G}) \\
&= \frac{(b-r)n}{b(n-r)}\binom{n}{r}^{-1} X_R(G) - \frac{(b-r)r!(n-r-1)}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G}) \\
&\quad + \frac{(n-r)r!(n-r-1)!}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G}) \\
&= \frac{(b-r)n}{b(n-r)}U_R(G) + \frac{r!(n-r-1)!(n-b)}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_1*})} X_R(\mathscr{G}).
\end{aligned}
$$

On the other hand, Equation (C.75) implies that $E_*\big[U_R(\mathbb{G}_b^*)\big] = U_R(G)$. Consequently,

$$
\begin{aligned}
h_{1,R}\left(\mathbb{V}_1\right) &\overset{(C.113)}{=} E_*\big[U_R(\mathbb{G}_b^*) - E_*[U_R(\mathbb{G}_b^*)] \mid \mathbb{V}_1\big] = E_*\big[U_R(\mathbb{G}_b^*) \mid \mathbb{V}_1\big] - E_*\big[U_R(\mathbb{G}_b^*)\big] \\
&= \frac{(b-r)n}{b(n-r)}U_R(G) + \frac{r!(n-r-1)!(n-b)}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}) - U_R(G) \\
&= \frac{r!(n-r-1)!(n-b)}{b(n-1)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}) - \frac{(n-b)r}{b(n-r)}U_R(G),
\end{aligned}
$$

and

$$
g_{1,R}\left(\mathbb{V}_1\right) \overset{(C.113)}{=} \frac{n-1}{n-b}h_{1,R}\left(\mathbb{V}_1\right) = \frac{r!(n-r-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}) - \frac{(n-1)r}{b(n-r)}U_R(G).
$$

We now further study $g_{1,R}(\mathbb{V}_1)$. For the first part, we have

$$\begin{aligned}
&\frac{r!(n-r-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_{1^*}})} X_R(\mathscr{G}) \\
\overset{(C.72)}{=}\ &\frac{r!(n-r-1)!}{b(n-2)!}\Big[\big|\{S: S\subset G, \mathbb{V}_1\in V(S), S\cong R\}\big|\Big] \\
=\ &\frac{r!(n-r-1)!}{b(n-2)!}\Big[\big|\{S: S\subset G, S\cong R\}\big| - \big|\{S: S\subset G, \mathbb{V}_1\notin V(S), S\cong R\}\big|\Big] \\
=\ &\frac{r!(n-r-1)!}{b(n-2)!}\Big[X_R(G) - X_R(G\setminus\mathbb{V}_1)\Big] \\
=\ &\frac{r!(n-r-1)!}{b(n-2)!}X_R(G) - \frac{r!(n-r-1)!}{b(n-2)!}X_R(G\setminus\mathbb{V}_1).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&g_{1,R}(\mathbb{V}_1) = \frac{r!(n-r-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_{1^*}})} X_R(\mathscr{G}) - \frac{(n-1)r}{b(n-r)}U_R(G) \\
=\ &\frac{r!(n-r-1)!}{b(n-2)!}X_R(G) - \frac{r!(n-r-1)!}{b(n-2)!}X_R(G\setminus\mathbb{V}_1) - \frac{(n-1)r}{b(n-r)}U_R(G) \\
=\ &\frac{n(n-1)}{b(n-r)}U_R(G) - \frac{(n-1)r}{b(n-r)}U_R(G) - \frac{(n-1)r!(n-r-1)!}{b(n-1)!}X_R(G\setminus\mathbb{V}_1) \\
=\ &\frac{(n-1)}{b}[U_R(G) - U_R(G\setminus\mathbb{V}_1)],
\end{aligned} \tag{C.116}$$

which gives (C.80). For completeness, we show that the expectation of $g_{1,R}(\mathbb{V}_1)$ is zero. Notice that $\mathbb{V}_1$ could be any element in $\{\mathfrak{v}_1,\cdots,\mathfrak{v}_n\}$, and we use $E_{\mathbb{V}_{1*}}$

to indicate probability calculations with respect to random $\mathbb{V}_1$.

$$\begin{aligned}
E_*\big[g_{1,R}(\mathbb{V}_1)\big] =& E_{\mathbb{V}_1*}\big[g_{1,R}(\mathbb{V}_1)\big] \\
=& E_{\mathbb{V}_1*}\Big[\frac{r!(n-r-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}) - \frac{(n-1)r}{b(n-r)}U_R(G)\Big] \\
=& \frac{r!(n-r-1)!}{b(n-2)!}E_{\mathbb{V}_1*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G})\Big] - \frac{(n-1)r}{b(n-r)}U_R(G) \\
=& \frac{r!(n-r-1)!}{nb(n-2)!}\sum_{i=1}^{n}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\Big] - \frac{(n-1)r}{b(n-r)}U_R(G) \\
\overset{(C.74)}{=}& \frac{r!(n-r-1)!}{nb(n-2)!}rX_R(G) - \frac{(n-1)r}{b(n-r)}U_R(G) \\
=& \frac{r!(n-r-1)!}{nb(n-2)!}\frac{rn!}{r!(n-r)!}U_R(G) - \frac{(n-1)r}{b(n-r)}U_R(G) \\
=& \frac{(n-1)r}{b(n-r)}U_R(G) - \frac{(n-1)r}{b(n-r)}U_R(G) = 0.
\end{aligned} \tag{C.117}$$

(a).ii Now we want to show (C.81). We use $\text{var}_{\mathbb{V}_1*}$ and $\text{cov}_{\mathbb{V}_1*}$ to indicate probability calculations with respect to random $\mathbb{V}_1$. Because the randomness in $\text{cov}_*\left[g_{1,R}(\mathbb{V}_1)\right]$ is from the random vertex $\mathbb{V}_1$, we have

$$\text{cov}_*\left[g_{1,R}(\mathbb{V}_1)\right] = \text{cov}_{\mathbb{V}_1*}\left[g_{1,R}(\mathbb{V}_1)\right]. \tag{C.118}$$

Then based on lemma 48, we have

$$\begin{aligned}
&\mathrm{cov}_{\mathbb{V}_1*}\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big] \\
\overset{(C.80)}{=}&\mathrm{cov}_{\mathbb{V}_1*}\Big[\frac{r!(n-r-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})}X_R(\mathscr{G}), \frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1*})}X_{R'}(\mathscr{G})\Big] \\
=&\frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\mathrm{cov}_{\mathbb{V}_1*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})}X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1*})}X_{R'}(\mathscr{G})\Big] \\
\overset{(C.112)}{=}&\frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}c_S\frac{nq-rr'}{n^2}X_S(G).
\end{aligned} \tag{C.119}$$

Thus, (C.81) holds as a special case,

$$\mathrm{var}_{\mathbb{V}_1*}\left[g_{1,R}\left(\mathbb{V}_1\right)\right] = \left(\frac{r!(n-r-1)!}{b(n-2)!}\right)^2\Big[\sum_{q=0}^{r}\sum_{S\in S_{R,R}^{(k)}}c_S\frac{nk-r^2}{n^2}X_S(G)\Big].$$

(a).iii Now we continue to show (C.82). Recall that the subscript $\mathbb{V}_i*$ indicates that the randomness is from the random vertex $\mathbb{V}_i$, and the subscripte $\mathbb{V}_i*, \mathbb{V}_j*$ indicates that the randomness comes from two random vertices $\mathbb{V}_i$ and $\mathbb{V}_j$. Notice that $(n-1)rU_R(G)/b(n-r)$ and $(n-1)r'U_{R'}(G)/b(n-r')$ are two

constants when G is given. We first break the covariance into

$$\begin{aligned}
&\mathrm{cov}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i), \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] = \sum_{i=1}^{b}\sum_{j=1}^{b} \mathrm{cov}_{\mathbb{V}_i,\mathbb{V}_{j*}}\Big[g_{1,R}(\mathbb{V}_i), g_{1,R'}(\mathbb{V}_j)\Big] \\
\overset{(C.80)}{=}& \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{i=1}^{b}\sum_{j=1}^{b} \mathrm{cov}_{\mathbb{V}_i,\mathbb{V}_{j*}}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big] \\
=& \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{i=1}^{b}\Bigg\{\mathrm{cov}_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i*})} X_{R'}(\mathscr{G})\Big] \\
&+ \sum_{j=1,j\neq i}^{b} \mathrm{cov}_{\mathbb{V}_i,\mathbb{V}_{j*}}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big]\Bigg\} \\
:=& \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{i=1}^{b}\Big(\mathrm{I}+\mathrm{II}\Big).
\end{aligned} \tag{C.120}$$

Part I can be further derived as

$$\begin{aligned}
\mathrm{I} &= \mathrm{cov}_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i*})} X_{R'}(\mathscr{G})\Big] \\
&= E_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i*})} X_{R'}(\mathscr{G})\Big] - E_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G})\Big] E_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i*})} X_{R'}(\mathscr{G})\Big] \\
&= \frac{1}{n}\sum_{k=1}^{n}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_k*})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_k*})} X_{R'}(\mathscr{G})\Big] - \Big[\frac{1}{n}\sum_{k=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_k*})} X_R(\mathscr{G})\Big]\Big[\frac{1}{n}\sum_{k=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_k*})} X_{R'}(\mathscr{G})\Big] \\
&\overset{(C.74)}{=} \frac{1}{n}\sum_{k=1}^{n}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_k*})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_k*})} X_{R'}(\mathscr{G})\Big] - \frac{rX_R(G)}{n}\frac{r'X_{R'}(G)}{n}.
\end{aligned} \tag{C.121}$$

For part II, first we have,

$$
\begin{aligned}
&\operatorname{cov}_{\mathbb{V}_i,\mathbb{V}_j*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G}),\ \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big]\\
=&E_{\mathbb{V}_i,\mathbb{V}_j*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G})\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big]-E_{\mathbb{V}_i*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G})\Big]E_{\mathbb{V}_j*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big]\\
=&E_{\mathbb{V}_i,\mathbb{V}_j*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i*})} X_R(\mathscr{G})\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j*})} X_{R'}(\mathscr{G})\Big]-\Big[\frac{1}{n}\sum_{k=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_k*})} X_R(\mathscr{G})\Big]\Big[\frac{1}{n}\sum_{k=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_k*})} X_{R'}(\mathscr{G})\Big]\\
\overset{(C.74)}{=}&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_j*})} X_{R'}(\mathscr{G})\Big]-\frac{rX_R(G)r'X_{R'}(G)}{n^2}\\
=&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\Big[\sum_{j=1,j\neq i}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_j*})} X_{R'}(\mathscr{G})\Big]-\frac{rX_R(G)r'X_{R'}(G)}{n^2}\\
=&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\Big[\sum_{j=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_j*})} X_{R'}(\mathscr{G})-\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i*})} X_{R'}(\mathscr{G})\Big]-\frac{rX_R(G)r'X_{R'}(G)}{n^2}\\
=&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\Big[rX(R',G)-\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i*})} X_{R'}(\mathscr{G})\Big]-\frac{rX_R(G)r'X_{R'}(G)}{n^2}\\
=&\frac{rX_{R'}(G)}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i*})} X_{R'}(\mathscr{G})-\frac{rX_R(G)r'X_{R'}(G)}{n^2}\\
\overset{(C.74)}{=}&\frac{rX_R(G)r'X_{R'}(G)}{n^2(n-1)}-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i*})} X_R(\mathscr{G})\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i*})} X_{R'}(\mathscr{G}).
\end{aligned}
\tag{C.122}
$$

Consequently, we have

$$\begin{aligned}
\mathrm{II} &= \sum_{j=1,j\neq i}^{b} \mathrm{cov}_{\mathbb{V}_i,\mathbb{V}_j *}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i *})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_j *})} X_{R'}(\mathscr{G})\Big] \\
&\overset{(C.122)}{=} \sum_{j=1,j\neq i}^{b} \Big\{\frac{rX_R(G)r'X(R',G)}{n^2(n-1)} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i *})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i *})} X_{R'}(\mathscr{G})\Big\} \\
&= (b-1)\Big\{\frac{rX_R(G)r'X(R',G)}{n^2(n-1)} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i *})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i *})} X_{R'}(\mathscr{G})\Big\}.
\end{aligned}$$

By adding I and II following previous results, we have

$$\begin{aligned}
\mathrm{I}+\mathrm{II} &= \frac{n-b}{n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{v_i *})} X_R(\mathscr{G}) \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{v_i *})} X_{R'}(\mathscr{G}) - \frac{rX_R(G)r'X_{R'}(G)}{n^2}\Big\} \\
&\overset{(C.121)}{=} \frac{n-b}{n-1}\mathrm{cov}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i *})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i *})} X_{R'}(\mathscr{G})\Big].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{cov}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i), \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] \\
\overset{(C.120)}{=}& \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{i=1}^{b}\Big(I+II\Big) \\
=& \frac{n-b}{n-1}\frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{i=1}^{b}\mathrm{cov}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_i *})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_i *})} X_{R'}(\mathscr{G})\Big] \\
=& \frac{b(n-b)}{n-1}\frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\mathrm{cov}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1 *})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1 *})} X_{R'}(\mathscr{G})\Big] \\
\overset{(C.80)}{=}& \frac{b(n-b)}{n-1}\mathrm{cov}_*\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big],
\end{aligned} \tag{C.123}$$

which gives (C.82).

(a).iv To show (C.83), we start by breaking break the variance into

$$
\begin{aligned}
&\operatorname{var}_* \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right) \overset{(C.82)}{=} \frac{b(n-b)}{(n-1)} \operatorname{var}_*\Big[g_{1,R}\left(\mathbb{V}_1\right)\Big] \\
\overset{(C.81)}{=}& \frac{b(n-b)}{(n-1)}\left(\frac{r!(n-r-1)!}{b(n-2)!}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in S_{R,R}^{(q)}} c_S \frac{nq-r^2}{n^2} X_S(G)\Big] \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\right)^2 \Big\{\Big[\sum_{q=0}^{r}\sum_{S\in S_{R,R}^{(q)}} c_S \frac{q}{n} X_S(G)\Big] - \Big[\sum_{q=0}^{r}\sum_{S\in S_{R,R}^{(q)}} c_S \frac{r^2}{n^2} X_S(G)\Big]\Big\} \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} c_S \frac{q}{n} X_S(G)\Big] \\
&- \frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} c_S \frac{r^2}{n^2} X_S(G)\Big] \\
=& \ \mathrm{I} - \mathrm{II}.
\end{aligned}
$$

For part I, we have

$$
\begin{aligned}
&\frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} c_S \frac{q}{n} X_S(G)\Big] \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}\binom{n-1}{r-1}^{-1}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} \frac{c_S q}{n}\binom{n}{2r-q} U_S(G)\Big] \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}\right)^2 \Big[\sum_{q=0}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} \frac{c_S q}{n}\binom{n-1}{r-1}^{-2}\binom{n}{2r-q} U_S(G)\Big] \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}\right)^2 \Big[\sum_{q=1}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} \frac{c_S q}{n}\left(\frac{(r-1)!(n-r)!}{(n-1)!}\right)^2 \frac{n!}{(2r-q)!(n-2r+q)} U_S(G)\Big] \\
=& \frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}\right)^2 \Big[\sum_{q=1}^{r}\sum_{S\in \mathcal{S}_{R,R}^{(q)}} \frac{q[(r-1)!]^2 c_S}{(2r-q)!}\frac{(n-r)(n-r-1)\cdots(n-2r+q+1)}{(n-1)(n-2)\cdots(n-r+1)} U_S(G)\Big].
\end{aligned}
$$

Notice that $(n-r)(n-r-1)\cdots(n-2r+q+1)/(n-1)(n-2)\cdots(n-r+1)$

has $r-q$ items in the numerator and $r-1$ items in the denominator, and $U_S(G)<1$ for all S. Thus,

$$\lim_{b,n\to\infty}\frac{(n-b)}{b(n-1)}\Big[\frac{r!(n-r-1)!}{(n-2)!}\Big]^2\Big[\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}c_S\frac{q}{n}X_S(G)\Big]$$
$$=\lim_{b,n\to\infty}\frac{(n-b)}{b(n-1)}\Big[\frac{r(n-1)}{(n-r)}\Big]^2\sum_{q=1}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}\frac{q[(r-1)!]^2c_S}{(2r-q)!}\frac{(n-r)(n-r-1)\cdots(n-2r+q+1)}{(n-1)(n-2)\cdots(n-r+1)}U_S(G)$$
$$=\lim_{b,n\to\infty}\frac{1}{b}\sum_{S\in\mathcal{S}_{R,R}^{(1)}}\frac{c_S(r!)^2}{(2r-1)!}=0.$$

For par II,

$$\frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\right)^2\Big[\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}c_S\frac{r^2}{n^2}X_S(G)\Big]$$
$$=\frac{(n-b)}{b(n-1)}\left(\frac{r!(n-r-1)!}{(n-2)!}\frac{r}{n}\right)^2\Big[\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}c_SX_S(G)\Big]\overset{(C.62)}{=}\left(\frac{r!(n-r-1)!}{b(n-2)!}\frac{r}{n}\right)^2X_R(G)X_R(G)$$
$$=\frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}\right)^2\binom{n}{r}^{-2}X_R(G)X_R(G)=\frac{(n-b)}{b(n-1)}\left(\frac{r(n-1)}{(n-r)}U_R(G)\right)^2\xrightarrow{n,b}0.$$

Therefore, $\lim_{b,n\to\infty}\text{var}_*\left[\sum_{1\leqslant i\leqslant b}g_{1,R}\left(\mathbb{V}_1\right)\right]=\lim_{b,n\to\infty}\text{I}-\text{II}=0$, which gives (C.83).

Next we continue to prove part (b) based on the results in part (a).

(b).i We now show (C.84), which is an trivial extension of (C.80). Because $U_R(\mathbb{G}_b^*)+U_{R'}(\mathbb{G}_b^*)$ is a symmetric finite population statistic, lemma (C.111) implies that

$$h_{1,R,R'}\left(\mathbb{V}_1\right)=E_*\Big[U_R(\mathbb{G}_b^*)+U_{R'}(\mathbb{G}_b^*)-E_*\big[U_R(\mathbb{G}_b^*)+U_{R'}(\mathbb{G}_b^*)\big]\mid\mathbb{V}_1\Big],$$
$$g_{1,R,R'}\left(\mathbb{V}_1\right)=\frac{n-1}{n-b}h_{1,R,R'}\left(\mathbb{V}_1\right).$$

By the linearity of conditional expectation, we have

$$
\begin{aligned}
g_{1,R,R'}(\mathbb{V}_1) &\overset{(C.113)}{=} \frac{n-1}{n-b} h_{1,R,R'}(\mathbb{V}_1) \\
&= \frac{n-1}{n-b}\left(E_*\Big\{U_R(\mathbb{G}_b^*) + U_{R'}(\mathbb{G}_b^*) - E_*\big[U_R(\mathbb{G}_b^*) + U_{R'}(\mathbb{G}_b^*)\big] \mid \mathbb{V}_1\Big\}\right) \\
&= \frac{n-1}{n-b}\left(E_*\Big\{U_R(\mathbb{G}_b^*) - E_*\big[U_R(\mathbb{G}_b^*)\big] \mid \mathbb{V}_1\Big\} + E_*\Big\{U_{R'}(\mathbb{G}_b^*) - E_*\big[U_{R'}(\mathbb{G}_b^*)\big] \mid \mathbb{V}_1\Big\}\right) \\
&\overset{(C.113)}{=} \frac{n-1}{n-b} h_{1,R}(\mathbb{V}_1) + \frac{n-1}{n-b} h_{1,R'}(\mathbb{V}_1) \\
&\overset{(C.113)}{=} g_{1,R}(\mathbb{V}_1) + g_{1,R'}(\mathbb{V}_1).
\end{aligned}
$$

(b).ii We now derive (C.85) based on (C.82) as follows.

$$
\begin{aligned}
\mathrm{var}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R,R'}(\mathbb{V}_i)\Big] &= \mathrm{var}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i) + \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] \\
&= \mathrm{var}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i)\Big] + \mathrm{var}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] + 2\mathrm{cov}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i), \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] \\
&\overset{(C.82)}{=} \frac{b(n-b)}{n-1}\Big\{\mathrm{var}_*\Big[g_{1,R}(\mathbb{V}_1)\Big] + \mathrm{var}_*\Big[g_{1,R'}(\mathbb{V}_1)\Big]\Big\} + 2\mathrm{cov}_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i), \sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i)\Big] \\
&\overset{(C.123)}{=} \frac{b(n-b)}{n-1}\Big\{\mathrm{var}_*\Big[g_{1,R}(\mathbb{V}_1)\Big] + \mathrm{var}_*\Big[g_{1,R'}(\mathbb{V}_1)\Big] + 2\mathrm{cov}_*\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big]\Big\} \\
&= \frac{b(n-b)}{(n-1)}\mathrm{var}_*\Big[g_{1,R,R'}(\mathbb{V}_1)\Big].
\end{aligned}
$$

Our result is consistent with Lemma 1 of Bloznelis and Götze (2001) which focusing on the variance of general finite population U-statistic.

(b).iii Now we turn to prove (C.86). First, we have

$$\begin{aligned}
\mathrm{var}_* \sum_{1\leqslant i\leqslant b} g_{1,R,R'}\left(\mathbb{V}_i\right) \overset{(C.82)}{=} & \frac{b(n-b)}{(n-1)}\mathrm{var}_*\Big[g_{1,R,R'}\left(\mathbb{V}_1\right)\Big] \\
= & \frac{b(n-b)}{(n-1)}\mathrm{var}_*\Big[g_{1,R}(\mathbb{V}_1)\Big] + \frac{b(n-b)}{(n-1)}\mathrm{var}_*\Big[g_{1,R'}(\mathbb{V}_1)\Big] \\
& + \frac{2b(n-b)}{(n-1)}\mathrm{cov}_*\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big] \\
\overset{(C.82)}{=} & \mathrm{var}_* \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right) + \mathrm{var}_* \sum_{1\leqslant i\leqslant b} g_{1,R'}\left(\mathbb{V}_i\right) \\
& + \frac{2b(n-b)}{(n-1)}\mathrm{cov}_*\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big] \\
= & \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

Because I and II are studied before when dealing with (C.83), we only need to focus on part III.

$$\begin{aligned}
\mathrm{III} = & \mathrm{cov}_*\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big] = \mathrm{cov}_{\mathbb{V}_1 *}\Big[g_{1,R}(\mathbb{V}_1), g_{1,R'}(\mathbb{V}_1)\Big] \\
= & \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\mathrm{cov}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1 *})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1 *})} X_{R'}(\mathscr{G})\Big] \\
\overset{(C.119)}{=} & \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{nq-rr'}{n^2} X_S(G) \\
= & \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\Big\{\Big[\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{q}{n} X_S(G)\Big] - \frac{rr'}{n^2} X_R(G) X_{R'}(G)\Big\} \\
= & \sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{r!(n-r-1)!}{b(n-2)!}\frac{r'!(n-r'-1)!}{b(n-2)!}\frac{q}{n} X_S(G) \\
& - \frac{rr'(n-1)(n-1)}{b^2(n-r)(n-r')} U_R(G) U_{R'}(G)
\end{aligned}$$

As $U_R(G)U_{R'}(G) < 1$, we have

$$\lim_{b,n\to\infty} \frac{rr'(n-1)(n-1)}{b^2(n-r)(n-r')} U_R(G)U_{R'}(G) = 0.$$

On the other hand,

$$\begin{aligned}
&\sum_{q=0}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{r!(n-r-1)!}{b(n-2)!} \frac{r'!(n-r'-1)!}{b(n-2)!} \frac{q}{n} X_S(G) \\
= &\sum_{q=0}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{r!(n-r-1)!}{b(n-2)!} \frac{r'!(n-r'-1)!}{b(n-2)!} \frac{q}{n} \binom{n}{r+r'-q} U_S(G) \\
= &\sum_{q=1}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{r!(n-r-1)!}{b(n-2)!} \frac{r'!(n-r'-1)!}{b(n-2)!} \frac{q}{n} \frac{n!}{(r+r'-q)!(n-r-r'+q)!} U_S(G) \\
= &\sum_{q=1}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S \frac{qr!r'!}{b^2(r+r'-q)!} \frac{(n-1)(n-r'-1)\cdots(n-r'-r+q+1)}{(n-2)(n-3)\cdots(n-r)} U_S(G)
\end{aligned}$$

Notice that $(n-1)(n-r'-1)\cdots(n-r'-r+q+1)/(n-2)(n-3)\cdots(n-r)$ has $r-q$ items in the numerator and $r-1$ items in the denominator, and $U_S(G) \leqslant 1$, we have

$$\begin{aligned}
&\sum_{q=0}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}_{RR'}^{(q)}} c_S \frac{r!(n-r-1)!}{b(n-2)!} \frac{r'!(n-r'-1)!}{b(n-2)!} \frac{q}{n} X_S(G) \\
\overset{n\to\infty}{\to} &\sum_{S\in\mathcal{S}_{RR'}^{(1)}} c_S \frac{r!r'!}{b^2(r+r'-1)!} U_S(G) \overset{b\to\infty}{\to} 0.
\end{aligned}$$

Consequently, we have $\lim_{b,n\to\infty} \mathrm{III} = 0$. Therefore,

$$\lim_{b,n\to\infty} var_*\Big[\sum_{1\leqslant i\leqslant b} g_{1,R,R'}\left(\mathbb{V}_i\right)\Big] = \lim_{b,n\to\infty} \mathrm{I}+\mathrm{II}+\mathrm{III} \overset{(C.83)}{=} \lim_{b,n\to\infty} \mathrm{III} = 0,$$

which gives (C.86).

□

C.5.2 Proof of Lemma 48

Proof. To begin with, we have

$$\begin{aligned}
&\mathrm{cov}_{\mathbb{V}_1*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}), \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1*})} X_{R'}(\mathscr{G})\Big] \\
\overset{(C.72)}{=}&\mathrm{cov}_{\mathbb{V}_1*}\Big[\big|\{S: S\subset G, \mathbb{V}_1\in V(S), S\cong R\}\big|, \big|\{S: S\subset G, \mathbb{V}_1\in V(S), S\cong R'\}\big|\Big] \\
\overset{(C.115)}{=}&\mathrm{cov}_{\mathbb{V}_1*}\Big[X_R\left(G\right)-X_R\left(G\setminus\mathbb{V}_1\right), X_{R'}\left(G\right)-X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] \\
=&\mathrm{cov}_{\mathbb{V}_1*}\Big[X_R\left(G\setminus\mathbb{V}_1\right), X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] \\
=&E_{\mathbb{V}_1*}\Big[X_R\left(G\setminus\mathbb{V}_1\right)X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] - E_{\mathbb{V}_1*}\Big[X_R\left(G\setminus\mathbb{V}_1\right)\Big]E\Big[X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] \\
=&\mathrm{I}-\mathrm{II}.
\end{aligned}$$

For part I, we have

$$
\begin{aligned}
&E_{\mathbb{V}_{1*}}\Big[X_R\left(G\setminus\mathbb{V}_1\right)X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big]=\frac{1}{n}\sum_{i=1}^{n}\Big[X_R\left(G\setminus v_i\right)X_{R'}\left(G\setminus v_i\right)\Big]\\
\overset{(C.62)}{=}&\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{S\in\mathscr{S}_{R,R'}}c_S X_S\left(G\setminus v_i\right)\Big]=\frac{1}{n}\sum_{S\in\mathscr{S}_{R,R'}}c_S\sum_{i=1}^{n}X_S\left(G\setminus v_i\right)\\
=&\frac{1}{n}\sum_{S\in\mathscr{S}_{R,R'}}c_S\sum_{i=1}^{n}\Big[X_S(G)-\big|\{H:H\subset G,v_i\in V(H),H\cong S\}\big|\Big]\\
=&\frac{1}{n}\sum_{S\in\mathscr{S}_{R,R'}}c_S\sum_{i=1}^{n}X_S(G)-\frac{1}{n}\sum_{S\in\mathscr{H}_{R,R'}}c_S\sum_{i=1}^{n}\big|\{H:H\subset G,v_i\in V(H),H\cong S\}\big|\\
\overset{(C.74)}{=}&\sum_{S\in\mathscr{S}_{R,R'}}c_S X_S(G)-\sum_{S\in\mathscr{S}_{R,R'}}c_S\frac{s}{n}X_S(G).
\end{aligned}
$$

For part II, we have

$$
\begin{aligned}
&E_{\mathbb{V}_{1*}}\Big[X_R\left(G\setminus\mathbb{V}_1\right)\Big]E_{\mathbb{V}_{1*}}\Big[X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big]=\frac{1}{n}\sum_{i=1}^{n}X_R\left(G\setminus v_i\right)\frac{1}{n}\sum_{i=1}^{n}X_{R'}\left(G\setminus v_i\right)\\
\overset{(C.115)}{=}&\frac{1}{n}\Big[\sum_{i=1}^{n}\Big(X_R\left(G\right)-\Big|\{S:S\subset G,v_i\in V(S),S\cong R\}\Big|\Big)\Big]\cdot\\
&\frac{1}{n}\Big[\sum_{i=1}^{n}\Big(X_{R'}\left(G\right)-\Big|\{S:S\subset G,v_i\in V(S),S\cong R'\}\Big|\Big)\Big]\\
\overset{(C.74)}{=}&\frac{1}{n}\Big[nX_R\left(G\right)-rX_R\left(G\right)\Big]\frac{1}{n}\Big[nX_{R'}\left(G\right)-r'X_{R'}\left(G\right)\Big]\\
=&\frac{(n-r)}{n}\frac{(n-r')}{n}X_R\left(G\right)X_{R'}\left(G\right)=\left(1-\frac{r+r'}{n}+\frac{rr'}{n^2}\right)X_R\left(G\right)X_{R'}\left(G\right)\\
\overset{(C.62)}{=}&\left(1-\frac{r+r'}{n}+\frac{rr'}{n^2}\right)\sum_{S\in\mathscr{S}_{R,R'}}c_S X_S(G).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\operatorname{cov}_{\mathbb{V}_1*}\Big[\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{\mathbb{V}_1*})} X_R(\mathscr{G}),\ \sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_{r'}^{\mathbb{V}_1*})} X_{R'}(\mathscr{G})\Big] \\
= \quad & E_{\mathbb{V}_1*}\Big[X_R\left(G\setminus\mathbb{V}_1\right)X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] - E_{\mathbb{V}_1*}\Big[X_R\left(G\setminus\mathbb{V}_1\right)\Big]E_{\mathbb{V}_1*}\Big[X_{R'}\left(G\setminus\mathbb{V}_1\right)\Big] \\
\overset{(C.62)}{=} & \sum_{S\in\mathcal{S}_{R,R'}} c_S X_S(G) - \sum_{S\in\mathcal{S}_{R,R'}} c_S\frac{s}{n}X_S(G) - \left(1-\frac{r+r'}{n}+\frac{rr'}{n^2}\right)\sum_{S\in\mathcal{S}_{R,R'}} c_S X_S(G) \\
= & \sum_{S\in\mathcal{S}_{R,R'}}\left(\frac{r+r'-s}{n}-\frac{rr'}{n^2}\right)c_S X_S(G) = \sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}} c_S\frac{nq-rr'}{n^2}X_S(G).
\end{aligned}
$$

□

C.6 Proofs for Results in Section C.1

C.6.1 Proof of Theorem 44

We start by introducing the setup of finite population asymptotic described in Bloznelis and Götze (2001): Suppose that there exist a sequence of finite populations $\{\mathscr{V}^{(k)}\}$, where $\mathscr{V}^{(k)} = \{\mathfrak{v}_1 \cdots \mathfrak{v}_{n_k}\}$ with $n_k \to \infty$ as $k \to \infty$. Consequently, $\{T_k\}$ is a sequence of finite population U-statistics, where $T_k = t_k(\mathbb{V}_1, \cdots, \mathbb{V}_{b_k})$ is based on samples $\{\mathbb{V}_1, \cdots, \mathbb{V}_{b_k}\}$ drawn without replacement from $\mathscr{V}^{(k)}$, with $b_k \to \infty$ as $k \to \infty$.

We next present some critical auxiliary lemmas for the proof. Lemma 49, 50 and 51 are the existed results. The details of Lemma 52 and 53 are deferred to Sections C.2 and C.2, respectively.

Lemma 49. *(Proposition 3 in Bloznelis and Götze (2001)) The Hoeffding's decomposition of T_k is*

$$T_k = E_*[T_k] + \sum_{1 \leqslant i \leqslant b_k} T_{1,k}(\mathbb{V}_i) + \Delta(T_k),$$

where $\sum_{1 \leqslant i \leqslant b_k} T_{1,k}(\mathbb{V}_i)$ is the linear term, and $\Delta(T_k)$ is the remainder. Suppose that

1) $E_*\Delta^2(T_k) = o(1)$.

2) There exist constants $c_1, c_2 > 0$ such that $0 < c_1 \leqslant \mathrm{var}_*(T_k) \leqslant c_2 < \infty$,

3) For every $\epsilon > 0$, $\lim_{k\to\infty} b_k E_*[T_{1,k}^2(\mathbb{V}_1) \mathbb{1}_{\{T_{1,k}^2(\mathbb{V}_1) > \epsilon\}}] = 0$.

Then $(T_k - E_*T_k)/(\mathrm{var}_*(T_k))$ is asymptotically standard normal. Note that the subscript $*$ is not used in Bloznelis and Götze (2001), and we add it here to distinguish the sourse of randomness.

Lemma 50. *For a general finite population U-statistic T, the Hoeffding's decomposition represents T as*

$$\begin{aligned} T =& E_* T + \sum_{1 \leqslant i \leqslant b} g_1(\mathbb{V}_i) + \sum_{1 \leqslant i < j \leqslant b} g_2(\mathbb{V}_i, \mathbb{V}_j) + \cdots \\ =& E_* T + S_1 + S_2 + S_3 + \cdots . \end{aligned} \tag{C.109}$$

Bloznelis and Götze (2002) demonstrated that the Hoeffding's decomposition of a finite population U-statistic is orthogonal in ℓ_2, leading to the following property:

$$E_*[S_a S_b] = 0, \qquad \text{if } a \neq b. \tag{C.124}$$

Lemma 51 (Theorem 1 of Bickel et al. (2011))**.** *Let $\iint w^2(u,v)dudv < \infty$.*

a) If $(n-1)\rho_n \to \infty$,

$$\begin{aligned} \frac{\widehat{\rho}_{\mathbb{G}_n}}{\rho_n} &\to 1 \text{ in probability,} \\ \sqrt{n}\Big(\frac{\widehat{\rho}_{\mathbb{G}_n}}{\rho_n} - 1\Big) &\to \mathcal{N}(0, \sigma^2) \text{ in distribution,} \end{aligned} \tag{C.125}$$

for some $\sigma^2 > 0$.

b) For any motif R, assume that $\iint w^{2\mathfrak{r}}(u,v)dudv < \infty$, also $\rho_n = \omega(n^{-1})$ if R is acyclic and $\rho_n = \omega(n^{-2/r})$ otherwise. Then

$$\begin{aligned} \widehat{\rho}_{\mathbb{G}_n}^{-\mathfrak{r}} U_R(\mathbb{G}_n) &\to \rho_n^{-\mathfrak{r}} E[U_R(\mathbb{G}_n)] \text{ in probability} \\ \sqrt{n}\Big[\widehat{\rho}_{\mathbb{G}_n}^{-\mathfrak{r}} U_R(\mathbb{G}_n) - \rho_n^{-\mathfrak{r}} E[U_R(\mathbb{G}_n)]\Big] &\to \mathcal{N}\big(0, \sigma_R^2\big) \text{ in distribution,} \end{aligned} \tag{C.126}$$

where σ_R^2 is defined in Assumption 3. In addition, Proposition 2 in Green and

Shalizi (2022) implies that

$$\sqrt{n}\Big[\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n) - \rho_n^{-\mathfrak{r}}E[U_R(\mathbb{G}_n)]\Big] \to \mathscr{N}\big(0, \sigma_R^2\big) \text{ in distribution.} \tag{C.127}$$

c) More generally, for m motifs $R_1, \cdots, R_m$ with sizes $\max\{r_1, \cdots, r_m\} \leqslant r$,

$$\sqrt{n}\Big\{\big[\widehat{\rho_{\mathbb{G}_n}}^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_n), \cdots, \widehat{\rho_{\mathbb{G}_n}}^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_n)\big] - \big[\rho_n^{-\mathfrak{r}_1}E[U_{R_1}(\mathbb{G}_n)], \cdots, \rho_n^{-\mathfrak{r}_m}E[U_{R_m}(\mathbb{G}_n)]\big]\Big\}$$
$$\to \mathscr{N}\big(0, \Sigma(\boldsymbol{R})\big) \text{ in distribution,} \tag{C.128}$$

where $\Sigma(\boldsymbol{R})$ depends on $R_1, \cdots, R_m$ and $(n-1)\rho_n$.

Lemma 52. *Let R and R' be two motifs with $\max\{r, r'\} \leq r_1$ and $\max\{\mathfrak{r}, \mathfrak{r}'\} \leq \mathfrak{r}_1$. Suppose that Assumption 2 holds after replacing r by r_1 and $\mathfrak{r}$ by $\mathfrak{r}_1$, and Assumption 1 holds. Then*

$$pr\Big\{\lim_{b\to\infty}\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}_*\big[\sqrt{b}U_R(\mathbb{G}_b^*), \sqrt{b}U_{R'}(\mathbb{G}_b^*)\big]$$
$$= \big(1-c_2\big)\lim_{b\to\infty}\rho_b^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}\big[\sqrt{b}U_R(\mathbb{G}_b), \sqrt{b}U_{R'}(\mathbb{G}_b)\big]\Big\} = 1.$$

Lemma 53. *For any motif R, under Assumptions 2,*

$$pr\left\{\lim_{n\to\infty}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n) = \frac{r!}{|\mathrm{Aut}(R)|}P_w(R)\right\} = 1. \tag{C.129}$$

Now we are in the position to show Theorem 44. For simplicity, let $k = n_k = n$ and $b_k = b_n$.

of Theorem 44. We start with part (a), and will prove the results one by one. For simplicity, we write $G = G^{(n)}$, $b = b_n$.

(a).i Since $U_R(\mathbb{G}_b^*)$ is a finite population U-statistic, $\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)$ is also a finite

population U-statistic. Thus, (C.87) can be obtained as

$$\begin{aligned}\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) \overset{(C.79)}{=}& \sqrt{b}\rho_n^{-\mathfrak{r}}\Big\{E_*[U_R(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right) + \sum_{1\leqslant i<j\leqslant b} g_{2,R}\left(\mathbb{V}_i, \mathbb{V}_j\right) + \cdots\Big\} \\ =& \sqrt{b}\rho_n^{-\mathfrak{r}}\Big\{E_*[U_R(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right)\Big\} + \Delta[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)] \\ =& E_*[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b} \sqrt{b}\rho_n^{-\mathfrak{r}} g_{1,R}\left(\mathbb{V}_i\right) + \Delta[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)] \\ \overset{(C.75)}{=}& \sqrt{b}\rho_n^{-\mathfrak{r}}U_R(G) + \sum_{1\leqslant i\leqslant b} \sqrt{b}\rho_n^{-\mathfrak{r}} g_{1,R}\left(\mathbb{V}_i\right) + \Delta[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)].\end{aligned} \tag{C.130}$$

(a).ii We now prove (C.88), which bound the accuracy of approximation of the linear part. To begin with, we have

$$E_*\Big(\Big\{\Delta[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\Big\}\Big\{\sum_{1\leqslant i\leqslant b} \sqrt{b}\rho_n^{-\mathfrak{r}} g_{1,R}\left(\mathbb{V}_i\right)\Big\}\Big) \overset{(C.124)}{=} 0. \tag{C.131}$$

Consequently,

$$
\begin{aligned}
&E_*\Big[\Delta^2[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\Big] \overset{(C.87)}{=} E_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) - E_*[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)] - \sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big]^2 \\
=&E_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) - E_*[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\Big]^2 + E_*\Big[\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big]^2 \\
&- 2E_*\Big[\Big(\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) - E_*[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\Big)\Big(\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big)\Big] \\
\overset{(C.87)}{=}&\mathrm{var}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)\Big] + E_*\Big[\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big]^2 \\
&- 2E_*\Big[\Big(\Delta[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)] + \sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big)\Big(\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big)\Big] \\
\overset{(C.131)}{=}&\mathrm{var}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)\Big] - E_*\Big[\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big]^2 \\
\overset{(C.117)}{=}&\mathrm{var}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)\Big] - \mathrm{var}_*\Big[\sum_{1\leqslant i\leqslant b}\sqrt{b}\rho_n^{-\mathfrak{r}}g_{1,R}\left(\mathbb{V}_i\right)\Big] = \mathrm{I} - \mathrm{II}.
\end{aligned}
\tag{C.132}
$$

Part I is the variance of the network moment. We study the covariance here

as it is helpful for the analysis later.

$$
\begin{aligned}
&\mathrm{cov}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*),\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*)\Big]=b\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\mathrm{cov}_*\Big[U_R(\mathbb{G}_b^*),U_{R'}(\mathbb{G}_b^*)\Big]\\
\overset{(C.76)}{=}&b\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\Big\{\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}c_S\binom{b}{s}U_S(G)-U_R(G)U_{R'}(G)\Big\}\\
=&b\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\Big\{\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}c_S\binom{b}{s}U_S(G)-\binom{n}{r}^{-1}\binom{n}{r'}^{-1}X_R(G)X_{R'}(G)\Big\}\\
\overset{(C.62)}{=}&b\rho_n^{-(\mathfrak{r}+\mathfrak{r}')}\Big\{\binom{b}{r}^{-1}\binom{b}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}c_S\binom{b}{s}U_S(G)\\
&-\binom{n}{r}^{-1}\binom{n}{r'}^{-1}\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}c_S\binom{n}{s}U_S(G)\Big\}\\
=&\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}\frac{c_Sr!r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'}}\Big(\frac{(b-r)!(b-r')!}{(b-1)!(b-s)!}-\frac{b(n-r)!(n-r')!}{n!(n-s)!}\Big)U_S(G)\\
=&\sum_{q=0}^{\min\{r,r'\}}\sum_{S\in\mathcal{S}_{R,R'}^{(q)}}\frac{c_Sr!r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'}}\Big(\frac{(b-r)!(b-r')!n!(n-s)!-b!(b-s)!(n-r)!(n-r')!]}{(b-1)!(b-s)!n!(n-s)!}\Big)U_S(G).
\end{aligned}
\tag{C.133}
$$

As a special case, we have

$$
\begin{aligned}
\mathrm{I}=&\mathrm{var}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)\Big]\\
=&\sum_{q=0}^{r}\sum_{S\in S_{R,R}^{(q)}}\frac{c_Sr!r!}{s!\rho_n^{2\mathfrak{r}}}\Big(\frac{(b-r)!(b-r)!n!(n-s)!-b!(b-s)!(n-r)!(n-r)!]}{(b-1)!(b-s)!n!(n-s)!}\Big)U_S(G).
\end{aligned}
\tag{C.134}
$$

For part II, based on Proposition 43, we have

$$
\mathrm{var}_*\Big[\sqrt{b}\rho_n^{-\mathfrak{r}}\sum g_{1,R}\left(\mathbb{V}_i\right)\Big]=b\rho_n^{-2\mathfrak{r}}\mathrm{var}_*\Big[\sum g_{1,R}\left(\mathbb{V}_i\right)\Big]\overset{(C.82)}{=}\rho_n^{-2\mathfrak{r}}b\frac{b(n-b)}{(n-1)}\mathrm{var}_*\Big[g_{1,R}\left(\mathbb{V}_1\right)\Big]
$$

$$
\begin{aligned}
&\overset{(C.81)}{=}\rho_n^{-2\mathfrak{r}}\frac{b^2(n-b)}{(n-1)}\Big[\frac{r!(n-r-1)!}{b(n-2)!}\Big]^2\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}c_S\frac{nq-r^2}{n^2}X_S(G)\\
=&\rho_n^{-2\mathfrak{r}}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}c_S\frac{b^2(n-b)}{(n-1)}\frac{r!(n-r-1)!r!(n-r-1)!}{b^2(n-2)!(n-2)!}\frac{(nq-r^2)}{n^2}\frac{n!}{(2r-q)!(n-2r+q)!}U_S(G)\\
=&\rho_n^{-2\mathfrak{r}}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}\frac{c_Sr!r!}{(2r-q)!}\Big[\frac{n-b}{n-1}\frac{n(n-1)}{n^2}\frac{(n-r-1)\cdots(n-2r+q+1)}{(n-2)\cdots(n-r)}(nq-r^2)\Big]U_S(G)\\
=&\rho_n^{-2\mathfrak{r}}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}\frac{c_Sr!r!}{(2r-q)!}\Big[\frac{(n-b)(n-r-1)\cdots(n-2r+q+1)(nq-r^2)}{n(n-2)\cdots(n-r)}\Big]U_S(G)\\
=&\rho_n^{-2\mathfrak{r}}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}\frac{c_Sr!r!}{s!}\Big[\frac{(n-b)(n-r-1)\cdots(n-s+1)(nq-r^2)}{n(n-2)\cdots(n-r)}\Big]U_S(G).
\end{aligned}
$$

Therefore, by combining part I and part II, we have

$$
E_*\big\{\Delta^2[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\big\}\overset{(C.132)}{=}\mathrm{I}-\mathrm{II}=\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}a_Sy_{n,S}z_{n,S},\tag{C.135}
$$

where

$$
\begin{aligned}
a_S&=\frac{c_Sr!r!}{s},\\
y_{n,S}&=\Big[\frac{(b-r)!(b-r)!n!(n-s)!-b!(b-s)!(n-r)!(n-r)!]}{(b-1)!(b-s)!n!(n-s)!}\\
&\quad-\frac{(n-b)(n-r-1)\cdots(n-s+1)(nq-r^2)}{n(n-2)\cdots(n-r)}\Big],\\
z_{n,S}&=\rho_n^{-2\mathfrak{r}}U_S(G).
\end{aligned}\tag{C.136}
$$

As S is given, a_s is a fixed quantity. Now we focus on the limiting behavior of $y_{n,S}z_{n,S}$ under different number of merged vertices q.

- When $q=0$, we have $\mathfrak{s}=2\mathfrak{r}$ and $s=2r$. We start with the first part of

$y_{n,S}$.

$$\begin{aligned}
&\frac{(b-r)!(b-r)!n!(n-s)! - b!(b-s)!(n-r)!(n-r)!}{(b-1)!(b-s)!n!(n-s)!} \\
=&\frac{[n\cdots(n-r+1)(b-r)\cdots(b-2r+1)] - [b\cdots(b-r+1)(n-r)\cdots(n-2r+1)]}{(b-1)\cdots(b-r+1)n\cdots(n-r+1)} \\
=&\frac{\mathrm{IV}-\mathrm{V}}{\mathrm{VI}}.
\end{aligned}$$

Now we study each component as follows:

$$\begin{aligned}
\mathrm{IV} =&b^r n^r - (r+r+1+\cdots r+r-1)b^{r-1}n^r - [1+2+\cdots+(r-1)]b^r n^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) + o(b^{r-1}n^{r-1}) \\
=&b^r n^r - \frac{(r+2r-1)r}{2}b^{r-1}n^r - \frac{(r-1)r}{2}b^r n^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) + o(b^{r-1}n^{r-1}). \\
\mathrm{V} =&b^r n^r - (r+r+1+\cdots r+r-1)n^{r-1}b^r - [1+2+\cdots+(r-1)]n^r b^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) + o(b^{r-1}n^{r-1}) \\
=&b^r n^r - \frac{(r+2r-1)r}{2}n^{r-1}b^r - \frac{(r-1)r}{2}n^r b^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) + o(b^{r-1}n^{r-1}).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\text{IV} - \text{V} =& b^r n^r - \frac{(r+2r-1)r}{2} b^{r-1} n^r - \frac{(r-1)r}{2} b^r n^{r-1} \\
&+ O(b^{r-2} n^r) + O(b^r n^{r-2}) + O(b^{r-1} n^{r-1}) \\
&- \Big[b^r n^r - \frac{(r+2r-1)r}{2} n^{r-1} b^r - \frac{(r-1)r}{2} n^r b^{r-1} \\
&+ O(b^{r-2} n^r) + O(b^r n^{r-2}) + O(b^{r-1} n^{r-1}) \Big] \\
=& \frac{-(r+2r-1)r}{2} n^{r-1} b^{r-1} (n-b) + \frac{(r-1)r}{2} n^{r-1} b^{r-1} (n-b) \\
&+ O(b^{r-2} n^r) + O(n^{r-2} b^r) + O(b^{r-1} n^{r-1}) \\
=& (-r^2) n^{r-1} b^{r-1} (n-b) + O(b^{r-2} n^r) + O(n^{r-2} b^r) + O(b^{r-1} n^{r-1}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{VI} =& b^{r-1} n^r + o(b^{r-1} n^r). \\
\frac{\text{IV} - \text{V}}{\text{VI}} =& (-r^2) \frac{n-b}{n} + o(1).
\end{aligned} \tag{C.137}$$

For the second part of $y_{n,S}$, we have

$$\begin{aligned}
& \Big[\frac{(n-b)(n-r-1) \cdots (n-s+1)(nq-r^2)}{n(n-2) \cdots (n-r)} \Big] \\
=& \Big[\frac{(n-b)(n-r-1)(n-r-2) \cdots (n-2r+1)(0-r^2)}{n(n-2)(n-3) \cdots (n-r)} \Big] \\
=& \Big[(1 - \frac{b}{n})(1 - \frac{r+1}{n-2})(1 - \frac{r+1}{n-3}) \cdots (1 - \frac{r+1}{n-r})(-r^2) \Big] \\
=& -r^2 \Big[\frac{n-b}{n} + o(1) \Big].
\end{aligned}$$

Therefore, we have

$$y_{n,S} = (-r^2) \frac{n-b}{n} + o(1) - (-r^2) \frac{n-b}{n} + o(1) = o(1).$$

The difficulty in studying $z_{n,S}$ is that we have no information for $\{G^{(n)}\}$.

To tackle this, we assume that $G \sim \mathbb{G}_{\ltimes}$. Lemma 53 implies that under Assumption 2,

$$pr\Big\{ \lim_{n\to\infty} \rho_n^{-\mathfrak{s}} U_S(\mathbb{G}_n) = \frac{s!}{|\mathrm{Aut}(S)|} P_w(S) \Big\} = 1.$$

Thus, we assert that $y_{n,S} z_{n,S} \to 0$ with probability one because $y_{n,S} = o(1)$ and $z_{n,S}$ is a realization of $\rho_n^{-\mathfrak{s}} U_S(\mathbb{G}_n)$.

- When $q = 1$, we have $\mathfrak{s} = 2\mathfrak{r}$ and $s = 2r - 1$. As before, we study the first part of $y_{n,S}$.

$$\begin{aligned}
&\frac{[(b-r)!(b-r)!n!(n-s)! - b!(b-s)!(n-r)!(n-r)!]}{(b-1)!(b-s)!n!(n-s)!} \\
=&\frac{[(b-r)!(b-r)!n!(n-2r+1)! - b!(b-2r+1)!(n-r)!(n-r)!]}{(b-1)!(b-2r+1)!n!(n-2r+1)!} = 1 - \frac{b}{n} + o(1).
\end{aligned}$$

For the second part of $y_{n,S}$, we have

$$\begin{aligned}
&\Big[\frac{(n-b)(n-r-1)\cdots(n-s+1)(nq-r^2)}{n(n-2)\cdots(n-r)}\Big] \\
=&\Big[\frac{(n-b)(n-r-1)(n-r-2)\cdots(n-2r+2)(n-r^2)}{n(n-2)(n-3)\cdots(n-r)}\Big] \\
=&1 - \frac{b}{n} + o(1).
\end{aligned}$$

Therefore, when $q = 1$, we have $y_{n,S} = o(1)$. Thus, we also have $y_{n,S} z_{n,S} \to 0$ with probability one.

- When $q > 1$, the first part of $y_{n,S}$ is

$$\frac{[(b-r)!(b-r)!n!(n-s)! - b!(b-s)!(n-r)!(n-r)!]}{(b-1)!(b-s)!n!(n-s)!} = O\Big[\frac{1}{b^{(q-1)}}\Big].$$

The second part of $y_{n,S}$ is

$$\begin{aligned}&\frac{(n-b)(n-r-1)\cdots(n-s+1)(nq-r^2)}{n(n-2)\cdots(n-r)}\\=&\frac{(n-b)(n-r-1)(n-r-2)\cdots(n-2r+q+1)(nq-r^2)}{n(n-2)(n-3)\cdots(n-r)}\\=&O\big[\frac{(n-b)}{n^q}\big].\end{aligned}$$

Therefore, $y_{n,s}=o(1)$. We have $y_{n,S}z_{n,S}\to 0$ for every $q\geq 2$, .

Finally, because r is a constant and $\mathcal{S}_{R,R}$ is a fixed set given R. For any random network sequence $\{G^{(n)}\}$, with probability one,

$$\lim_{n\to\infty} E_*\big\{\Delta^2[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)]\big\}=\lim_{n\to\infty}\sum_{q=0}^{r}\sum_{S\in\mathcal{S}_{R,R}^{(q)}}a_S y_{n,s}z_{n,s}=0.$$

(a).iii Now we want to show (C.89), which is related to non-degeneration, and is termed as the non-lattice assumption in Zhang and Xia (2022). From lemma 52, under Assumptions 2 and 1,

$$pr\Big\{\lim_{b\to\infty}\rho_n^{-2\mathfrak{r}}\mathrm{var}_*\big[\sqrt{b}U_R(\mathbb{G}_b^*)\big]=\big(1-c_2\big)\lim_{b\to\infty}\rho_b^{-2\mathfrak{r}}\mathrm{var}\big[\sqrt{b}U_R(\mathbb{G}_b)\big]\Big\}=1.$$

Since $c_2<1$ is a constant, (C.89) holds by Assumption 3.

(a).iv Next, we want to show (C.90), which is a Lindeberg-Feller typed condition (Bloznelis and Götze, 2001). To verify this, We want to show that

$$b\rho_n^{-2\mathfrak{r}}g_{1,R}^2(\mathbb{V}_1)=o(1).$$

Let's begin by considering the following expression:

$$b\rho_n^{-2\mathfrak{r}} g_{1,R}^2(\mathbb{V}_1) \overset{(C.80)}{=} \frac{\rho_n^{-2\mathfrak{r}}}{b}\left\{\frac{r(n-1)}{(n-r)}\binom{n-1}{r-1}^{-1}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})} X_R(\mathscr{G}) - \frac{(n-1)r}{(n-r)}U_R(G)\right\}^2. \tag{C.138}$$

Recall that K_r denotes the complete graph of r vertices. Clearly, for any $\mathscr{G} \in \mathcal{S}(\mathbb{G}_r^{V_1*})$,

$$X_R(\mathscr{G}) \leqslant X_R(K_r).$$

Equation (2.7) in Bhattacharya et al. (2022) shows that $X_R(K_r) = r!/|Aut(R)|$. Therefore,

$$\begin{aligned}
&\frac{r!(n-r-1)!}{(n-2)!}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})} X_R(\mathscr{G}) = \frac{r(n-r)}{(n-1)}\binom{n-1}{r-1}^{-1}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})} X_R(\mathscr{G}) \\
\leqslant &\frac{r(n-r)}{(n-1)}\binom{n-1}{r-1}^{-1}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})}\left(\frac{r!}{|\mathrm{Aut}(R)|}\right)\mathbb{1}_{\{R\subset\mathscr{G}\}} \\
= &\frac{r(n-r)}{(n-1)}\left(\frac{r!}{|\mathrm{Aut}(R)|}\right)\binom{n-1}{r-1}^{-1}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})}\mathbb{1}_{\{R\subset\mathscr{G}\}} \leqslant \frac{r(n-r)}{(n-1)}\left(\frac{r!}{|\mathrm{Aut}(R)|}\right).
\end{aligned} \tag{C.139}$$

By (C.138),

$$\begin{aligned}
b\rho_n^{-2\mathfrak{r}} g_{1,R}^2(\mathbb{V}_1) &\overset{(C.138)}{\leqslant} \frac{\rho_n^{-2\mathfrak{r}}}{b}\left\{\left[\frac{r(n-1)}{(n-r)}\binom{n-1}{r-1}^{-1}\sum_{\mathscr{G}\in\mathcal{S}(\mathbb{G}_r^{V_1*})} X_R(\mathscr{G})\right]^2 + \left[\frac{(n-1)r}{(n-r)}U_R(G)\right]^2\right\} \\
&\overset{(C.139)}{\leqslant} \frac{\rho_n^{-2\mathfrak{r}}}{b}\left\{\left[\frac{r(n-r)}{(n-1)}\left(\frac{r!}{|\mathrm{Aut}(R)|}\right)\right]^2 + \left[\frac{(n-1)r}{(n-r)}U_R(G)\right]^2\right\}.
\end{aligned}$$

Given that both r and$|\mathrm{Aut}(R)|$ are constants, and considering $U_R(G) \leqslant 1$, it follows that if $\rho_n^{-2\mathfrak{r}}/b \to 0$, then for any given $\epsilon > 0$, there exists a $K > 0$ such that when $k > K$,

$$\mathbb{1}_{\left\{b\rho_n^{-2\mathfrak{r}} g_{1,R}^2(\mathbb{V}_1) > \epsilon\right\}} = 0. \tag{C.140}$$

Consequently, under these conditions, it can be deduced that

$$\lim_{n\to\infty} bE_*\left[b\rho_n^{-2\mathfrak{r}} g_{1,R}^2(\mathbb{V}_1) \mathbb{1}_{\{b\rho_n^{-2\mathfrak{r}} g_{1,R}^2(\mathbb{V}_1)>\epsilon\}}\right] = 0.$$

(a).v Finally, we arrive at (C.91). To prove this, we first recall (42), which implies $E_*[U_R(\mathbb{G}_b^*)] = U_R(G)$. In addition, since (C.88), (C.89), and (C.90) are satisfied, lemma 49 implies that

$$\frac{\sqrt{b_n}\left[\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*) - \rho_n^{-\mathfrak{r}} U_R(G^{(n)})\right]}{\mathrm{var}_*(\sqrt{b_n}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_{b_n}^*))},$$

is asymptotically standard normal.

Now we continue to prove part (b). Let $R_1, \cdots, R_m$ be m motifs. Consider the following linear combination

$$\Theta_{R_1,\ldots,R_m}^{(a_1,\cdots,a_m)} = a_1\sqrt{b}\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*) + \cdots + a_m\sqrt{b}\rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*),$$

where $a_1, \cdots, a_m$ are constants. For simplicity, we write $\Theta = \Theta_{R_1,\ldots,R_m}^{(a_1,\cdots,a_m)}$. The statistic Θ is a symmetric finite population statistic with the following Hoeffding's decomposition

$$\Theta = E_*(\Theta) + \sum_{1\leqslant i\leqslant b} g_{1,\Theta}(\mathbb{V}_i) + \Delta(\Theta).$$

Now we want to show that every linear combination Θ is asymptotically normal. Following 49, we need to verify the following conditions.

$$\lim_{b\to\infty} E_*\Delta^2(\Theta) = 0, \text{ where } \Delta(\Theta) = \Theta - E_*(\Theta) - \sum_{1\leqslant i\leqslant b} g_{1,\Theta}(\mathbb{V}_i). \tag{C.141}$$

$$0 < c_1 \leqslant \lim_{b\to\infty} \mathrm{var}_*(\Theta) \leqslant c_2 < \infty, \qquad \text{for some } c_1, c_2 > 0. \tag{C.142}$$

$$\text{For every } \epsilon > 0 \qquad \lim_{b\to\infty} bE_*\left[g^2_{1,\Theta}\left(\mathbb{V}_1\right) \mathbb{1}_{\left\{g^2_{1,\Theta}(\mathbb{V}_1)>\epsilon\right\}}\right] = 0. \tag{C.143}$$

(b).i To show (C.141), we first consider a pair of motifs R and R' associate with two constants α and β. In this case,

$$\Theta = \Theta^{(\alpha,\beta)}_{R,R'} = \alpha\sqrt{b}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}^*_b) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}^*_b). \tag{C.144}$$

The Hoeffding's decomposition of $\Theta^{(\alpha,\beta)}_{R,R'}$ can be expressed as follows

$$\Theta^{(\alpha,\beta)}_{R,R'} = E_*\big(\Theta^{(\alpha,\beta)}_{R,R'}\big) + \sum_{1\leqslant i\leqslant b} g_{1,\Theta^{(\alpha,\beta)}_{R,R'}}\left(\mathbb{V}_i\right) + \sum_{1\leqslant i<j\leqslant b} g_{2,\Theta^{(\alpha,\beta)}_{R,R'}}\left(\mathbb{V}_i,\mathbb{V}_j\right) + \cdots. \tag{C.145}$$

The Proposition 43 implies

$$E_*(\Theta^{(\alpha,\beta)}_{R,R'}) = \alpha\sqrt{b}\rho_n^{-\mathfrak{r}} E_*\big[U_R(\mathbb{G}^*_b)\big] + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'} E_*\big[U_{R'}(\mathbb{G}^*_b)\big], \tag{C.146}$$

and

$$\sum_{1\leqslant i\leqslant b} g_{1,\Theta^{(\alpha,\beta)}_{R,R'}}\left(\mathbb{V}_i\right) \overset{(C.84)}{=} \alpha\sqrt{b}\rho_n^{-\mathfrak{r}} \sum_{1\leqslant i\leqslant b} g_{1,R}\left(\mathbb{V}_i\right) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'} \sum_{1\leqslant i\leqslant b} g_{1,R'}\left(\mathbb{V}_i\right), \tag{C.147}$$

where $g_{1,R}(\mathbb{V}_i)$ and $g_{1,R'}(\mathbb{V}_i)$ are defined in (C.80). Consequently, we have

$$\begin{aligned}
\Delta[\Theta_{R,R'}^{(\alpha,\beta)}] &= \Theta_{R,R'}^{(\alpha,\beta)} - E_*\big[\Theta_{R,R'}^{(\alpha,\beta)}\big] - \sum_{1\leqslant i\leqslant b} g_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_i) \\
&\overset{(C.144)}{=} \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*) - E_*\big[\Theta_{R,R'}^{(\alpha,\beta)}\big] - \sum_{1\leqslant i\leqslant b} g_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_i) \\
&\overset{(C.146)}{=} \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*) \\
&\quad - \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}E_*\big[U_R(\mathbb{G}_b^*)\big] - \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}E_*\big[U_{R'}(\mathbb{G}_b^*)\big] - \sum_{1\leqslant i\leqslant b} g_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_i) \\
&\overset{(C.147)}{=} \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*) \\
&\quad - \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}E_*\big[U_R(\mathbb{G}_b^*)\big] - \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}E_*\big[U_{R'}(\mathbb{G}_b^*)\big] \\
&\quad - \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i) - \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}\sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i) \\
&= \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*) - \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}E_*\big[U_R(\mathbb{G}_b^*)\big] - \alpha\sqrt{b}\rho_n^{-\mathfrak{r}}\sum_{1\leqslant i\leqslant b} g_{1,R}(\mathbb{V}_i) \\
&\quad + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*) - \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}E_*\big[U_{R'}(\mathbb{G}_b^*)\big] - \beta\sqrt{b}\rho_n^{-\mathfrak{r}'}\sum_{1\leqslant i\leqslant b} g_{1,R'}(\mathbb{V}_i) \\
&\overset{(C.130)}{=} \alpha\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)) + \beta\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*)).
\end{aligned} \tag{C.148}$$

Consequently,

$$E_*\Big[\Delta^2(\Theta_{R,R'}^{(\alpha,\beta)})\Big] \overset{(C.148)}{\leq} 2E_*\Big[\alpha\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*))\Big]^2 + 2E_*\Big[\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*))\Big]^2.$$

From (C.88), under Assumptions 3 2 and 1, we have

$$\begin{aligned}
\lim_{b\to\infty} E_*\Big[\Delta^2(\Theta_{R,R'}^{(\alpha,\beta)})\Big] &\leqslant \lim_{b\to\infty} 2E_*\Big[\alpha\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*))\Big]^2 \\
&\quad + \lim_{b\to\infty} 2E_*\Big[\beta\Delta(\sqrt{b}\rho_n^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b^*))\Big]^2 = 0.
\end{aligned}$$

For m motifs, due to the linearity of expectation, we have

$$E_*\big[\Theta\big] = a_1 E_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*)\big] + \cdots + a_m E_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*)\big]. \tag{C.149}$$

On the other hand, we have

$$\sum_{1\leqslant i\leqslant b} g_{1,\Theta}\left(\mathbb{V}_i\right) \overset{(C.147)}{=} a_1\sqrt{b}\rho_n^{-\mathfrak{r}_1} \sum_{1\leqslant i\leqslant b} g_{1,R_1}\left(\mathbb{V}_i\right) + \cdots + a_m\sqrt{b}\rho_n^{-\mathfrak{r}_m} \sum_{1\leqslant i\leqslant b} g_{1,R_m}\left(\mathbb{V}_i\right). \tag{C.150}$$

Similar to the derivation of (C.148), we have:

$$\Delta(\Theta) = a_1\Delta[\sqrt{b}\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*)] + \cdots a_m\Delta[\sqrt{b}\rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*)].$$

Thus, the accuracy of approximation of the linear part could be bounded as:

$$\begin{aligned}\lim_{b\to\infty} E_*\Delta^2(\Theta) \leqslant m\Big\{ a_1^2 \lim_{b\to\infty} E_*\Delta^2[\sqrt{b}\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*)] + \\ \cdots + a_m^2 \lim_{b\to\infty} E_*\Delta^2[\sqrt{b}\rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*)]\Big\} \overset{(C.88)}{=} 0.\end{aligned} \tag{C.151}$$

(b).ii Now we prove (C.142). We also first consider a pair of motifs R, R' as follows.

$$\begin{aligned}\mathrm{var}_*\big[\Theta_{R,R'}^{(\alpha,\beta)}\big] =& \mathrm{var}_*\big[\alpha\sqrt{b}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_b^*) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b^*)\big] \\ =& \alpha^2 \mathrm{var}_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_b^*)\big] + \beta^2 \mathrm{var}_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b^*)\big] \\ &+ 2\alpha\beta\,\mathrm{Cov}_*\left[\sqrt{b}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_b^*), \sqrt{b}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b^*)\right].\end{aligned} \tag{C.152}$$

By lemma 52, under Assumptions 2 and 1,

$$pr\left\{ \lim_{b\to\infty} \mathrm{var}_*\big[\Theta_{R,R'}^{(\alpha,\beta)}\big] = (1-c_2) \lim_{b\to\infty} \mathrm{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b)\Big] \right\} = 1. \tag{C.153}$$

Therefore, for any sequence of networks, condition in (C.142) holds with probability one if

$$0 < c_1 \leqslant \lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b)\Big] \leqslant c_2 < \infty. \quad \text{(C.154)}$$

Now we define:

$$\begin{aligned}
\tilde{\sigma}_R^2 &= \lim_{b\to\infty} \text{var}\left[\sqrt{b}\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b)\right], \\
\tilde{\sigma}_{R'}^2 &= \lim_{b\to\infty} \text{var}\left[\sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b)\right], \\
\tilde{\sigma}_{R,R'} &= \lim_{b\to\infty} \text{cov}\left[\sqrt{b}\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b), \sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b)\right],
\end{aligned}$$

where $\tilde{\sigma}_R^2$, $\tilde{\sigma}_{R'}^2$, and $\tilde{\sigma}_{R,R'}^2$ are all constants as derived in Proposition 40. Consequently,

$$\begin{aligned}
\lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b)\Big] &= \tilde{\sigma}_R^2\alpha^2 + 2\tilde{\sigma}_{R,R'}^2\alpha\beta + \tilde{\sigma}_{R'}^2\beta^2 \\
&= \beta^2\left[\tilde{\sigma}_R^2\frac{\alpha^2}{\beta^2} + 2\tilde{\sigma}_{R,R'}\frac{\alpha}{\beta} + \tilde{\sigma}_{R'}^2\right].
\end{aligned}$$

Lemma 51 implies that $(\sqrt{b}\rho_b^{-\mathfrak{r}}(U_R(\mathbb{G}_b), \sqrt{b}\rho_b^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b))$ converges in distribution to a bivariate Gaussian distribution. The properties of the probability density function of the bivariate Gaussian distribution implies that:

$$\frac{\tilde{\sigma}_{R,R'}}{\tilde{\sigma}_R\tilde{\sigma}_{R'}} < 1.$$

Consequently, $\beta^2\left[\tilde{\sigma}_R^2\frac{\alpha^2}{\beta^2} + 2\tilde{\sigma}_{R,R'}\frac{\alpha}{\beta} + \tilde{\sigma}_{R'}^2\right]$ has no real root due to $\tilde{\sigma}_{R,R'} - \tilde{\sigma}_R\tilde{\sigma}_{R'} <$

0. This implies that

$$\lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b)\Big] > 0.$$

Let

$$c_1 = \frac{1}{2}\lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b)\Big],$$

then

$$\lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b)\Big] > c_1 > 0.$$

On the other hand, we set the upper bound $c_2 = \max\{4\alpha^2\tilde{\sigma}_R^2, 4\beta^2\tilde{\sigma}_{R'}^2\}$ based on

$$\begin{aligned}
&\lim_{b\to\infty} \text{var}\Big[\alpha\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b) + \beta\sqrt{b}\rho_b^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b)\Big] \\
\leqslant &4\max\Big\{\lim_{b\to\infty}\alpha^2\text{var}\Big[\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)\Big], \lim_{b\to\infty}\beta^2\text{var}\Big[\sqrt{b}\rho_b^{-\mathfrak{r}'}U_{R'}(\mathbb{G}_b)\Big]\Big\}.
\end{aligned}$$

Thus, the condition in (C.142) holds.

For m motifs, we first apply Proposition 52 to have

$$pr\Big\{\lim_{b\to\infty}\text{var}_*[\Theta] = \lim_{b\to\infty}\text{var}\Big[a_1\sqrt{b}\rho_b^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b) + \cdots + a_m\sqrt{b}\rho_b^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b)\Big]\Big\} = 1. \tag{C.155}$$

From Lemma 51, we have

$$\Big(a_1\sqrt{b}\rho_b^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b), \cdots, a_m\sqrt{b}\rho_b^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b)\Big) \xrightarrow{d} \mathcal{N}\Big(0, \Sigma(\mathbf{R})\Big).$$

Since the covariance matrix $\Sigma(\mathbf{R})$ of multivariate Gaussian has to be a positive

definite matrix, let q be a $1 \times m$ vector with all elements equal to one, we have:

$$\lim_{b\to\infty} \operatorname{var}\Big[a_1\sqrt{b}\rho_b^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b) + \cdots + a_m\sqrt{b}\rho_b^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b)\Big] = q\Sigma(\mathbf{R})q^\top > 0.$$

Consequently, the non-lattice condition in Equation (C.142) holds by setting

$$c_1 = \frac{1}{2}\lim_{b\to\infty} \operatorname{var}\Big[a_1\sqrt{b}\rho_b^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b) + \cdots + a_m\sqrt{b}\rho_b^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b)\Big],$$

$$c_2 = \max\{ma_1^2\tilde{\sigma}_{R_1}^2, \cdots, ma_m^2\tilde{\sigma}_{R_m}^2\}.$$

(b).iii To show (C.143), first considering two motifs R and R',

$$g_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_1) \overset{(C.150)}{=} \alpha\sqrt{b}\rho_n^{-\mathfrak{r}} g_{1,R}(\mathbb{V}_1) + \beta\sqrt{b}\rho_n^{-\mathfrak{r}'} g_{1,R'}(\mathbb{V}_1).$$

Consequently,

$$\begin{aligned}\lim_{b\to\infty} g^2_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_1) &\leqslant \lim_{b\to\infty} \alpha^2 b\rho_n^{-2\mathfrak{r}} g^2_{1,R}(\mathbb{V}_1) + \lim_{b\to\infty} \beta^2 b\rho_n^{-2\mathfrak{r}'} g^2_{1,R'}(\mathbb{V}_1) \\ &\overset{(C.140)}{=} 0.\end{aligned} \tag{C.156}$$

Therefore, for any $\epsilon > 0$, there exists a sufficiently large N and a sufficiently large B, such that when $n > N$ and $b > B$,

$$\mathbb{1}_{\left\{g^2_{1,\Theta_{R,R'}^{(\alpha,\beta)}} > \epsilon\right\}} = 0.$$

Consequently,

$$\lim_{b\to\infty} bE_*\left[g^2_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_1)\right]\mathbb{1}_{\left\{g^2_{1,\Theta_{R,R'}^{(\alpha,\beta)}}(\mathbb{V}_1) > \epsilon\right\}} = 0. \tag{C.157}$$

For m motifs, condition in Equation (C.143) also holds as

$$\lim_{b\to\infty} g^2_{1,\Theta}\left(\mathbb{V}_1\right) \overset{(C.150)}{\leqslant} \lim_{b\to\infty} a_1^2 b\rho_n^{-2\mathfrak{r}} g^2_{1,R}\left(\mathbb{V}_1\right) + \cdots + \lim_{b\to\infty} a_m^2 b\rho_n^{-2\mathfrak{r}'} g^2_{1,R'}\left(\mathbb{V}_1\right) \overset{(C.140)}{=} 0. \tag{C.158}$$

Therefore, Lemma 49 implies that every linear combination:

$$a_1\sqrt{b}\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*) + \cdots + a_m\sqrt{b}\rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*), \tag{C.159}$$

is asymptotically normal, which implies

$$\sqrt{b}\Big\{\left[\rho_n^{-\mathfrak{r}_1} U_{R_1}(\mathbb{G}_b^*), \cdots, \rho_n^{-\mathfrak{r}_m} U_{R_m}(\mathbb{G}_b^*)\right] - \left[\rho_n^{-\mathfrak{r}_1} U_{R_1}(G), \cdots, \rho_n^{-\mathfrak{r}_m} U_{R_m}(G)\right]\Big\}$$
$$\to \mathcal{N}\left[0, \Sigma(*\mathbf{R})\right] \text{ in distribution.}$$

□

C.6.2 Proof of Lemma 52

Proof. The following result was developed in (C.133).

$$\begin{aligned}
&\operatorname{cov}_*\left[\sqrt{b}\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_b^*), \sqrt{b}\rho_n^{-\mathfrak{r}'} U_{R'}(\mathbb{G}_b^*)\right] \\
&\overset{(C.133)}{=} \sum_{q=0}^{\min\{r,r'\}} \sum_{S\in\mathcal{S}^{(q)}_{R,R'}} \frac{c_S r! r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'}} \left[\frac{(b-r)!(b-r')!n!(n-s)! - b!(b-s)!(n-r)!(n-r')!}{(b-1)!(b-s)!n!(n-s)!}\right] U_S(G).
\end{aligned}$$

Notice c_S, $r!$, $r'!$ are fixed quantities. We are now going to study the limiting behaviors for other quantities under different number of merged vertices q.

- When $q = 0$, we have $\mathfrak{s} = \mathfrak{r} + \mathfrak{r}'$ and $s = r + r'$. Thus,

$$\begin{aligned}
&\frac{(b-r)!(b-r')!n!(n-s)! - b!(b-s)!(n-r)!(n-r')!}{(b-1)!(b-s)!n!(n-s)!} \\
=&\frac{(b-r)!(b-r')!n!(n-r-r')! - b!(b-r-r')!(n-r)!(n-r')!}{(b-1)!(b-r-r')!n!(n-r-r')!} \\
=&\frac{n\cdots(n-r+1)(b-r')\cdots(b-r-r'+1) - b(b-1)\cdots(b-r+1)(n-r')\cdots(n-r-r'+1)}{(b-1)\cdots(b-r+1)n(n-1)\cdots(n-r+1)} \\
=&\frac{\mathrm{IV}-\mathrm{V}}{\mathrm{VI}}.
\end{aligned}$$

Similar to (C.137), Now we study each component as follows:

$$\begin{aligned}
\mathrm{IV} =& b^r n^r - (r' + r' + 1 + \cdots r' + r - 1)b^{r-1}n^r - [1 + 2 + \cdots + (r-1)]b^r n^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) \\
=& b^r n^r - \frac{(r+2r'-1)r}{2}b^{r-1}n^r - \frac{(r-1)r}{2}b^r n^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}).
\end{aligned}$$

$$\begin{aligned}
\mathrm{V} =& b^r n^r - (r + r + 1 + \cdots r + r - 1)n^{r-1}b^r - [1 + 2 + \cdots + (r-1)]n^r b^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) \\
=& b^r n^r - \frac{(r+2r'-1)r}{2}n^{r-1}b^r - \frac{(r-1)r}{2}n^r b^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathrm{IV}-\mathrm{V} =& b^r n^r - \frac{(r+2r-1)r}{2} b^{r-1} n^r - \frac{(r-1)r}{2} b^r n^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) \\
&- \Big[b^r n^r - \frac{(r+2r'-1)r}{2} n^{r-1} b^r - \frac{(r-1)r}{2} n^r b^{r-1} \\
&+ O(b^{r-2}n^r) + O(b^r n^{r-2}) + O(b^{r-1}n^{r-1}) \Big] \\
=& \frac{-(r+2r'-1)r}{2} n^{r-1} b^{r-1} (n-b) + \frac{(r-1)r}{2} n^{r-1} b^{r-1} (n-b) \\
&+ O(b^{r-2}n^r) + O(n^{r-2}b^r) + O(b^{r-1}n^{r-1}) \\
=& (-rr') n^{r-1} b^{r-1} (n-b) + O(b^{r-2}n^r) + O(n^{r-2}b^r) + O(b^{r-1}n^{r-1}).
\end{aligned}$$

Thus, we have

$$\frac{\mathrm{IV}-\mathrm{V}}{\mathrm{VI}} = (-rr')\frac{n-b}{n} + o(1),$$

and by Assumption 1,

$$\lim_{b\to\infty} \frac{\mathrm{IV}-\mathrm{V}}{\mathrm{VI}} = (-rr')(1-c_2). \tag{C.160}$$

On the other hand, since $\rho_n^{\mathfrak{r}+\mathfrak{r}'} = \rho_n^{\mathfrak{s}} \leqslant \rho_n^{2\mathfrak{r}_1}$, by lemma 53.

$$\begin{aligned}
\lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(0)}_{R,R'}} \frac{c_S r!r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'}} \Big(\frac{\mathrm{IV}-\mathrm{V}}{\mathrm{VI}}\Big) U_S(G) &\overset{(C.137)}{=} (1-c_2) \lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(0)}_{R,R'}} \frac{-c_S r!r'!rr'}{s!} \rho_n^{-\mathfrak{s}} U_S(G) \\
&\overset{(C.129)}{=} (1-c_2) \sum_{S\in\mathcal{S}^{(0)}_{R,R'}} \frac{-c_S r!r'!rr'}{|\mathrm{Aut}(S)|} P_w(S).
\end{aligned}$$

- when $q = 1$, we have $\mathfrak{s} = \mathfrak{r} + \mathfrak{r}'$ and $s = r + r' - 1$. Thus,

$$\begin{aligned}
&\frac{(b-r)!(b-r')!n!(n-s)! - b!(b-s)!(n-r)!(n-r')!}{(b-1)!(b-s)!n!(n-s)!} \\
=&\frac{(b-r)!(b-r')!n!(n-r-r'+1)! - b!(b-r-r'+1)!(n-r)!(n-r')!}{(b-1)!(b-r-r'+1)!n!(n-r-r'+1)!} \\
=&1 - \frac{b}{n} + o(1).
\end{aligned} \tag{C.161}$$

  Since $\rho_n^{\mathfrak{r}+\mathfrak{r}'} = \rho_n^{\mathfrak{s}} \leqslant \rho_n^{2\mathfrak{r}_1}$, by Assumption 1,

$$\begin{aligned}
&\lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(1)}_{R,R'}} \frac{c_S r!r'!}{s!\rho_n^{\mathfrak{s}}} \Big[\frac{(b-r)!(b-r')!n!(n-s)! - b!(b-s)!(n-r)!(n-r')!}{(b-1)!(b-s)!n!(n-s)!}\Big] U_S(G) \\
\overset{(C.161)}{=}&(1-c_2)\lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(1)}_{R,R'}} \frac{c_S r!r'!}{s!\rho_n^{\mathfrak{s}}} U_S(G) \overset{(C.129)}{=} (1-c_2) \sum_{S\in\mathcal{S}^{(1)}_{R,R'}} c_S \frac{r!r'!}{|\mathrm{Aut}(S)|} P_w(S).
\end{aligned}$$

- when $q > 1$, we have

$$\frac{[(b-r)!(b-r)!n!(n-s)! - b!(b-s)!(n-r)!(n-r)!]}{(b-1)!(b-s)!n!(n-s)!} = O(\frac{1}{b^{(q-1)}}).$$

  In addition, since $\rho_n^{(\mathfrak{s}-\mathfrak{r}-\mathfrak{r}')}$ is at most $O(\rho_n^{-(q-1)q/2})$ and $\rho_n^{q/2} b \to \infty$, we have

$$\begin{aligned}
&\lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(q)}_{R,R'}} \frac{c_S r!r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'}} \Big[\frac{(b-r)!(b-r')!n!(n-s)! - b!(b-s)!(n-r)!(n-r')!}{(b-1)!(b-s)!n!(n-s)!}\Big] U_S(G) \\
=&\lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(q)}_{R,R'}} \frac{c_S r!r'!}{s!\rho_n^{\mathfrak{r}+\mathfrak{r}'-\mathfrak{s}}} O(\frac{1}{b^{(q-1)}}) \rho_n^{-\mathfrak{s}} U_S(G) = \lim_{b\to\infty} \sum_{S\in\mathcal{S}^{(q)}_{R,R'}} \frac{c_S r!r'!}{s!} O(\frac{1}{b\rho_n^{q/2}})^{(q-1)} \rho_n^{-\mathfrak{s}} U_S(G) \overset{(C.129)}{=} 0.
\end{aligned}$$

Therefore,

$$pr\Big\{ \lim_{b\to\infty} \rho_n^{-(\mathfrak{r}+\mathfrak{r}')} \mathrm{cov}_*\big[\sqrt{b}U_R(\mathbb{G}_b^*), \sqrt{b}U_{R'}(\mathbb{G}_b^*)\big]$$

$$= \big(1-c_2\big)\Big[\sum_{S\in\mathcal{S}_{R,R'}^{(1)}} c_S \frac{r!r'!}{|\mathrm{Aut}(S)|} P_w(S) - \sum_{S\in\mathcal{S}_{R,R'}^{(0)}} \frac{c_S r!r'!rr'}{|\mathrm{Aut}(S)|} P_w(S)\Big]$$

$$= \big(1-c_2\big) \lim_{b\to\infty} \rho_b^{-(\mathfrak{r}+\mathfrak{r}')} \mathrm{cov}\big[\sqrt{b}U_R(\mathbb{G}_b), \sqrt{b}U_{R'}(\mathbb{G}_b)\big]\Big\} = 1.$$

In particular,

$$pr\Big\{ \lim_{b\to\infty} \mathrm{var}_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_b^*)\big] = \big(1-c_2\big) \lim_{b\to\infty} \mathrm{var}\big[\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)\big]\Big\} = 1. \quad \text{(C.162)}$$

□

C.6.3 Proof of Lemma 53

Proof. Proposition 53 adopts the technique used in Theorem 2.5 of Lovász and Szegedy (2006) and Theorem 4.4.5 in Zhao (2023). We start by generating a sequence of graphs $\{\mathbb{G}^{(i)}\}_{i=1}^n$. The first graph $\mathbb{G}^{(1)}$ is only a vertex v_1 with latent position $\xi_1 \sim \mathrm{Unif}[0,1]$, the second graph $\mathbb{G}^{(2)}$ contains two vertex v_1, v_2 with latent positions ξ_1 and $\xi_2 \sim \mathrm{Unif}[0,1]$. The probability of an edge between v_1, v_2 is $h_n(\xi_1, \xi_2)$. In this way, $\{\mathbb{G}^{(i)}\}_{i=1}^n$ is generated by incrementally adding one vertex and the corresponding edges at a time, and previously selected vertices and edges are not revisited. Furthermore, we have $\mathbb{G}^{(i)} \sim \mathbb{G}_i^{h_n}$.

Let ϕ: $V(R) \to V(G)$ be an injective mapping, and let A_ϕ denote the event that

ϕ is a homomorphism from R to graph $\mathbb{G}^{(n)}$. We define the sequence $\{A_i\}_{i=1}^{n}$ as

$$A_i = (n)_r^{-1} \sum_{\phi} pr(A_\phi \mid G^{(i)}). \tag{C.163}$$

Based on the definition of A_l, we have

$$A_n = (n)_r^{-1} \sum_{\phi} pr(A_\phi \mid \mathbb{G}^{(n)}) = t(R, \mathbb{G}^{(n)}), \tag{C.164}$$

$$A_0 = (n)_r^{-1} \sum_{\phi} pr(A_\phi) = \int_{[0,1]^r} \prod_{(v_i,v_j)\in\mathscr{E}(R)} h_n(\xi_i, \xi_j) \prod_{v_i \in V(R)} d\xi_i \overset{(C.65)}{=} P_{h_n}(R). \tag{C.165}$$

Lovász and Szegedy (2006) showed that $\{A_n\}$ is a martingale and $|A_i - A_{i-1}| \leqslant r/n$ in their Theorem 2.5. Then by invoking Azuma's inequality, they showed that, for every $\delta > 0$,

$$pr\Big(\big|t(R, \mathbb{G}_n) - P_{h_n}(R)\big| > \delta\Big) \leqslant 2\exp\left(-\delta^2 n/2r^2\right).$$

On the other hand, the Proposition 1 of Amini et al. (2012) implies that

$$X_R(G) = \mathrm{inj}(R, G)/|\mathrm{Aut}(R)|.$$

Therefore, we have

$$\rho_n^{-\mathfrak{r}} U_R(G) = \rho_n^{-\mathfrak{r}} \binom{n}{r}^{-1} X_R(G) = \rho_n^{-\mathfrak{r}} \binom{n}{r}^{-1} \frac{\mathrm{inj}(R, G)}{|\mathrm{Aut}(R)|} = \rho_n^{-\mathfrak{r}} \frac{r! t(R, G)}{|\mathrm{Aut}(R)|}. \tag{C.166}$$

Consequently, we have

$$pr\left(\left|\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n) - \frac{\rho_n^{-\mathfrak{r}}r!}{|\mathrm{Aut}(R)|}P_{h_n}(R)\right| > \frac{r!\rho_n^{-\mathfrak{r}}\delta}{|\mathrm{Aut}(R)|}\right) \leqslant 2\exp\left(-\frac{\delta^2 n}{2r^2}\right).$$

For every $0 < \epsilon < 1$, let $\delta = \rho_n^{\mathfrak{r}}\epsilon|\mathrm{Aut}(R)|/r!$, then by (C.68) we have

$$pr\left(\left|\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n) - E[\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n)]\right| > \epsilon\right) \leqslant 2\exp\left(-\frac{\epsilon^2|\mathrm{Aut}(R)|^2\rho_n^{2\mathfrak{r}}n}{2(rr!)^2}\right).$$

When $\rho_n w(u,v) \leq 1$ for all u, v, the quantity $\rho_n^{-\mathfrak{r}}P_{h_n}(R) = P_w(R)$, which is a constant that does not depend on n. Thus, if there exist some $c_1 > 1$, such that $n\rho_n^{2\mathfrak{r}} > c_1 \log n$, then the summation of the following series converge:

$$\sum_n 2\exp\left(-\frac{\epsilon^2|\mathrm{Aut}(R)|^2\rho_n^{2\mathfrak{r}}n}{2(rr!)^2}\right).$$

By Borel-Cantelli lemma, we have

$$\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}_n) \xrightarrow{\text{a.s}} \frac{\rho_n^{-\mathfrak{r}}r!}{|\mathrm{Aut}(R)|}P_{h_n}(R) = \frac{r!}{|Aut(R)}P_w(R).$$

As a special case, when motif R is an edge, $\widehat{\rho}_G = U_R(G)$. Thus, Lemma 53 implies that with probability one:

$$\lim_{n\to\infty}\rho_n^{-\mathfrak{r}}\widehat{\rho}_G = \frac{r!}{|\mathrm{Aut}(R)|}P_w(R).$$

Since $\mathfrak{r} = 1$, $r = 2$, $|\mathrm{Aut}(R)| = 2$ (Rodriguez, 2014) and $P_w(R) = 1$ (Bickel et al., 2011), with probability one, we have

$$\lim_{n\to\infty}\rho_n^{-1}\widehat{\rho}_G = 1. \tag{C.167}$$

□

C.7 Proof of of Theorem 9

The following lemma is used for the proof.

Lemma 54 (Durrett (2019)). *Let $\{X_n\}$ be a sequence of random variables with CDF $\{F_{X_n}\}$, and let X be another random variable. If $X_n \xrightarrow{d} X$, and F is a continuous function, then the following convergence property holds:*

$$\sup_t |F_{X_n}(t) - F(t)| \to 0.$$

Lemma 55. *Let $\{X_n\}$ be a sequence of random variables with $\{F_{X_n}\}$ and $\lim_{n\to\infty} E[X_n^2] = 0$. That is, $\{X_n\}$ converge to zero in L_2. Let*

$$F_Y(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{if } t \geq 0. \end{cases}$$

Then

$$\sup_t |F_{X_n}(t) - F_Y(t)| \to 0.$$

of Lemma 55. Since $\lim_{n\to\infty} E[X_n^2] = 0$ by Chebyshev's inequality, for any $\epsilon > 0$,

$$P(|X_n| \geq \epsilon) \leq \frac{E[X_n^2]}{\epsilon^2} \to 0 \text{ as } n \to \infty. \tag{C.168}$$

Consequently,

- For any $t \geq 0$, (C.168) directly implies that

$$\lim_{n\to\infty} P(|X_n| \leq t) = 1$$

.

- For any $t < 0$, let $\epsilon = -t > 0$, then with probability one:

$$\lim_{n\to\infty} P(|X_n| \geq -t) = 0,$$

which leads to

$$\lim_{n\to\infty} P(|X_n| \leq t) = 0.$$

Consequently,

$$\lim_{n\to\infty} \sup_{t<0} |F_{X_n}(t)| = 0,$$

$$\lim_{n\to\infty} \sup_{t\geq 0} |F_{X_n}(t) - 1| = 0,$$

which implies

$$\lim_{n\to\infty} \sup_{t} |F_{X_n}(t) - F_Y(t)| = 0.$$

□

of Theorem 9.

We first focus on a single motif R, and consider two scenarios: non-degeneration and degeneration. We want to show that, with probability one:

$$\sup_{t\in\mathbb{R}} \left| J^R_{*,n,b}(t) - J^R_{b,(1-\frac{b}{n})}(t) \right| \to 0. \tag{C.169}$$

i First, (C.169) can be upper bounded as

$$\begin{aligned}\sup_{t\in\mathbb{R}}\big|J^R_{*,n,b}(t)-J^R_{b,(1-\frac{b}{n})}(t)\big| &= \sup_{t\in\mathbb{R}}\big|J^R_{b,(1-\frac{b}{n})}(t)-J^R_{*,n,b}(t)\big| \\ &\leqslant \sup_{t\in\mathbb{R}}\big|J^R_{b,(1-\frac{b}{n})}(t)-J^R_{b,c}(t)\big|+\sup_{t\in\mathbb{R}}\big|J^R_{b,c}(t)-\Phi\big(\frac{t}{\sigma_{c,R}}\big)\big| \\ &\quad+\sup_{t\in\mathbb{R}}\big|\Phi\big(\frac{t}{\sigma_{c,R}}\big)-\Phi\big(\frac{t}{\sigma_{*R}}\big)\big|+\sup_{t\in\mathbb{R}}\big|J^R_{*,n,b}(t)-\Phi\big(\frac{t}{\sigma_{*R}}\big)\big|,\end{aligned} \tag{C.170}$$

where $c=1-c_2$, $\sigma^2_{c,R}=c\sigma^2_R$ with $\sigma^2_R=\lim_{b\to\infty}\operatorname{var}\big[\sqrt{b}\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)\big]$, and

$$\sigma^2_{*R}=\lim_{b\to\infty}\operatorname{var}_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_R(\mathbb{G}^*_b)\big].$$

Now we are in the position to show that all four components on the right-hand side of (C.170) go to zero. By Slutsky's theorem, we have

$$\sup_{t\in\mathbb{R}}\big|J^R_{b,(1-\frac{b}{n})}(t)-J^R_{b,c}(t)\big|\to 0.$$

Then Lemma 51 implies that:

$$\sqrt{b}\big\{\widehat{\rho}^{-\mathfrak{r}}_{\mathbb{G}_b}U_R(\mathbb{G}_b)-E\big[\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)\big]\big\}\xrightarrow{d}\mathcal{N}(0,\sigma^2_R),$$

which gives:

$$\sqrt{bc}\big\{\widehat{\rho}^{-\mathfrak{r}}_{\mathbb{G}_b}U_R(\mathbb{G}_b)-E\big[\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)\big]\big\}\xrightarrow{d}\mathcal{N}(0,c\sigma^2_R).$$

Consequently, since σ_R is fixed and Φ is continuous, Lemma 54 implies that

$$\sup_{t\in\mathbb{R}}\big|J^R_{b,c}(t)-\Phi\big(\frac{t}{\sigma_{c,R}}\big)\big|\to 0.$$

For the third term, recall that $G \sim \mathbb{G}_n$. Lemma 52 implies that σ^2_{*R} equals to $c\sigma^2_R$ almost surely. Thus, by using Lemma 54 again, we have

$$\sup_{t\in\mathbb{R}} \left|\Phi\Big(\frac{t}{\sigma_{*R}}\Big) - \Phi\Big(\frac{t}{\sigma_{c,R}}\Big)\right| \to 0.$$

Next, since Assumptions 3, 2, 1 are satisfied, then by (C.91), with probability one:

$$\sqrt{b}\big[\rho_n^{-\mathfrak{r}} U_R(\mathbb{G}_b^*) - \rho_n^{-\mathfrak{r}} U_R(G)\big] \to \mathcal{N}(0, \sigma^2_{*R}) \text{ in distribution.}$$

In addition, (C.167) implies that $\lim_{n\to\infty} \rho_n^{-1}\widehat{\rho}_G = 1$ with probability one: . Consequently, by Slutsky's theorem, with probability one:

$$\sqrt{b}\big[\widehat{\rho}_G^{-\mathfrak{r}} U_R(\mathbb{G}_b^*) - \widehat{\rho}_G^{-\mathfrak{r}} U_R(G)\big] \to \mathcal{N}\big(0, \sigma^2_{*R}\big) \text{ in distribution,}$$

which further implies with probability one:

$$\sup_{t\in\mathbb{R}} \left|J^R_{*,n,b}(t) - \Phi\Big(\frac{t}{\sigma_{*R}}\Big)\right| \to 0.$$

Therefore, with probability one:

$$\sup_{t\in\mathbb{R}} \left|J^R_{*,n,b}(t) - J^R_{b,c}(t)\right| \to 0.$$

ii Suppose that $\sigma^2_R = 0$. Then by definition

$$\sqrt{bc}\Big\{\widehat{\rho}_{\mathbb{G}_b}^{-\mathfrak{r}} U_R(\mathbb{G}_b) - E[\rho_b^{-\mathfrak{r}} U_R(\mathbb{G}_b)]\Big\}$$

converge to zero in L_2. Thus, by Lemma 55, we have

$$\sup_{t\in\mathbb{R}}\left|J^R_{b,c}(t) - F_Y(t)\right| \to 0.$$

On the other hand, Lemma 52 implies that $\sigma^2_{*R} = 0$ almost surely. In addition, (C.75) implies that

$$E_*\left[U_R(\mathbb{G}^*_b)\right] = U_R(G), \tag{C.171}$$

Thus, we have

$$\sqrt{bc}\Big\{\widehat{\rho}^{-\mathfrak{r}}_{\mathbb{G}_b}U_R(\mathbb{G}_b) - E[\rho_b^{-\mathfrak{r}}U_R(\mathbb{G}_b)]\Big\}$$

converge to zero in L_2. Since (C.167) implies that $\lim_{b\to\infty}\rho_b^{-1}\widehat{\rho}_{\mathbb{G}_b} = 1$ with probability one, we have

$$\sqrt{b}\big[\widehat{\rho}^{-\mathfrak{r}}_{G}U_R(\mathbb{G}^*_b) - \widehat{\rho}^{-\mathfrak{r}}_{G}U_R(G)\big]$$

converge to zero in L_2 with probability one. Again by Lemma 55, with probability one:

$$\sup_{t\in\mathbb{R}}\left|J^R_{*,n,b}(t) - F_Y(t)\right| \to 0.$$

Therefore, with probability one:

$$\sup_{t\in\mathbb{R}}\left|J^R_{*,n,b}(t) - J^R_{b,c}(t)\right| \to 0.$$

Now we turn to consider m motifs $\{R_1, \cdots, R_m\}$. For simplicity, let $[t_m] = \{t_1, \cdots, t_m\}$ and $[R_m] = \{R_1, \cdots, R_m\}$. We still consider non-degeneration and degeneration separately as follows.

i Under Assumption 3, similar as before, we break the Kolmogorov-Smirnov

distance into three parts:

$$\sup_{[t_m]\in\mathbb{R}^m}\left|J^{[R_m]}_{*,n,b}([t_m])-J^{[R_m]}_{b,c}([t_m])\right|\leqslant\sup_{[t_m]\in\mathbb{R}^m}\left|J^{[R_m]}_{b,c}([t_m])-\Phi_{\Sigma_c(\mathbf{R})}\big(\frac{t_1}{\sigma_{c,R_1}},\cdots,\frac{t_m}{\sigma_{c,R_m}}\big)\right|$$
$$+\sup_{[t_m]\in\mathbb{R}^m}\left|\Phi_{\Sigma_c(\mathbf{R})}\big(\frac{t_1}{\sigma_{c,R_1}},\cdots,\frac{t_m}{\sigma_{c,R_m}}\big)-\Phi_{\Sigma(*\mathbf{R})}\big(\frac{t_1}{\sigma_{*R_1}},\cdots,\frac{t_m}{\sigma_{*R_m}}\big)\right|$$
$$+\sup_{[t_m]\in\mathbb{R}^m}\left|J^{[R_m]}_{*,n,b}([t_m])-\Phi_{\Sigma(*\mathbf{R})}\big(\frac{t_1}{\sigma_{*R_1}},\cdots,\frac{t_m}{\sigma_{*R_m}}\big)\right|,$$

where for each $i\in[m]$, $\sigma^2_{c,R_i}=c\sigma^2_{R_i}$ with $\sigma^2_{R_i}=\lim_{n\to\infty}\mathrm{var}\big[\sqrt{n}\rho_n^{-\mathfrak{r}}U_{R_i}(\mathbb{G}_n)\big]$, and

$$\sigma^2_{*R_i}=\lim_{b\to\infty}\mathrm{var}_*\big[\sqrt{b}\rho_n^{-\mathfrak{r}}U_{R_i}(\mathbb{G}^*_b)\big].$$

As before, we now want to to show that all three components go to zero. By Lemma 51:

$$\sqrt{b}\Big\{\big[\widehat{\rho^{-}}^{\mathfrak{r}_1}_{\mathbb{G}_b}U_{R_1}(\mathbb{G}_b),\cdots,\widehat{\rho^{-}}^{\mathfrak{r}_m}_{\mathbb{G}_b}U_{R_m}(\mathbb{G}_b)\big]-\big[\rho_b^{-\mathfrak{r}_1}E[U_{R_1}(\mathbb{G}_b)],\cdots,\rho_b^{-\mathfrak{r}_m}E[U_{R_m}(\mathbb{G}_b)]\big]\Big\}$$
$$\to\mathcal{N}[0,\Sigma(\mathbf{R})]\text{ in distribution.}$$

Since $c>0$ is a constant, by Slutsky's theorem:

$$\sqrt{bc}\Big\{\big[\widehat{\rho^{-}}^{\mathfrak{r}_1}_{\mathbb{G}_b}U_{R_1}(\mathbb{G}_b),\cdots,\widehat{\rho^{-}}^{\mathfrak{r}_m}_{\mathbb{G}_b}U_{R_m}(\mathbb{G}_b)\big]-\big[\rho_b^{-\mathfrak{r}_1}E[U_{R_1}(\mathbb{G}_b)],\cdots,\rho_b^{-\mathfrak{r}_m}E[U_{R_m}(\mathbb{G}_b)]\big]\Big\}$$
$$\to\mathcal{N}[0,\Sigma_c(\mathbf{R})]\text{ in distribution.}$$

Thus, by Lemma 54:

$$\sup_{[t_m]\in\mathbb{R}^m}\left|J^{[R_m]}_{b,c}([t_m])-\Phi_{\Sigma_c(\mathbf{R})}\big(\frac{t_1}{\sigma_{c,R_1}},\cdots,\frac{t_m}{\sigma_{c,R_m}}\big)\right|\to 0. \tag{C.172}$$

The second term goes to zero because Lemma 52 implies that $\Sigma(*\mathbf{R})$ converge to $\Sigma_c(\mathbf{R})$ almost surely.

Under Assumptions 3, 2, 1, by Theorem 44, for any sequence $\{G^{(n)}\}$, with probability one:

$$\sqrt{b}\Big\{\big[\rho_n^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b^*),\cdots,\rho_n^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b^*)\big]-\big[\rho_n^{-\mathfrak{r}_1}U_{R_1}(G),\cdots,\rho_n^{-\mathfrak{r}_m}U_{R_m}(G)\big]\Big\}$$
$$\to\mathcal{N}\big[0,\Sigma(*\mathbf{R})\big] \text{ in distribution,}$$

Since $\lim_{n\to\infty}\rho_n^{-\mathfrak{r}}\widehat{\rho}_G=1$ with probability one,

$$\sqrt{b}\Big\{\big[\widehat{\rho_G}^{-\mathfrak{r}_1}U_{R_1}(\mathbb{G}_b^*),\cdots,\widehat{\rho_G}^{-\mathfrak{r}_m}U_{R_m}(\mathbb{G}_b^*)\big]-\big[\widehat{\rho_G}^{-\mathfrak{r}_1}U_{R_1}(G),\cdots,\widehat{\rho_G}^{-\mathfrak{r}_m}U_{R_m}(G)\big]\Big\}$$
$$\to\mathcal{N}\big[0,\Sigma(*\mathbf{R})\big] \text{ in distribution with probability one,}$$

which further implies that with probability one:

$$\sup_{[t_m]\in\mathbb{R}^m}\big|J_{*,n,b}^{[R_m]}([t_m])-\Phi_{\Sigma(*\mathbf{R})}(\frac{t_1}{\sigma_{*R_1}},\cdots,\frac{t_m}{\sigma_{*R_m}})\big|\to 0. \tag{C.173}$$

Therefore, with probability one:

$$\sup_{[t_m]\in\mathbb{R}^m}\big|J_{*,n,b}^{[R_m]}([t_m])-J_{b,c}^{[R_m]}([t_m])\big|\to 0,$$

which implies

$$\sup_{[t_m]\in\mathbb{R}^m}\big|J_{*,n,b}^{[R_m]}([t_m])-J_{b,(1-\frac{b}{n})}^{[R_m]}([t_m])\big|\to 0,$$

as $c=1-c_2$.

ii The proof of degenerate case is an natural extension of the previous result, replacing the random variables by random vectors, let Y be a zero vector with

CDF F_Y and then one can show both

$$\sup_{[t_m]\in\mathbb{R}^m} \left|J_{*,n,b}^{[R_m]}([t_m]) - F_Y([t_m])\right| \to 0,$$

$$\sup_{[t_m]\in\mathbb{R}^m} \left|J_{b,c}^{[R_m]}([t_m]) - F_Y([t_m])\right| \to 0,$$

and then the convergence follows directly as before.

□

C.8 Proof of the empirical consistency

As before, let $c = 1 - c_2$, $[t_m] = \{t_1, \cdots, t_m\}$ and $[R_m] = \{R_1, \cdots, R_m\}$. The following lemma is used for the proof.

Lemma 56 (Theorem 1 in Lunde and Sarkar (2023))**.** *Suppose that that there exists a CDF $J([t_m])$, such that for all continuity points of $J(\cdot)$,*

$$|J_{b,c}^{[R_m]}([t_m]) - J([t_m])| \to 0,$$
$$|J_{*,n,b}^{[R_m]}([t_m]) - J([t_m])| \to 0.$$

Then

$$\hat{J}_{*,n,b}^{[R_m]}([t_m]) \to J([t_m]) \text{ in probability.}$$

of Theorem 46. To begin with, let

$$J([t_m]) = \Phi_{\Sigma_c(\mathbf{R})}\left(\frac{t_1}{\sigma_{c,R_1}}, \cdots, \frac{t_m}{\sigma_{c,R_m}}\right)$$
$$J_*([t_m]) = \Phi_{\Sigma(*\mathbf{R})}\left(\frac{t_1}{\sigma_{*R_1}}, \cdots, \frac{t_m}{\sigma_{*R_m}}\right).$$

Under Assumptions 3, 2, 1, we have

$$|J_{b,c}^{[R_m]}([t_m]) - J([t_m])| \overset{(C.172)}{\to} 0,$$
$$|J_{*,n,b}^{[R_m]}([t_m]) - J_*([t_m])| \overset{(C.173)}{\to} 0.$$

Moreover, Lemma 52 implies that $\Sigma(*\mathbf{R})$ converge to $\Sigma_c(\mathbf{R})$ almost surely. Thus, $J_*([t_m])$ converge to $J([t_m])$ almost surely. Therefore, by Lemma 56, with probability one (measure on the random network sequence):

$$\hat{J}_{*,n,b}^{[R_m]}([t_m]) \to J([t_m]) \text{ in probability.}$$

Consequently, by Lemma 54:

$$\sup_{[t_m]\in\mathbb{R}^m} \left|\hat{J}_{*,n,b}^{[R_m]}([t_m]) - J([t_m])\right| \to 0.$$

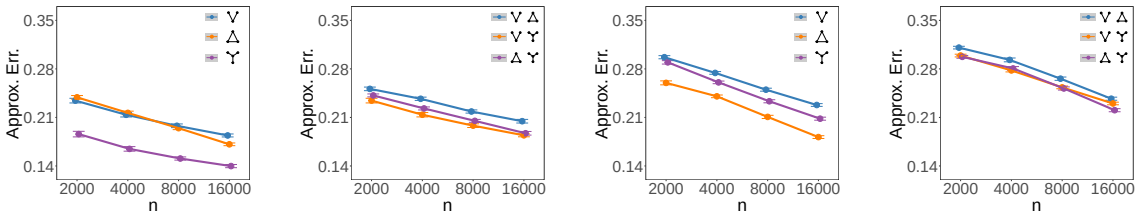
Finally, based on (C.172) and (C.173), we arrived at

$$\begin{aligned}
&\sup_{[t_m]\in\mathbb{R}^m} \left|\hat{J}_{*,n,b}^{[R_m]}([t_m]) - J_{*,n,b}^{[R_m]}([t_m])\right| \leqslant \sup_{[t_m]\in\mathbb{R}^m} \left|\hat{J}_{*,n,b}^{[R_m]}([t_m]) - \Phi_{\Sigma_c(\mathbf{R})}\Big(\frac{t_1}{\sigma_{c,R_1}}, \cdots, \frac{t_m}{\sigma_{c,R_m}}\Big)\right| \\
&+ \sup_{[t_m]\in\mathbb{R}^m} \left|\Phi_{\Sigma_c(\mathbf{R})}\Big(\frac{t_1}{\sigma_{c,R_1}}, \cdots, \frac{t_m}{\sigma_{c,R_m}}\Big) - \Phi_{\Sigma(*\mathbf{R})}\Big(\frac{t_1}{\sigma_{*R_1}}, \cdots, \frac{t_m}{\sigma_{*R_m}}\Big)\right| \\
&+ \sup_{[t_m]\in\mathbb{R}^m} \left|J_{*,n,b}^{[R_m]}([t_m]) - \Phi_{\Sigma(*\mathbf{R})}\Big(\frac{t_1}{\sigma_{*R_1}}, \cdots, \frac{t_m}{\sigma_{*R_m}}\Big)\right| \to 0.
\end{aligned}$$

□

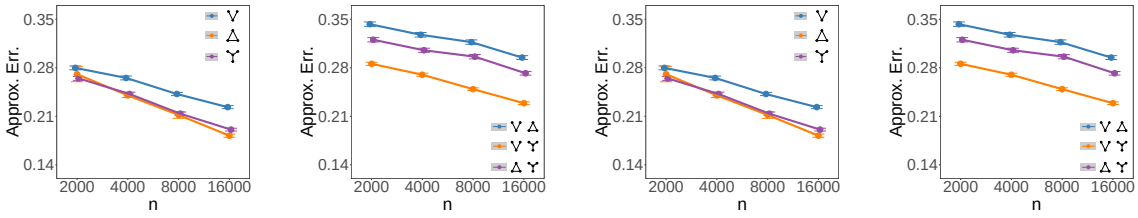
C.9 Additional simulation results

C.9.1 Additional simulation results for Section 4.4



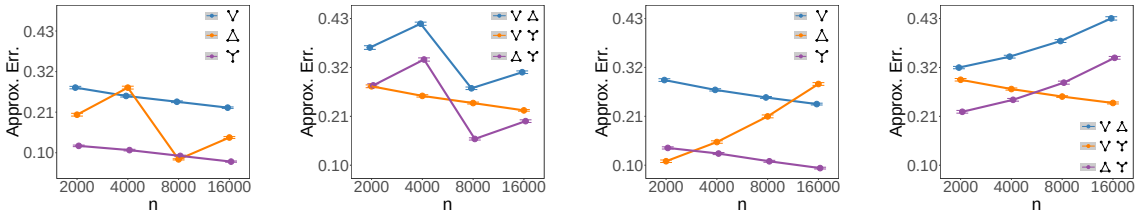
(a) Graphon 1: marginal CDFs **(b)** Graphon 1: joint CDFs **(c)** Graphon 2: marginal CDFs **(d)** Graphon 2: joint CDFs

Figure C.6: Empirical approximation errors of the CDFs under $b = \lceil 2n^{1/2} \rceil$ and $\rho_n = 0.25n^{-0.1}$.



(a) Graphon 1: marginal CDFs **(b)** Graphon 1: joint CDFs **(c)** Graphon 2: marginal CDFs **(d)** Graphon 2: joint CDFs

Figure C.7: Empirical approximation errors of the CDFs under $b = \lceil 2n^{1/2} \rceil$ and $\rho_n = 0.25n^{-0.25}$.



(a) Graphon 1: marginal CDFs **(b)** Graphon 1: joint CDFs **(c)** Graphon 2: marginal CDFs **(d)** Graphon 2: joint CDFs

Figure C.8: Empirical approximation errors of the CDFs under $b = \lceil n^{2/3} \rceil$ and $\rho_n = 0.25n^{-0.5}$.

List of Figures

List of Tables

Reference

J. Agterberg, M. Tang, and C. Priebe. Nonparametric two-sample hypothesis testing for random graphs with negative and repeated eigenvalues. *arXiv preprint arXiv:2012.09828*, 2020.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

A. A. Alyakin, J. Agterberg, H. S. Helm, and C. E. Priebe. Correcting a nonparametric two-sample graph hypothesis test for graphs with different numbers of vertices with applications to connectomics. *Applied Network Science*, 9(1):1, 2024.

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

O. Amini, F. V. Fomin, and S. Saurabh. Counting subgraphs via homomorphisms. *SIAM Journal on Discrete Mathematics*, 26(2):695–717, 2012.

E. Austin, W. Pan, and X. Shen. A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. *Statistica Sinica*, 30(2):783, 2020.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6), 2008.

A.-L. Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

R. L. Barter, S.-J. Schramm, G. J. Mann, and Y. H. Yang. Network-based biomarkers enhance classical approaches to prognostic gene expression signatures. *BMC systems biology*, 8:1–16, 2014.

S. Basu, A. Shojaie, and G. Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

B. B. Bhattacharya, A. Chatterjee, and S. Janson. Fluctuations of subgraph counts in graphon based random graphs. *Combinatorics, Probability and Computing*, pages 1–37, 2022.

S. Bhattacharyya and P. Bickel. Supplement to "subsampling bootstrap of count features of networks.", 2015a.

S. Bhattacharyya and P. J. Bickel. Subsampling bootstrap of count features of networks. *The Annals of Statistics*, 43(6):2384–2411, 2015b.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

M. Bloznelis and F. Götze. Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *The Annals of Statistics*, 29 (3):899–917, 2001.

M. Bloznelis and F. Götze. An edgeworth expansion for symmetric finite population statistics. *The Annals of Probability*, 30(3):1238–1265, 2002.

B. Bollobás and O. Riordan. Metrics for sparse graphs. *arXiv preprint arXiv:0708.1919*, 2007.

C. Borgs, J. Chayes, and L. Lovász. Moments of two-variable functions and the uniqueness of graph limits. *Geometric and functional analysis*, 19:1597–1619, 2010.

S. Boyd, N. Parikh, E. Chu, B.Peleato, and J.Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3:1–122, 2011.

T. T. Cai and W. Liu. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016.

T. T. Cai and A. Zhang. Inference for high-dimensional differential correlation matrices. *Journal of multivariate analysis*, 143:107–126, 2016.

T. T. Cai, A. R. Zhang, and Y. Zhou. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*, 2022.

F. Campbell and G. I. Allen. Within group variable selection through the exclusive lasso. *Electronic Journal of Statistics*, 11(2):4220–4257, 2017.

Y. Cao, W. Lin, and H. Li. Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association*, 114(526):759–772, 2019.

S. Chan and E. Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216. PMLR, 2014.

J. Chang, W. Zhou, W.-X. Zhou, and L. Wang. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics*, 73(1):31–41, 2017.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

S. Chatterjee, D. Saha, S. Dan, and B. B. Bhattacharya. Two-sample tests for inhomogeneous random graphs in l_r norm: Optimality and asymptotics. In *International Conference on Artificial Intelligence and Statistics*, pages 6903–6911. PMLR, 2023.

J. Chen, C. Liu, J. Cen, T. Liang, J. Xue, H. Zeng, Z. Zhang, G. Xu, C. Yu, Z. Lu, et al. Kegg-expressed genes and pathways in triple negative breast cancer: Protocol for a systematic review and data mining. *Medicine*, 99(18), 2020.

K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113 (521):241–251, 2018.

X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378, 2017a.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26, 2017b.

D. Choi and P. J. Wolfe. Co-clustering separately exchangeable network data1. *The Annals of Statistics*, 42(1):29–63, 2014.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17:1–19, 2016.

S. A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158, 1971.

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

A. Dedieu. An error bound for lasso and group lasso in high dimensions. *arXiv:1912.11398*, 2019.

L. Deng, L. Ma, K.-K. Cheng, X. Xu, D. Raftery, and J. Dong. Sparse pls-based method for overlapping metabolite set enrichment analysis. *Journal of proteome research*, 20(6):3204–3213, 2021.

W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. *Proceedings of the SPIE*, 2013.

D. Deshpande, K. Chhugani, Y. Chang, A. Karlsberg, C. Loeffler, J. Zhang, A. Muszyńska, V. Munteanu, H. Yang, J. Rotman, et al. Rna-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14:997383, 2023.

J. Ding, V. Tarokh, and Y. Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018.

X. Du and M. Tang. Hypothesis testing for equality of latent positions in random graphs. *Bernoulli*, 29(4):3221–3254, 2023.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

H. Fang, C. Huang, H. Zhao, and M. Deng. gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708, 2017.

E. K. Fischer, Y. Song, K. A. Hughes, W. Zhou, and K. L. Hoke. Nonparallel transcriptional divergence during parallel adaptation. *Molecular Ecology*, 30(6): 1516–1530, 2021.

J. H. Friedman, T. J. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv: Statistics Theory*, 2010.

C. Gao and J. Lafferty. Testing network structure using relations between small subgraph probabilities. *arXiv preprint arXiv:1704.06742*, 2017.

C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, pages 2624–2652, 2015.

S. Ghazanfar, D. Strbenac, J. T. Ormerod, J. Y. Yang, and E. Patrick. Dcars: differential correlation across ranked samples. *Bioinformatics*, 35(5):823–829, 2019.

D. Ghoshdastidar and U. Von Luxburg. Practical methods for graph two-sample testing. *Advances in Neural Information Processing Systems*, 31, 2018.

D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg. Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory*, pages 954–977. PMLR, 2017.

M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 11 2021.

M. Gonen, D. Ron, and Y. Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM Journal on Discrete Mathematics*, 25(3):1365–1411, 2011.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, second edition,

1994. ISBN 0201558025 9780201558029 0201580438 9780201580433 0201142368 9780201142365.

E. Grave, G. R. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. *Advances in Neural Information Processing Systems*, 24, 2011.

A. Green and C. R. Shalizi. Bootstrapping exchangeable random graphs. *Electronic Journal of Statistics*, 16(1):1058–1095, 2022.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

D. N. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2:275, 1979.

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38 (4):1978–2004, 2010. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/20744481.

L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. *Proceedings of the 26th Annual International Conference on Machine Learning,ICML*, 09:433–440, 2009.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011a.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12:2297–2334, 2011b.

P. Ji and J. Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, pages 1779–1812, 2016.

P. Ji, J. Jin, Z. T. Ke, and W. Li. Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, 40(2):469–485, 2022.

M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44 (D1):D457–D462, 10 2015.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

S. Kim and E. P. Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6:1095–1117, 2012.

J. M. Klusowski and Y. Wu. Estimating the number of connected components in a graph via subgraph sampling. *Bernoulli*, 26(3):1635–1664, 2020.

S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19): 2373–2380, 2006.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

S. Lee and E. Xing. Screening rules for overlapping group lasso. *arXiv:1410.6880*, 2014.

S. Lee and E. P. Xing. Leveraging input and output structures for joint mapping of epistatic and marginal eqtls. *Bioinformatics*, 28(12):i137–i146, 2012.

K. Levin and E. Levina. Bootstrapping networks with latent space structure. *arXiv preprint arXiv:1907.10821*, 2019.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008.

J. Li and S. X. Chen. Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940, 2012.

T. Li and C. M. Le. Network estimation by mixing: Adaptivity and more. *Journal of the American Statistical Association*, pages 1–16, 2023.

T. Li, E. Levina, and J. Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

T. Li, L. Lei, S. Bhattacharyya, K. Van den Berge, P. Sarkar, P. J. Bickel, and E. Levina. Hierarchical community detection by recursive partitioning. *Journal of the American Statistical Association*, 117(538):951–968, 2022.

T. Li, X. Tang, and A. Chatrath. Compressed spectral screening for large-scale differential correlation analysis with application in selecting glioblastoma gene modules. *The Annals of Applied Statistics*, 17(4):3450–3475, 2023.

Y. Li and H. Li. Two-sample test of community memberships of weighted stochastic block models. *arXiv preprint arXiv:1811.12593*, 2018.

F. Liu, W. Xu, J. Lu, and D. J. Sutherland. Meta two-sample testing: Learning kernels for testing with limited data. *Advances in Neural Information Processing Systems*, 34:5848–5860, 2021.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009a.

J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009b. URL http://www.public.asu.edu/~jye02/Software/SLEP.

A. Livshits, A. Git, G. Fuks, C. Caldas, and E. Domany. Pathway-based personalized analysis of breast cancer expression data. *Molecular oncology*, 9(7):1471–1483, 2015.

P.-L. Loh. *High-dimensional statistics with systematically corrupted data*. University of California, Berkeley, 2014.

K. Lounici, M. Pontil, S. V. D. Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

L. Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

M. Love, S. Anders, and W. Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.

R. Lunde and P. Sarkar. Subsampling sparse graphons under minimal assumptions. *Biometrika*, 110(1):15–32, 2023.

R. Lunde, E. Levina, and J. Zhu. Conformal prediction for network-assisted regression. *arXiv preprint arXiv:2302.10095*, 2023.

S. Ma and M. R. Kosorok. Detection of gene pathways with predictive power for breast cancer prognosis. *BMC bioinformatics*, 11(1):1–11, 2010.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. *Journal of Machine Learning Research*, 14(8), 2013.

J. Mairal, R. Jenatton, F. Bach, and G. R. Obozinski. Network flow algorithms for structured sparsity. *Advances in Neural Information Processing Systems*, 23, 2010.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, R. Jenatton, and G. Obozinski. Spams: A sparse modeling software, v2. 3. *URL http://spams-devel. gforge. inria. fr/downloads. html*, 2014.

C. Mao, Y. Wu, J. Xu, and S. H. Yu. Testing network correlation efficiently via counting trees. *arXiv preprint arXiv:2110.11816*, 2021.

P.-A. Maugis, S. Olhede, C. Priebe, and P. Wolfe. Testing for equivalence of network distribution using subgraph counts. *Journal of Computational and Graphical Statistics*, 29(3):455–465, 2020.

A. T. McKenzie, I. Katsyv, W.-M. Song, M. Wang, and B. Zhang. Dgca: a comprehensive r package for differential gene correlation analysis. *BMC systems biology*, 10:1–25, 2016.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71, 2008.

I. Menashe, D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer research*, 70(11):4453–4459, 2010.

R. Miao and T. Li. Informative core identification in complex networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):108–126, 2023.

H. E. Miller and A. J. Bishop. Correlation analyzer: functional predictions from gene co-expression correlations. *BMC bioinformatics*, 22:1–19, 2021.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

K. P. Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

M. Newman. *Networks*. Oxford university press, 2018.

M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

S. Nowakowski, P. Pokarowski, W. Rejchel, and A. Sołtys. Improving group lasso for high-dimensional categorical data. In *International Conference on Computational Science*, pages 455–470. Springer, 2023.

S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

H. Park, A. Niida, S. Miyano, and S. Imoto. Sparse overlapping group lasso for integrative multi-omics analysis. *Journal of Computational Biology*, 22(2):73–84, 2015.

R. N. Perry, D. Albarracin, R. Aherrahrou, and M. Civelek. Network preservation analysis reveals dysregulated metabolic pathways in human vascular smooth muscle cell phenotypic switching. *Circulation: Genomic and Precision Medicine*, 16 (4):372–381, 2023.

Z. Qin, K.Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the grouplasso. *Mathematical Programming Computation*, 5(2), 2013.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5): 1009–1030, 2009.

P. Ribeiro and F. Silva. G-tries: an efficient data structure for discovering network motifs. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 1559–1566, 2010.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

L. Rodriguez. Automorphism groups of simple graphs, 2014.

C. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. Buetow. Pid: The pathway interaction database. *Nature Precedings*, 3, 08 2008. doi: 10.1038/npre.2008.2243.1.

J. R. Schott. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis*, 51 (12):6535–6542, 2007.

M. Schroeder, B. Haibe-Kains, A. Culhane, C. Sotiriou, G. Bontempi, and J. Quackenbush. *breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI).*, 2021. URL http://compbio.dfci.harvard.edu/. R package version 1.32.0.

M. Shao, D. Xia, Y. Zhang, Q. Wu, and S. Chen. Higher-order accurate two-sample network inference and network hashing. *arXiv preprint arXiv:2208.07573*, 2022.

A. Shojaie. Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1508, 2021.

D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafati,

M. Martens, R. Miller, K. Nishida, L. Rieswijk, A. Waagmeester, L. M. T. Eijssen, C. T. Evelo, A. R. Pico, and E. L. Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, 11 2017. ISSN 0305-1048.

C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14:1–18, 2013.

R. Stark, M. Grzelak, and J. Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017a.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23 (3):1599 – 1630, 2017b.

A. Tank, E. B. Fox, and A. Shojaie. An efficient admm algorithm for structural break detection in multivariate time series. *arXiv preprint arXiv:1711.08392*, 2017.

D. A. Tarzanagh and G. Michailidis. Estimation of graphical models through structured norm minimization. *Journal of machine learning research*, 18(1), 2018.

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

M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

S. Xiang, X. Shen, and J. Ye. Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence*, 224:28–50, 2015. ISSN 0004-3702.

B. Yan and P. Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 116(534):734–745, 2021.

X. Yan and J. Bien. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.

C. Yang, X. Wan, Q. Yang, H. Xue, and W. Yu. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC bioinformatics*, 11(1):1–11, 2010.

J. Yang and J. Peng. Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, 29(1):191–202, 2020.

J. Yang, C. Han, and E. Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *Artificial Intelligence and Statistics*, pages 1060–1067. PMLR, 2014.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

S. J. Young and E. R. Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

G. Yu and J. Bien. Learning local dependence in ordered data. *The Journal of Machine Learning Research*, 18(1):1354–1413, 2017.

H. Yuan, R. Xi, C. Chen, and M. Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. *Advances in Neural Information Process Systems*, pages 352–360, 2011.

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

M. Yuan and Q. Wen. A practical two-sample test for weighted random graphs. *Journal of Applied Statistics*, 50(3):495–511, 2023.

M. Yuan, R. Liu, Y. Feng, and Z. Shang. Testing community structure for hypergraphs. *The Annals of Statistics*, 50(1):147–169, 2022.

Y. Zeng and P. Breheny. Overlapping group logistic regression with applications to genetic pathway selection. *Cancer informatics*, 15:CIN–S40043, 2016.

Y. Zhang and D. Xia. Edgeworth expansions for network moments. *The Annals of Statistics*, 50(2):726–753, 2022.

Y. Zhang, E. Levina, and J. Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, 2017.

L. Zhao and X. Chen. Normal approximation for finite-population u-statistics. *Acta mathematicae applicatae Sinica*, 6(3):263–272, 1990.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009a.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009b.

Y. Zhao. *Graph Theory and Additive Combinatorics: Exploring Structure and Randomness*. Cambridge University Press, 2023.

Y. Zhou, B. Xu, Y. Zhou, J. Liu, X. Zheng, Y. Liu, H. Deng, M. Liu, X. Ren, J. Xia, et al. Identification of key genes with differential correlations in lung adenocarcinoma. *Frontiers in cell and developmental biology*, 9:675438, 2021.

L. Zhu, J. Lei, B. Devlin, and K. Roeder. Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *The annals of applied statistics*, 11(3):1810, 2017.