ABSTRACT

SONG, YUKUN. Statistical Inference on Random Graphs. (Under the direction of Minh Tang).

Graph or network data is widely studied across various fields such as Statistics, Machine Learning, Computer Science, and Biology, due to its ability to represent complex systems through vertices (objects) and edges (relationships). These graphs can model social networks, biological networks, and transportation systems. By analyzing these graphs, we can uncover underlying structures, such as identifying user groups in social platforms or predicting future connections. As the volume of graph data grows, the analysis has extended from single to multiple graphs, enabling us to identify common properties, compare similarities and differences, and perform sophisticated hypothesis testing for graph equivalence and correlation.

In Chapter 2, we propose an improved methodology for testing differences between two random graphs. By enhancing existing semiparametric methods, we derive new test statistics that are more powerful under alternative hypotheses. These statistics allow for the detection of smaller differences between two graphs, improving the reliability and accuracy of hypothesis testing in random graph models.

In Chapter 3, we consider the problem of testing for independence between two inhomogeneous random graphs on the same vertex set. We introduce a notion of pairwise edge correlations and derive a necessary condition for their detectability. We also show that the problem can exhibit a statistical vs. computational tradeoff. Additionally, we propose an asymptotically valid and consistent test procedure with the polynomial time complexity for a latent space model.

In Chapter 4, we consider statistical inference on multiple graphs. Some related research shows that an Omnibus matrix leads to the inference of the latent position for multiple graphs jointly and simultaneously without pairwise Procrustes alignment. We derive a new approach to recover latent positions, overcoming the limitation in existing methods that all graphs have the same or similar latent positions.

In Chapter 5, we summarize the contributions of the advanced methods developed for statistical inference on random graphs, including improved two-sample hypothesis testing, independence testing procedures, and a novel joint inference method for multiple graphs, and we discuss potential future directions.

Statistical Inference on Random Graphs

by
Yukun Song

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2024

APPROVED BY:

Ryan Martin	Luo Xiao
Nathaniel Josephs	Alexander Chouljenko

Minh Tang
Chair of Advisory Committee

DEDICATION

To my family.

BIOGRAPHY

The author was born in Chengdu, Sichuan, China. He earned a Bachelor of Science degree in Mathematics from Nankai University in 2015. Subsequently, he conducted research in Mathematics under the guidance of Dr. Wojbor A. Woyczynski and earned his Master of Science degree in Mathematics from Case Western Reserve University in 2017. Afterwards, he moved to Raleigh to begin his graduate studies in Statistics at NC State University, obtaining his Master of Statistics in 2020. He is currently pursuing a Ph.D. in Statistics, working under the guidance of Dr. Minh Tang on research topics related to statistical inference on random graphs.

ACKNOWLEDGEMENTS

First, I would like to express my deepest thanks to my advisor, Dr. Minh Tang, for his patient guidance over all these years. He has always encouraged me to pursue my research interests and provided insightful suggestions when I encountered research problems. I'd also like to thank my other committee members, Dr. Ryan Martin, Dr. Luo Xiao, Dr. Nathaniel Josephs, Dr. Eric Chi, and Dr. Alexander Chouljenko. I am grateful to them all for sharing their time and knowledge with me during my graduate studies, providing valuable feedback and advice during my written and oral prelim exams, and assisting in the development of my dissertation.

My appreciation also goes to the faculty members and staff in the Department of Statistics for all their guidance and help during the past five years.

Last but not least, I am truly grateful to my family for their continuous and unconditional support. Thanks to all my friends for their selfless help and support over all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Graph or network data is widely studied in Statistics, Machine Learning, Computer Science, and Biology. Graphs are ideal for representing complex systems where there are different objects represented by vertices and their pairwise relationships represented by edges, such as social networks, biological networks, and transportation systems. In a social platform, the vertices can represent users, and the edges correspond to the relationship between them. In a neural system the vertices can represent neurons, and the edges correspond to the state of sending signals between them. We can extract underlying structures from such data. For example, in a social platform, we can identify groups of users with similar characteristics and predict future friendships (Bedi and Sharma 2016; Hasan and Zaki 2011). There are many previous works on a single graph to explore the underlying structures. For example, many researches focus on the relationships among vertices, such as community detection (Clauset et al. 2004; Blondel et al. 2008; Bae et al. 2017) and vertex classification (Zhu et al. 2003; Madhawa and Murata 2020).

As graph data increases, it is common to focus on analysis on multiple graphs. First, there are some similar things as for single graph, such as community detection but on multiple graphs (Bhattacharyya and Chatterjee 2018; Lei and Lin 2023; Huang et al. 2023). Here, we can identify common properties across graphs by considering them jointly. Additionally, we can analyze graphs separately to compare their similarities and differences. As for similarity,

the graphs constructed from the wiring diagrams of the mushroom body for the left and right hemispheres of the *Drosophila* larva are highly correlated as a pair of neurons are more likely to send signals to each other in the left hemisphere if their correspondences also send signals to each other in the right hemisphere (Eichler et al. 2017; Winding et al. 2023). As for difference, there arises problems on graph comparison (Bullmore and Sporns 2009; Richiardi et al. 2011, 2013) to determine whether graphs are statistically equivalent. Tang et al. (2017a,b); Levin et al. (2017) provide a sophisticated two or multiple sample hypothesis test for random graphs.

A graph on n vertices is defined as $\mathcal{G} = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of vertices and E is the set of edges. In the dissertation, we focus on the simple graph, where the edges are undirected and unweighted, without self-loops as shown in Fig 1.1. Hence, a pair of vertices $(v_1, v_2) \in E$ means that the vertex v_i and the vertex v_j are connected with an edge. We usually use an adjacency matrix A to express the graph G. For the simple graph, $A_{ij} = 1$ if there exists an edge between vertices i and j and $A_{ij} = 0$ if there is no edge between them. Hence, the adjacency matrix A is symmetric with $A_{ii} = 0$ for no self loops. For the simple graph in Fig 1.1, we can express it by using an adjacency matrix

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

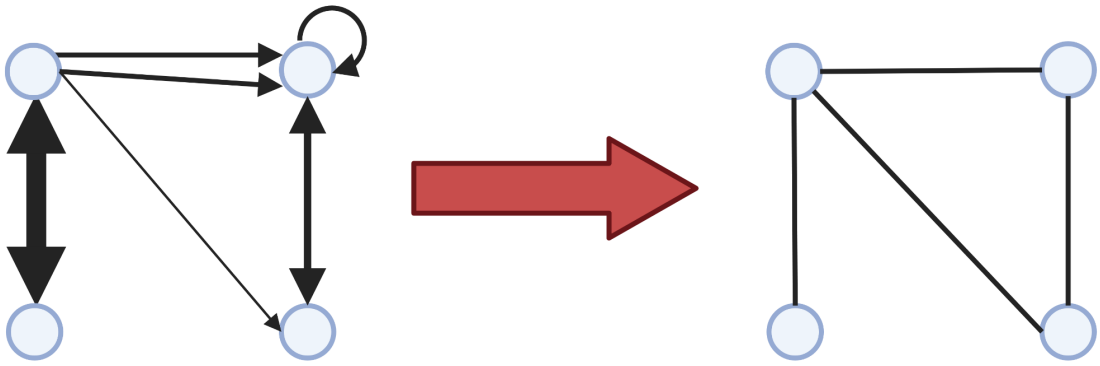


Figure 1.1: Simple Graph

When we have one or more graphs with known vertex correspondence, such as in Fig 1.2, we are interested in the relationships within the vertices or the relationships among graphs.

For example, we may ask how similar the two graphs are. This question can be translated into comparing the similarity of the corresponding adjacency matrices as shown in Fig 1.3. Hence, one line of inquiry is whether the two matrices are generated from the same latent distribution, while another is whether the two matrices are generated independently.

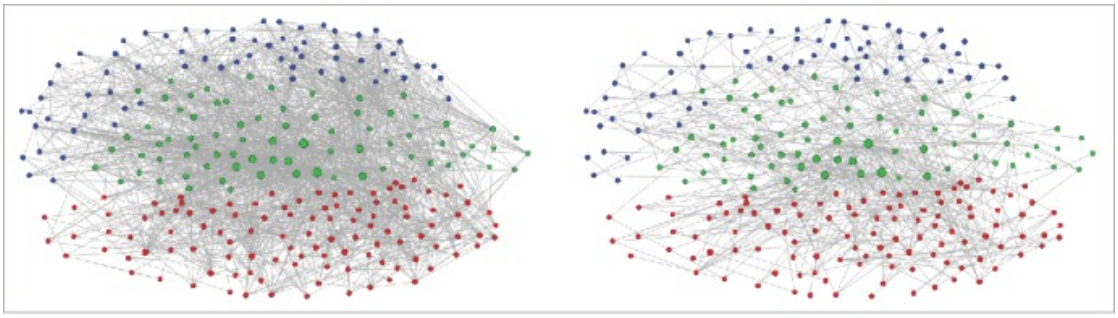


Figure 1.2: Red nodes correspond to motor neurons, green nodes correspond to interneurons, and blue nodes correspond to sensory neurons. The chemical connectome is on the left. The electrical gap junctional connectome is on the right. (Chen et al. 2016)

In Chapter 2, we focus on hypothesis testing problems on two graphs. Tang et al. (2017a) introduces a semiparametric two-sample hypothesis testing procedure for random dot product graphs (RDPG), which effectively approximates a broad spectrum of random graphs, ranging from simple to complex (Athreya et al. 2018). This method assumes that the two random dot product graphs share the same vertex set with known vertex correspondence. The primary goal of these tests is to determine whether the two models share identical generating latent positions or whether their generating latent positions are scaled or diagonally transformed versions of each other. Tang et al. (2017b) proposes statistics based on the estimated latent positions, with the statistics having a supremum limit not exceeding 1 under the null hypothesis. However, the limited information provided by these statistics can lead to indeterminacy and conservatism in hypothesis testing. We improve the existing statistics by setting the limiting value to 1. Consequently, these statistics are more likely to be significant and asymptotically greater than 1 under the alternative hypothesis when the two graphs are independent. This improvement enables the detection of smaller differences between two graphs, transitioning from a "bounded" to an "approximating" approach. When the two graphs are not independent, the limiting value may vary, but it remains straightforward to calculate this limit under the condition that, for all pairs of vertices, the edges between them exhibit the same correlation across the two graphs. We also conduct related simulations to support our results.

In Chapter 3, we address independence testing for inhomogeneous Erdős-Rényi random

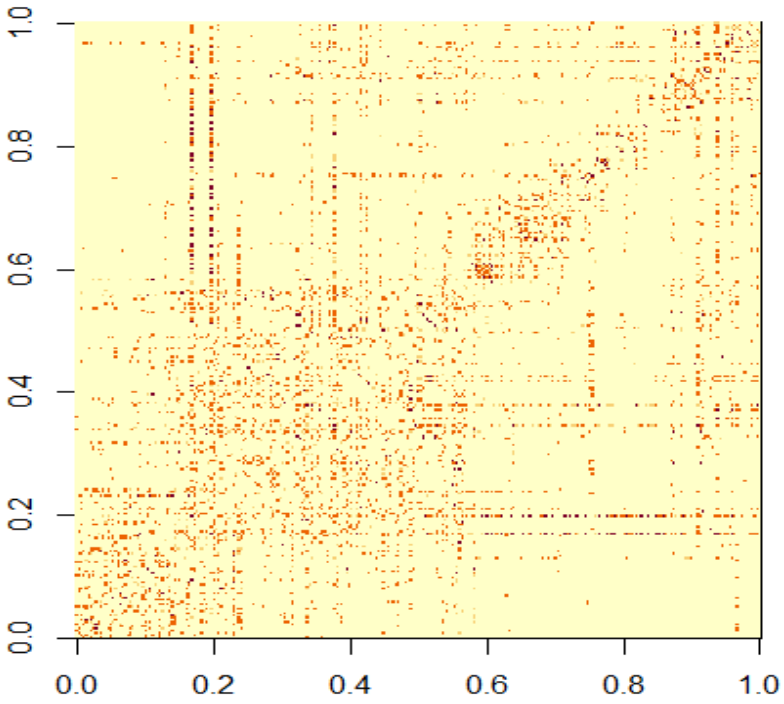


Figure 1.3: Adjacency Matrices of Sample Graphs. The orange is for the chemical connectome only. The yellow one is for junctional connectome. The dark is for both.

graphs on the same vertex set. Detecting correlation and testing for independence between Euclidean vectors is one of the most classical and widely studied inference problems in multivariate statistics. In contrast, independence testing for graphs is much less studied. Part of the difficulty lies in choosing an appropriate notion of correlation for graph-valued data. We are motivated by the problem of, given two graphs on the same vertex set, determining whether or not the presence of an edge between any pair of vertices in the first graph is *stochastically* independent of the presence of the corresponding edge in the second graph. Two graphs are thus said to be independent if all edges in the first are *pairwise* independent of the corresponding edges in the second, and are said to be correlated otherwise. We consider independence testing in the setting where the graphs are, *marginally*, inhomogeneous Erdős-Rényi random graphs while the pairwise edges correlations are described by the entries of a symmetric matrix R (see Definition 4). Two graphs are then independent if $R_{ij} \equiv 0$ for all $\{i, j\}$ and are correlated if $|R_{ij}| > 0$ for *some* $\{i, j\}$. Equivalently, we are interested in testing the null hypothesis $\mathbb{H}_0 \colon \|R\|_F = 0$ against the alternative hypothesis $\mathbb{H}_A \colon \|R\|_F > 0$. Our

contributions are as follows. We first show that there exists a procedure for deciding between $\mathbb{H}_0$ and $\mathbb{H}_A$ with asymptotically *vanishing* type-I and type-II errors only if $\|R\|_F \to \infty$ as $n \to \infty$ (see Section 3.3). We also describe two related examples where the condition $\|R\|_F \to \infty$ is sufficient for one example but not sufficient for the other. We next show that the problem exhibits a statistical vs. computational tradeoff, i.e., there are regimes for which $\|R\| \to \infty$ that are statistically detectable but may require running time which scales exponentially with n. We achieve this through a polynomial time reduction of the planted clique problem, which is well-known to be computationally hard (Alon et al. 1998; Barak et al. 2019), to our correlation testing problem (see Theorem 5). Finally, we consider a special case of our correlation testing problem in which the graphs are sampled from the graphon or latent space model (Hoff et al. 2002; Lovász 2012; Bollobás et al. 2007) and propose an asymptotically valid and consistent test procedure that also runs in time polynomial in n (see Section 3.4). We evaluate the performance of the proposed test procedures through simulations and show that they exhibit power even for moderate values of n and small values of $n^{-1}\|R\|_F$. We then apply these procedures to two real data experiments. In the first experiment, we analyze the electrical and chemical connectomes for the *C. elegans* worm and show that there are significant correlations between the two connectomes; this confirms the observation made in Chen et al. (2016) wherein the authors showed that, for their vertex nomination tasks, using both connectomes leads to better accuracy. In the second experiment, we analyze two Wikipedia hyperlink graphs constructed using documents on the same topics but in different languages and show that, by estimating the correlations between the graphs, we obtain more accurate links prediction.

In Chapter 4, we consider statistical inference on multiple graphs. Inference typically relies on effective low-dimensional representations of graphs, often achieved through spectral decompositions (Belkin and Niyogi 2003; Sussman et al. 2012). Our focus is also on node-aligned graphs that share a common vertex set. Tang et al. (2017a, 2014) have developed tests to assess the similarity between two such graphs with latent models. As the number of graphs increases, the complexity of inference grows, necessitating research that extends beyond simple pairwise comparisons. In this regard, Levin et al. (2017); Draves and Sussman (2020) have broadened the scope of these tests to include generalized multiple graphs through joint low-dimensional embedding with the Omnibus matrix. The approach enables them to access latent positions across all graphs and simultaneously infer issues related to these positions. Additionally, the embedding is particularly useful for various inferences, such as community detection, vertex classification, hypothesis testing, and anomaly detection (Pantazis et al. 2022; Jones and Rubin-Delanchy 2020; Chen et al. 2020). Despite these advances, most existing approaches presume that all graphs share the same distribution or have significantly

similar distributions. Specifically, the theoretical properties of the Omnibus embedding are established for each random adjacency matrix marginally distributed according to a random dot product graph model with the same latent positions or for a group of random dot product graph models with a specific structure, such as eigen-scaling random dot product graph models. Under these conditions, the column spaces of the latent positions or the probability matrices for all random graphs are identical. Additionally, the rank of combination of latent positions for all random graphs is kept as the dimension of the latent position, so it is sufficient to use the same dimension for the Omnibus embedding. Our work diverges from this norm by not imposing any distributional constraints across the graphs. Without such constraints among graphs, the column spaces can differ, and the rank can be larger. Therefore, we need an Omnibus embedding with a possible larger dimension to include the information for all graphs. We adopt a generalized random dot product graph model, presenting a more flexible and encompassing framework compared to those typically explored in the literature. This approach allows for a more nuanced exploration of the underlying structures and relationships within and between multiple graph datasets. Furthermore, we demonstrate the effectiveness of our method through extensive simulations and real data analysis, confirming its practical applicability and robustness.

In summary, graph models have widespread applications in statistics, machine learning, and biology. Understanding and analyzing these models is crucial for research in these fields. In the following chapters, we will explore the performance and effectiveness of these models in specific applications.

CHAPTER

2

TWO-SAMPLE HYPOTHESIS TEST FOR RANDOM GRAPHS

2.1 Introduction

In the previous chapter, we provided an overview of the fundamental theories and application background of graph models. This chapter delves into the application of graph models in two-sample hypothesis testing. Specifically, we focus on improving existing results in this field. We will introduce enhanced methodologies and demonstrate their effectiveness through theoretical analysis and simulation experiments.

Testing differences between graphs is a crucial aspect of statistical inference on random graphs. For instance, analyzing variations among brain graphs is a prominent research area in neuroscience and machine learning (Bullmore and Sporns 2009; Richiardi et al. 2011, 2013).

Tang et al. (2017a) introduces a semiparametric two-sample hypothesis testing procedure for random dot product graphs (RDPG), which effectively approximates a broad spectrum of random graphs, ranging from simple to complex (Athreya et al. 2018). This method assumes that the two random dot product graphs share the same vertex set with known vertex correspondence. The primary goal of these tests is to determine whether the two models share identical generating latent positions or whether their generating latent positions are scaled or

diagonally transformed versions of each other. Tang et al. (2017b) proposes statistics based on the estimated latent positions, with the statistics having a supremum limit not exceeding 1 under the null hypothesis.

However, the limited information provided by these statistics can lead to indeterminacy and conservatism in hypothesis testing. We improve the existing statistics by setting the limiting value to 1. Consequently, these statistics are more likely to be significant and asymptotically greater than 1 under the alternative hypothesis when the two graphs are independent. This improvement enables the detection of smaller differences between two graphs, transitioning from a "bounded" to an "approximating" approach. When the two graphs are not independent, the limiting value may vary, but it remains straightforward to calculate this limit under the condition that, for all pairs of vertices, the edges between them exhibit the same correlation across the two graphs.

2.2 Preliminary

2.2.1 Notation

For a positive integer n, set $[n] := \{1, 2, \ldots, n\}$.

For a matrix $M \in \mathbb{R}^{p_1 \times p_2}$, let $\sigma_i(M)$ denote the ith largest singular value of M, for $1 \leq i \leq \min\{p_1, p_2\}$. Let M_i denote the ith row of M, and m_{ij} denote the element in the i-th row and j-th column, for $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$.

Let $\mathcal{O}(d)$ representing the collection of orthonormal matrices in $\mathbb{R}^{d \times d}$, which means $\mathcal{O}(d) = \{W \in \mathbb{R}^{d \times d} : WW^T = W^TW = I\}$, where I is the identity matrix.

Let $f(n)$ and $g(n)$ be two real valued functions of n. Define $f(n) = O(g(n))$ if there exist a positive real number M and a real number n_0 such that for all $n > n_0$, $|f(n)| \leq M \cdot g(n)$. Define $f(n) = o(g(n))$ if for every positive real number m, there exists a real number N_0 such that for all $n > N_0$, $|f(n)| \leq m \cdot g(n)$. We also use $f(n) \ll g(n)$ to express $f(n) = o(g(n))$. Define $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$. Define $f(n) = \omega(g(n))$ if $g(n) = o(g(n))$. Define $f(n) = \theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$. In this paper, two functions with the notation are treated as functions of n without additional remarks. Besides, define $f(n) = O_{\mathbb{P}}(g(n))$ if $f(n) = O(g(n))$ with the probability at least $1 - n^{-C}$ for any $C > 0$. The analogous notation applies for $o_{\mathbb{P}}(\cdot), \omega_{\mathbb{P}}(\cdot)$. Also, define $f(n) \sim g(n)$ if $\frac{f(n)}{g(n)} \to 1$.

2.2.2 Background

The Random Dot Product Graph (RDPG) is a common random graph model, which can successfully approximate a wide range of random graphs, from simple to complex random

graphs. It is defined as the following.

Definition 1 (Random Dot Product Graph)**.** Let χ_d^n be defined by

$$\chi_d^n = \{M \in \mathbb{R}^{n\times d} : MM^T \in [0,1]^{n\times n} \text{ and } \operatorname{rank}(M) = d\},$$

and let $X = [X_1|\dots|X_n]^T \in \chi_d^n$. Suppose A is a random adjacency matrix given by

$$\mathbb{P}[A|X] = \prod_{i<j} (X_i^T X_j)^{A_{ij}} (1 - X_i^T X_j)^{1-A_{ij}},$$

then it is denoted by $A \sim \text{RDPG}(X)$ and say that A is the adjacency matrix of a random dot product graph with latent position X of rank at most d. Also, A can be presented as $A \sim \text{Bernoulli}(P)$, where $P = XX^T$. The matrix X represents the latent position.

Two-sample hypothesis testing for random graphs is widely used in many aspects, like neuroscience, social networks, and machine learning. We are focusing on a two-sample hypothesis testing problem for RDPG, where two random dot product graphs are on the same vertex set, with known vertex correspondence. The test aims to detect whether two models have the same generating latent position or have generating latent positions that are scaled or diagonal transformations of one another as below.

- (Equality, up to an orthogonal transformation)

$$H_0 : X =_W Y \qquad \text{vs.} \qquad H_a : X \neq_W Y;$$

- (Scaling)

$$\begin{aligned} &H_0 : X =_W cY \text{ for some } c > 0 \qquad \text{vs.} \\ &H_a : X \neq_W cY \text{ for any } c > 0; \end{aligned}$$

- (Diagonal transformation)

$$\begin{aligned} &H_0 : X =_W DY \text{ for some diagonal } D \qquad \text{vs.} \\ &H_a : X \neq_W DY \text{ for any diagonal } D, \end{aligned}$$

where $=_W$ denotes existence of an orthogonal matrix $W \in \mathbb{R}^{d\times d}$ such that $X = YW$, $X = cYW$ and $X = DYW$, respectively. Those three kinds of tests refer to testing $P = Q$, $P = cQ$ for some $c > 0$ or $P = DQD$ for some diagonal D, respectively. The three cases allow

us to test for different requirements. For example, for the stochastic block model (Holland et al. 1983), the equality case can test whether two models are the same, while the diagonal transformation case can test whether two models have the same structure, that is, the same block assignment.

For the three testing procedures, we have the following three test statistics,

$$
\begin{aligned}
T_1 &= \frac{\min\limits_{W\in\mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F}{\sqrt{d\gamma_2^{-1}(A)} + \sqrt{d\gamma_2^{-1}(B)}},\\
T_2 &= \frac{\min\limits_{W\in\mathcal{O}(d)} \|\frac{\hat{X}}{\|\hat{X}\|_F}W - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F}{2\sqrt{d\gamma_2^{-1}(A)}/\|\hat{X}\|_F + 2\sqrt{d\gamma_2^{-1}(B)}/\|\hat{Y}\|_F},\\
T_3 &= \frac{\min\limits_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F}{2\sqrt{d\gamma_2^{-1}(A)}\|\mathcal{D}^{-1}(\hat{X})\| + 2\sqrt{d\gamma_2^{-1}(B)}\|\mathcal{D}^{-1}(\hat{Y})\|}.
\end{aligned}
$$

Here $A \sim \text{RDPG}(X)$, $B \sim \text{RDPG}(Y)$. The estimators $\hat{X}$, $\hat{Y}$ are the adjacency spectral embedding (ASE) of A, B (Tang et al. 2017a). For a matrix M with singular values $\sigma_1 \geq \sigma_2 \geq \ldots$, we set

$$
\delta(M) = \max_{1\leq i\leq n}\sum_{j=1}^{n} M_{ij}; \qquad \gamma_1(M) = \min_{1\leq i\leq d}\frac{\sigma_i(M) - \sigma_{i+1}(M)}{\delta(M)}; \qquad \gamma_2(M) = \frac{\sigma_d(M) - \sigma_{d+1}(M)}{\delta(M)}.
$$

Also, we set $\mathcal{D}(M)$ as the diagonal matrix whose diagonal entries are the Euclidean norm of the rows of M and let $\mathcal{P}(M)$ be the matrix whose rows are the projection of the rows of M onto the unit sphere (Tang et al. 2017a). For example, if $M = \begin{pmatrix} 0.3 & 0.4 \\ 0.05 & 0.12 \\ 0.4 & 0.3 \end{pmatrix}$, then

$$
\mathcal{D}(M) = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.13 & 0 \\ 0 & 0 & 0.5 \end{pmatrix} \qquad \text{and} \qquad \mathcal{P}(M) = \begin{pmatrix} 3/5 & 4/5 \\ 5/13 & 12/13 \\ 4/5 & 3/5 \end{pmatrix}.
$$

We consider a sequence of such tests for $n \in \mathbb{N}$. For this, we have sequences $\{X_n\}$ and $\{Y_n\}$, where the latent positions X_n, Y_n do not need to be related to $X_{n'}, Y_{n'}$ for any $n' \neq n$. Additionally, sequences exist for all corresponding statistics. Tang et al. (2017a) yields that with high probability, for any $\epsilon > 0, i \in \{1, 2, 3\}$, statistics $T_i^{(n)} \leq 1 + \epsilon$ under H_0, provided that n is large enough. Consider the following definition.

Definition 2. For the three kinds of tests, some statistics T and associated rejection regions

R are consistent, asymptotically level α tests, which are defined as the following: if for any $\eta > 0$, there exists $N = N(\eta)$ such that

(1) if $n > N$ and H_a is true, then $\mathbb{P}(T \in R) > 1 - \eta$,

(2) if $n > N$ and H_0 is true, then $\mathbb{P}(T \in R) \leq \alpha + \eta$.

Tang et al. (2017a) show that the three statistics and associated rejection regions $\{T \geq 1 + \epsilon\}$ are consistent for the three kinds of tests above, respectively, if $d_n \to \infty$, where

- Equality Case: $d_n = \min_{W \in \mathcal{O}(d)} \|X_n W - Y_n\|_F$,
- Scaling Case: $d_n = \frac{\min_{W \in \mathcal{O}(d)} \| \frac{X_n}{\|X_n\|_F} W - \frac{Y_n}{\|Y_n\|_F} \|_F}{1/\|X_n\|_F + 1/\|Y_n\|_F}$,
- Diagonal Transformation: $d_n = \frac{\min_{W \in \mathcal{O}(d)} \|\mathcal{P}(X_n) W - \mathcal{P}(Y_n)\|_F}{\|\mathcal{D}^{-1}(X_n)\| + \|\mathcal{D}^{-1}(Y_n)\|}$.

2.3 Improvement

For a matrix $M \in \chi_d^n$, we define

$$C(M) = \mathrm{tr} S_{MM^T}^{-1/2} U_{MM^T}^T \mathbb{E}\big[\big(\text{Bernoulli}(MM^T) - MM^T\big)^2\big] U_{MM^T} S_{MM^T}^{-1/2},$$

where $MM^\top = U_{MM^T} S_{MM^T} U_{MM^T}^\top$ is the eigendecomposition of $MM^\top$. Then we know

$$C(X) = \mathrm{tr} S_P^{-1/2} U_P^T \mathbb{E}[(A - P)^2] U_P S_P^{-1/2}.$$

Tang et al. (2017a) show that if assuming that there exists a fixed $d \in \mathbb{N}$ such that for all n, P_n is of rank d with d distinct positive eigenvalues and there exist constants $\epsilon > 0$, $c_0 > 0$ and $n_0(\epsilon, c) \in \mathbb{N}$ such that for all $n \geq n_0$:

$$\gamma_1(P_n) \geq c_0; \qquad \delta(P_n) \geq (\log n)^{2+\epsilon},$$

there exists a deterministic sequence of orthogonal matrices W_n such that

$$\|\hat{X}_n - X_n W_n\|_F - \sqrt{C(X_n)} \xrightarrow{a.s.} 0,$$

and

$$\sqrt{C(X_n)} \leq \sqrt{d \gamma_2^{-1}(P_n)}; \qquad \sqrt{C(X_n)} = \Omega(1).$$

Consider the equality case, the above supremum limit result is also based on

$$\min_{W\in\mathcal{O}(d)} \|\hat{X}-\hat{Y}W\|_F \leq \min_{W_x\in\mathcal{O}(d)} \|\hat{X}-XW_x\|_F + \min_{W_y\in\mathcal{O}(d)} \|\hat{Y}-YW_y\|_F + \min_{W_0\in\mathcal{O}(d)} \|X-YW_0\|_F.$$

Furthermore, in most conditions, the difference between them is away from zero as $n \to \infty$. The other two cases have the similar inequality relation.

The inequality relationship leads to Supremum limits rather than limits. The test statistics have only a Supremum limit which is not greater than 1, which could lead to very conservative results if using 1 as the threshold in testing procedures. If we could get an equal relation, it is possible to derive statistics having a known limit.

If assuming A and B are independent, we found the following stronger relationship,

$$\min_{W\in\mathcal{O}(d)} \|\hat{X}-\hat{Y}W\|_F^2 \approx \min_{W_x\in\mathcal{O}(d)} \|\hat{X}-XW_x\|_F^2 + \min_{W_y\in\mathcal{O}(d)} \|\hat{Y}-YW_y\|_F^2 + \min_{W_0\in\mathcal{O}(d)} \|X-YW_0\|_F^2.$$

Furthermore, the difference between them converges to 0 almost surely. Based on the relation between $\min_{W_x\in\mathcal{O}(d)} \|\hat{X}-XW_x\|_F^2$ and $C(X)$ as well as the relation between $\min_{W_y\in\mathcal{O}(d)} \|\hat{Y}-YW_y\|_F^2$ and $C(Y)$, we have

$$\min_{W\in\mathcal{O}(d)} \|\hat{X}-\hat{Y}W\|_F^2 \approx C(X)+C(Y)+\min_{W_0\in\mathcal{O}(d)} \|X-YW_0\|_F^2. \tag{2.1}$$

Then we can consider new statistics:

$$\begin{aligned}
T_1 &= \frac{\min_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2}{C(\hat{X})+C(\hat{Y})},\\
T_2 &= \frac{\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}}{\|\hat{X}\|_F}W-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2}{C(\hat{X})/\|\hat{X}\|_F^2+C(\hat{Y})/\|\hat{Y}\|_F^2},\\
T_3 &= \frac{\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2}{C^*(\hat{X})+C^*(\hat{Y})}.
\end{aligned}$$

The three statistics and associated rejection regions $\{T \geq 1+\epsilon\}$ are consistent for the three kinds of tests above, respectively, if $\liminf d_n > 0$.

When A and B are not independent, there would be another difference between two sides in Equation 2.1. Under some conditions, such as for all pairs of vertices, the edges between them exhibit the same correlation across the two graphs, we can derive the difference so that we can also obtain the limit of the statistics under H_0.

2.4 Results

Recall the model, $X \in \chi_d^n$ and $Y \in \chi_d^n$. $A \sim \text{Bernoulli}(P)$, where $P = XX^T$, i.e., $A \sim \text{RDPG}(X)$. $B \sim \text{Bernoulli}(Q)$, where $Q = YY^T$, i.e., $B \sim \text{RDPG}(Y)$. Also, A_{ij} and B_{ij} are independent for any $1 \leq i < j \leq n$. There exists a fixed $d \in \mathbb{N}$ such that for all n, P is of rank d with d distinct positive eigenvalues and there exist constants $\epsilon > 0$, $c_0 > 0$ and $n_0(\epsilon, c) \in \mathbb{N}$ such that for all $n \geq n_0$:

$$\gamma_1(P) \geq c_0; \qquad \delta(P) \geq (\log n)^{2+\epsilon}.$$

The first condition ensures that the smallest non-zero eigenvalues of P is large enough and the non-zero eigenvalues are significantly distinct. The second condition ensure that the maximum expected degree of a graph is large enough.

Given a graph with adjacency matrix A, the adjacency spectral embedding (ASE) of A into R^d is defined by $\hat{X} = \text{ASE}(A) = U_A S_A^{1/2}$ where

$$|A| = [U_A|\tilde{U}_A] \begin{pmatrix} S_A & 0 \\ 0 & \tilde{S}_A \end{pmatrix} [U_A|\tilde{U}_A]$$

is the eigendecomposition of $|A| = (A^T A)^{1/2}$ in which S_A is the matrix of the d largest eigenvalues of $|A|$, $\tilde{S}_A$ is for the remaining eigenvalues and U_A, $\tilde{U}_A$ are the matrices whose columns are the corresponding eigenvectors. Also let $P = U_P S_P U_P^T$ be the eigendecomposition of P with S_P being the $d \times d$ diagonal matrix of nonzero eigenvalues of P.

For a matrix $M \in \chi_d^n$, we define

$$\begin{aligned} C(M) &= \text{tr} S_{MM^T}^{-1/2} U_{MM^T}^T \mathbb{E}\big[\big(\text{Bernoulli}(MM^T) - MM^T\big)^2\big] U_{MM^T} S_{MM^T}^{-1/2}, \\ C^*(M) &= \mathbb{E}[\text{tr} S_{MM^T}^{-1/2} U_{MM^T}^T (\text{Bernoulli}(MM^T) - MM^T) D^{-2} (\text{Bernoulli}(MM^T) - MM^T) U_{MM^T} S_{MM^T}^{-1/2}], \\ D &= \mathcal{D}(X). \end{aligned}$$

Then we have the following results.

Theorem 1. In our setting, with the probability at least $1 - n^C$ for any $C > 0$,
under H_0,

$$\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \sim 2C(\hat{X});$$

under H_1, assuming $\min_{W\in\mathcal{O}(d)}\|XW-Y\|_F^2=\mathbb{O}(1)$,

$$\min_{W\in\mathcal{O}(d)}\|\hat{X}W-\hat{Y}\|_F^2\sim C(\hat{X})+C(\hat{Y})+\min_{W\in\mathcal{O}(d)}\|XW-Y\|_F^2.$$

Theorem 2. In our setting, with the probability at least $1-n^C$ for any $C>0$, under H_0,

$$\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2\sim\frac{2C(\hat{X})}{\|\hat{X}\|_F^2};$$

under H_1, assuming $\min_{W\in\mathcal{O}(d)}\|\frac{XW}{\|X\|_F}-\frac{Y}{\|Y\|_F}\|_F^2=\mathbb{O}((\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})^2)$,

$$\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2\sim\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}+\min_{W\in\mathcal{O}(d)}\|\frac{XW}{\|X\|_F}-\frac{Y}{\|Y\|_F}\|_F^2.$$

Following Tang et al. (2017a), we modify the assumption for the diagonal case. There exist constants $\epsilon_1,\epsilon_2,\epsilon_3>0$, $c_0>0$ and $n_0(\epsilon_1,\epsilon_2,\epsilon_3,c)\in\mathbb{N}$ such that for all $n\geq n_0$:

$$\gamma_1(P)\geq c_0;\quad \delta(P)\geq n^{1/2}(\log n)^{\epsilon_1};\quad \min_i\|X_i\|\geq(\frac{\log n}{\sqrt{\delta(P)}})^{1-\epsilon_2};\quad \frac{\max\|X_i\|}{\min\|X_i\|}\leq n^{1/4-\epsilon_3}.$$

The first new assumption ensures that X_i is not too small, preventing vertices from having excessively low degrees. The second new assumption ensures that all X_i values are of similar magnitude, so that the degrees of all vertices are close.

Theorem 3. In our setting, with the probability at least $1-n^C$ for any $C>0$, under H_0,

$$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\sim 2C^*(\hat{X});$$

under H_1, assuming $\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F^2=\mathbb{O}((\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)^2)$,

$$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\sim C^*(\hat{X})+C^*(\hat{Y})+\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F^2,$$

where $C^*(X)=\mathbb{E}[\mathrm{tr}S_P^{-1/2}U_P^T(A-P)D^{-2}(A-P)U_PS_P^{-1/2}]$, $D=\mathcal{D}(X)$.

2.5 Tests

For Equality Case, based on Theorem 1, we have under H_0 with high probability

$$\frac{\min\limits_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2}{C(\hat{X})+C(\hat{Y})} \to 1,$$

while under H_A with high probability

$$\frac{\min\limits_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2}{C(\hat{X})+C(\hat{Y})} \to 1+c,$$

for some $c \in (0,\infty]$, when $\min_{W\in\mathcal{O}(d)} \|XW-Y\|_F = \Omega(1)$, because

$$\frac{\min\limits_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2}{C(\hat{X})+C(\hat{Y})} - (1+\frac{\min_{W\in\mathcal{O}(d)} \|XW-Y\|_F^2}{C(X)+C(Y)}) \to 0.$$

We can apply Theorem 1 when $\min_{W\in\mathcal{O}(d)} \|XW-Y\|_F = \theta(1)$. Additionally, we can extend this result to the case where $\min_{W\in\mathcal{O}(d)} \|XW-Y\|_F = \omega(1)$, and also obtain $\min\limits_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2 \to \infty$.

With the above results, the test procedure is consistent if $\min_{W\in\mathcal{O}(d)} \|XW-Y\|_F = \Omega(1)$ as the description in Theorem 2 in Tang et al. (2017a). This is an improved result compared to the one in Tang et al. (2017a). $\Omega(1)$ is a weaker condition than $\omega(1)$.

For Scaling Case, with the similar way, based on Theorem 2, we have under H_0 with high probability

$$\frac{\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2}{\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}} \to 1,$$

while under H_A with high probability

$$\frac{\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2}{\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}} \to 1+c,$$

for some $c \in (0,\infty]$, when $\min_{W\in\mathcal{O}(d)} \|\frac{XW}{\|X\|_F}-\frac{Y}{\|Y\|_F}\|_F = \Omega(\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})$, because

$$\frac{\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2}{\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}} - (1+\frac{\min_{W\in\mathcal{O}(d)} \|\frac{XW}{\|X\|_F}-\frac{Y}{\|Y\|_F}\|_F^2}{\frac{C(X)}{\|X\|_F^2}+\frac{C(Y)}{\|Y\|_F^2}}) \to 0.$$

With the above results, the test procedure is consistent if $\min_{W\in\mathcal{O}(d)}\|XW-Y\|_F = \Omega(\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})$ as the description in Theorem 3 in Tang et al. (2017a). This is an improved result compared to the one in Tang et al. (2017a). $\Omega(\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})$ is a weaker condition than $\omega(\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})$.

For Diagonal Transformation Case, with the similar way, based on Theorem 3, we have under H_0 with high probability

$$\frac{\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2}{C^*(\hat{X})+C^*(\hat{Y})}\to 1,$$

while under H_A with high probability

$$\frac{\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2}{C^*(\hat{X})+C^*(\hat{Y})}\to 1+c,$$

for some $c\in(0,\infty]$, when $\min\limits_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F = \Omega(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)$, because

$$\frac{\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2}{C^*(\hat{X})+C^*(\hat{Y})}-(1+\frac{\min\limits_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F^2}{C^*(X)+C^*(Y)})\to 0.$$

With the above results, the test procedure is consistent if $\min\limits_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F = \Omega(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)$ as the description in Theorem 4 in Tang et al. (2017a). This is an improved result compared to the one in Tang et al. (2017a). $\Omega(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)$ is a weaker condition than $\omega(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)$.

2.6 Dependence Case

In the previous procedures, we assume A and B are independent. The independence between A and B ensures the following crucial part. We have with high probability,

$$\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y)\to 0,$$

where $W_x := \arg\min_{W\in\mathcal{O}(d)}\|XW-\hat{X}\|_F^2$, $W_y := \arg\min_{W\in\mathcal{O}(d)}\|YW-\hat{Y}\|_F^2$. The details can be find in the proof part of Lemma 9.

We also know that

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)} \|\hat{X}-\hat{Y}W\|_F^2 \\
\approx &\min_{W_x\in\mathcal{O}(d)} \|\hat{X}-XW_x\|_F^2 + \min_{W_y\in\mathcal{O}(d)} \|\hat{Y}-YW_y\|_F^2 + \min_{W_0\in\mathcal{O}(d)} \|X-YW_0\|_F^2 \\
&- 2\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y).
\end{aligned}$$

Now we assume A_{jk} and B_{jk} have a correlation r_{jk}. In general, as r_{jk}'s increase, $\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y)$ increase, so that the statistics will decrease.

Furthermore, if $P=Q$ (under H_0), and all $r_{jk}=r$ $(i\neq j)$, we have with high probability,

$$\begin{aligned}
&\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y) \\
\sim &\mathrm{tr}S_P^{-1/2}U_P^T(A-P)(B-P)U_PS_P^{-1/2} \\
\sim &\mathrm{tr}S_P^{-1/2}U_P^T\mathbb{E}[(A-P)(B-P)]U_PS_P^{-1/2} \\
\sim &\mathrm{tr}S_P^{-1/2}U_P^Tr\mathbb{E}(A-P)^2U_PS_P^{-1/2} \\
\sim &r\cdot C(X).
\end{aligned}$$

Then under H_0,

$$\min_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2 \sim 2(1-r)C(X).$$

So under H_0,

$$T_{\mathrm{new}} \sim 1-r.$$

On this condition, we still have consistent test procedures even A and B are dependent.

2.7 Simulations

2.7.1 Convergence

We now conduct simulation experiments to evaluate the performance of our statistics under H_0 in Equality case

$$T_{\mathrm{new}} = \frac{\min_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F^2}{C(\hat{X})+C(\hat{Y})},$$

compared with

$$T_{\mathrm{old}} = \frac{\min_{W\in\mathcal{O}(d)} \|\hat{X}W-\hat{Y}\|_F}{\sqrt{d\gamma_2^{-1}(A)}+\sqrt{d\gamma_2^{-1}(B)}}.$$

We generate $\{X_i\}_{i=1}^n$ as iid sample from a bivariate normal with mean 0 and identity covariance matrix. We then consider the link function of the cosine similarity for P as below,

$$P_{ij} = \frac{|X_i^\top X_j|}{2\|X_i\|\|X_j\|}.$$

Then we have $A, B \overset{\text{iid}}{\sim} \text{RDPG}(X)$, $\hat{X} = \text{ASE}(A)$, and $\hat{Y} = \text{ASE}(B)$. We repeat 100 times to obtain sampled statistics. The results are presented in Figs 2.1 and 2.2 for $n \in \{100, 200, 500, 1000, 2000\}$. We observe that T_{new} significantly converges to 1 while T_{old} does not and remains significantly less than 1.

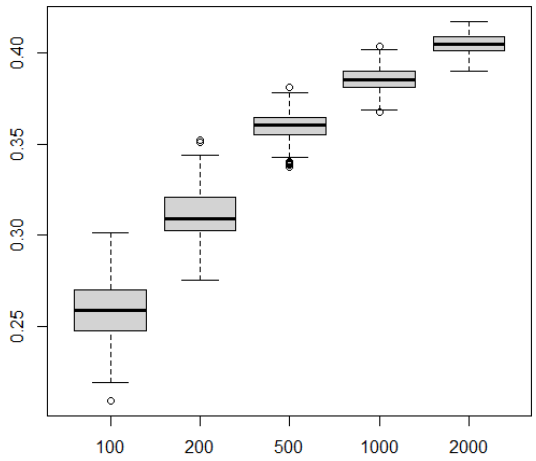


Figure 2.1: T_{old}

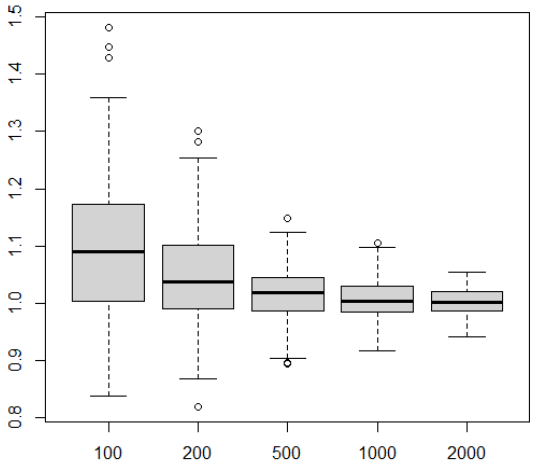


Figure 2.2: T_{new}

Also, we consider dependence case, assuming all $r_{ij} = r$ for $i \neq j$. We generate $A_{ij}, B_{ij} \sim$ Bernoulli(P_{ij}) with correlation r for $i < j$, and then complete A, B with symmetry and 0 in diagonal. Utilizing the same procedure, the new statistics are presented in Figs 2.3 and 2.4 for the combination of $n \in \{100, 200, 500, 1000, 2000\}$ and $r \in \{0.1, 0.5\}$. We observe that $T_{0.1}$ significantly converges to $1 - r = 0.9$ and $T_{0.5}$ significantly converges to $1 - r = 0.5$.

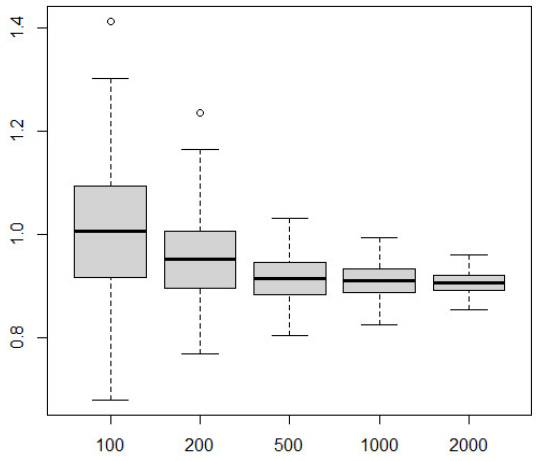


Figure 2.3: $T_{0.1}$, where A_{ij} and B_{ij} have the correlation of 0.1 for all $i < j$.

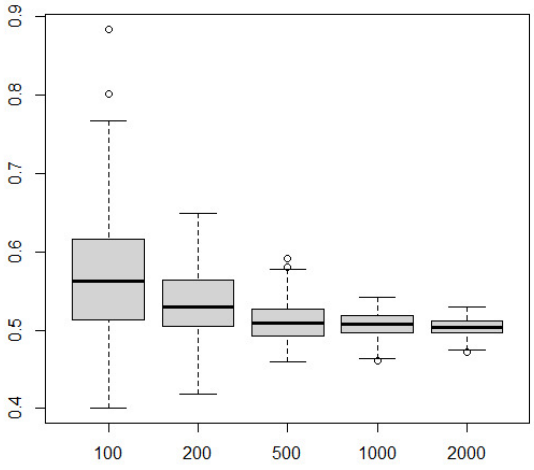


Figure 2.4: $T_{0.5}$, where A_{ij} and B_{ij} have the correlation of 0.5 for all $i < j$.

2.7.2 Hypothesis Testing

In this section, we conduct related simulations to compare the performance of the testing procedures between new statistics and old statistics in Equality case. We follow the model in Section 4.1 in Tang et al. (2017a) and start with the same setting. Here we use the stochastic block model (SBM) (Holland et al. 1983), which we know is also a specific RDPG under some conditions.

Recall a simple definition of SBM.

Definition 3 (Stochastic Block Model)**.** Let B be a symmetric $K \times K$ matrix with entries in $[0, 1]$. An undirected graph G on n vertices is a stochastic block model (SBM) graph with block probabilities B and blocks assignment $\tau : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ if G has symmetric adjacency matrix A where $a_{ij} \sim \text{Bernoulli}(b_{\tau(i),\tau(j)})$ independently, for $1 \leq i < j \leq n$, and denoted by $A \sim \text{SBM}(B, \tau)$.

Now we know that the $n \times K$ matrix Z whose elements $z_{ij} = 1$ if $\tau(i) = k$ and $z_{ij} = 0$ otherwise denotes the assignments situation. Then the probability matrix $P = \rho ZBZ^T$ and $P_{ij} = \rho b_{\tau(i),\tau(j)}$. So we know if B is positive semi-definite, the SBM can be a special case of RDPG.

We consider the problem of testing the null hypothesis $H_0 : X =_W Y$ against $H_A : X \neq_W Y$. We consider random graphs generated according to two stochastic block models with the

same block membership but different block probability matrices B_0 and B_ϵ for $\epsilon \geq 0$, where $B_\epsilon = 0.2 \times 11^\top + (0.3+\epsilon)I$, 1 is 2×1 matrix with all entries equaling to 1, and I is the 2×2 identity matrix. That is,

$$B_\epsilon = B_0 = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix} \quad \text{under } H_0; \qquad B_\epsilon = \begin{pmatrix} 0.5+\epsilon & 0.2 \\ 0.2 & 0.5+\epsilon \end{pmatrix} \quad \text{under } H_A.$$

Each vertex is assigned to some a block with the same probability 0.5.

We then test the hypothesis $H_0 : X =_W Y^{(\epsilon)}$ against $H_A : X \neq_W Y^{(\epsilon)}$, where X corresponds to B_0 and $Y^{(\epsilon)}$ corresponds to B_ϵ. We evaluate the performance of the test procedures by estimating the level and power of the test statistics through Monte Carlo simulation. We first consider the rejection region $\{T > 1\}$ as dictated by the asymptotic theory: $\limsup_{n\to\infty} T_{\text{old}} \leq 1$ and $\lim_{n\to\infty} T_{\text{new}} = 1$ under H_0. We conduct 100 Monte Carlo replicates to estimate the power of the test procedures. The results are presented in Table 2.1 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $\epsilon \in \{0, 0.05, 0.1, 0.2\}$. We observe that the old statistics are much more conservative than new statistics. Also, the significance level α (for $\epsilon = 0$) of the new test statistic significantly converges to 0.5, which satisfies the performance of Fig 2.2.

Table 2.1: Rejection Region: $\{T > 1\}$. Elements in tables are corresponding powers

	$\epsilon = 0$		$\epsilon = 0.05$		$\epsilon = 0.1$		$\epsilon = 0.2$	
Statistics	old	new	old	new	old	new	old	new
n=100	0	0.78	0	0.80	0	0.98	0	1
n=200	0	0.67	0	0.93	0	1	0	1
n=500	0	0.66	0	1	0	1	0	1
n=1000	0	0.62	0	1	0	1	0.71	1
n=2000	0	0.54	0	1	0	1	1	1

We also know that the rejection region $\{T > 1+\epsilon\}$ for any $\epsilon > 0$ is consistent theoretically. Now we consider the rejection region $\{T > 1 + 0.05\}$. The results are presented in Table 2.2 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $\epsilon \in \{0, 0.05, 0.1, 0.2\}$. Besides the previous performance, we also observe that the significance level α (for $\epsilon = 0$) of the new test statistic significantly converges to 0, which means that it is well controlled. Still, the new test has good power.

Table 2.2: Rejection Region: $\{T > 1.05\}$. Elements in tables are corresponding powers

	$\epsilon = 0$		$\epsilon = 0.05$		$\epsilon = 0.1$		$\epsilon = 0.2$	
Statistics	old	new	old	new	old	new	old	new
n=100	0	0.61	0	0.67	0	0.97	0	1
n=200	0	0.52	0	0.79	0	1	0	1
n=500	0	0.23	0	1	0	1	0	1
n=1000	0	0.14	0	1	0	1	0	1
n=2000	0	0.02	0	1	0	1	1	1

2.8 Discussion

In this chapter, we explored the two-sample hypothesis testing problem for random dot product graphs (RDPG), improving upon the existing methodologies in the field. By addressing the limitations of previous statistics and introducing new test statistics with better performance, we have advanced the ability to detect differences between two random graphs.

Our new statistics are designed to be more powerful under alternative hypotheses. Through both theoretical analysis and simulation experiments, we demonstrated that the new statistics converge to known limits, enhancing the reliability and accuracy of hypothesis testing in random graphs.

The application of the new test statistics to real-world data, such as brain graphs in neuroscience, could yield more accurate and insightful results. The enhanced ability to detect smaller differences between two graphs is particularly useful in fields like neuroscience and social network analysis, where subtle variations can be crucial.

In summary, this chapter contributes to the field of statistical inference on random graphs by providing improved methodologies for two-sample hypothesis testing. The new test statistics offer better performance, consistency, and applicability, making them valuable tools for researchers and practitioners working with random graphs. Future research could explore the distributions of statistics when the correlation between two graphs are more complicated and also seek to weaken our assumptions, such as without distinct eigenvalues or known vertex correspondence.

CHAPTER

3

INDEPENDENCE TESTING FOR INHOMOGENEOUS RANDOM GRAPHS

3.1 Introduction

In the previous chapter, we discussed the basic concepts and applications of two-sample hypothesis testing. This chapter will shift the focus to another important issue—independence testing. We will introduce the theoretical foundations, algorithm implementation, and demonstrate its application through real data analysis.

Detecting correlation and testing for independence between Euclidean vectors is one of the most classical and widely studied inference problems in multivariate statistics. See Anderson (2003, Chapter 11) and Kendall (1990) for standard references when the data are low-dimensional and Leung and Drton (2018); Gretton et al. (2007); Székely et al. (2007) for some examples of recent results in the high-dimensional setting. In contrast to the above mentioned literature, independence testing for graphs is much less studied. Part of the difficulty lies in choosing an appropriate notion of correlation for graph-valued data. In this dissertation, we are motivated by the problem of, given two graphs on the same vertex set, determining whether or not the presence of an edge between any pair of vertices in the first graph is *stochastically* independent of the presence of the corresponding edge in the

second graph. Two graphs are thus said to be independent if all edges in the first are *pairwise* independent of the corresponding edges in the second, and are said to be correlated otherwise.

The above notion of pairwise edge correlation, while rather simple, does appear in many real data applications. For example, two networks on different social platforms are likely to share a common subset of users whose induced sub-networks are correlated in that two users are more likely to be linked in one social platform if they are already linked in the other social platform. As another example, the graphs constructed from the wiring diagrams of the mushroom body for the left and right hemispheres of the *Drosophila* larva are highly correlated as a pair of neurons are more likely to send signals to each other in the left hemisphere if their correspondences also send signals to each other in the right hemisphere (Eichler et al. 2017; Winding et al. 2023). Finally, two knowledge graphs constructed using documents on the same topics but written in different languages are also highly correlated (Ma et al. 2012; Haghighi et al. 2005).

In this dissertation, we consider independence testing in the setting where the graphs are, *marginally*, inhomogeneous Erdős-Rényi random graphs while the pairwise edges correlations are described by the entries of a symmetric matrix R (see Definition 4). Two graphs are then independent if $R_{ij} \equiv 0$ for all $\{i, j\}$ and are correlated if $|R_{ij}| > 0$ for *some* $\{i, j\}$. Equivalently, we are interested in testing the null hypothesis $\mathbb{H}_0 \colon \|R\|_F = 0$ against the alternative hypothesis $\mathbb{H}_A \colon \|R\|_F > 0$.

Our contributions are as follows. We first show that there exists a procedure for deciding between $\mathbb{H}_0$ and $\mathbb{H}_A$ with asymptotically *vanishing* type-I and type-II errors only if $\|R\|_F \to \infty$ as $n \to \infty$ (see Section 3.3). We also describe two related examples where the condition $\|R\|_F \to \infty$ is sufficient for one example but not sufficient for the other. We next show that the problem exhibits a statistical vs. computational tradeoff, i.e., there are regimes for which $\|R\| \to \infty$ that are statistically detectable but may require running time which scales exponentially with n. We achieve this through a polynomial time reduction of the planted clique problem, which is well-known to be computationally hard (Alon et al. 1998; Barak et al. 2019), to our correlation testing problem (see Theorem 5). Finally, we consider a special case of our correlation testing problem in which the graphs are sampled from the graphon or latent space model (Hoff et al. 2002; Lovász 2012; Bollobás et al. 2007) and propose an asymptotically valid and consistent test procedure that also runs in time polynomial in n (see Section 3.4).

We evaluate the performance of the proposed test procedures through simulations and show that they exhibit power even for moderate values of n and small values of $n^{-1}\|R\|_F$. We then apply these procedures to two real data experiments. In the first experiment, we analyze the electrical and chemical connectomes for the *C. elegans* worm and show that there

are significant correlations between the two connectomes; this confirms the observation made in Chen et al. (2016) wherein the authors showed that, for their vertex nomination tasks, using both connectomes leads to better accuracy. In the second experiment, we analyze two Wikipedia hyperlink graphs constructed using documents on the same topics but in different languages and show that, by estimating the correlations between the graphs, we obtain more accurate links prediction.

3.1.1 Related Works

Existing research on independence testing for a pair of graphs is quite limited. Among the current literature, only Xiong et al. (2019) considered independence testing where the notion of correlation is similar to that described here, but the setting in Xiong et al. (2019) is much more restrictive as they assumed that the pairwise correlations are the same for all edges, i.e., $R_{ij} \equiv c$ for some constant c and furthermore each observed graph is, *marginally*, distributed according to a stochastic blockmodel (Holland et al. 1983). Their results will be a special case of Theorem 7 in this dissertation.

Detecting pairwise edges correlation between two graphs is also a central focus in many graph matching algorithms. More specifically, given a pair of *unlabeled* adjacency matrices A and B, graph matching aims to find a mapping between their vertex sets that best preserves their common structures, for example by minimizing the Frobenius norm error $\|\Pi A \Pi^\top - B\|_F$ over all *permutation* matrices Π. Recent results show that many graph matching algorithms have *average* running times that are polynomial in n, the number of vertices, whenever the graphs are sufficiently correlated; see e.g., (Wu et al. 2023; Lyzinski and Sussman 2020; Lyzinski et al. 2016; Pedarsani and Grossglauser 2011; Onaran et al. 2016; Korula and Lattanzi 2014) and the references therein. In contrast, the worst case running time for these algorithms could be exponential in n if the graphs are independent. Because the correspondence between the vertex sets is assumed *unknown* in graph matching but is assumed known in the context of the dissertation, our technical challenges and results are quite different from those for graph matching.

3.2 Background and Setting

We now formally introduce the hypothesis testing problem considered in this dissertation. We begin by describing the notion of edge correlated inhomogeneous Erdős-Rényi graphs. Note that all graphs considered in this dissertation are simple, undirected graphs, i.e., there are no self-loops and no multiple edges.

Definition 4. Let $n \in \mathbb{N}$. Let $P \in [0,1]^{n\times n}$ and $R \in [-1,1]^{n\times n}$ be *symmetric* matrices where R satisfies the constraint

$$R_{ij} \geq -\min\Big\{\frac{1-P_{ij}}{P_{ij}}, \frac{P_{ij}}{1-P_{ij}}\Big\}, \quad \text{for all } i,j. \tag{3.1}$$

We say that (A,B) are R-correlated heterogeneous Erdős-Rényi graphs on n vertices with probability matrix P and correlation matrix R, denoted by $(A,B) \sim R\text{-ER}(P)$, if

1. A is the adjacency matrix for an inhomogeneous Erdős-Rényi graph on n vertices with probability matrix P, that is, A is a $n \times n$ symmetric binary matrix whose (upper triangular) entries are independent Bernoulli random variables with success probabilities $\{P_{ij}\}_{i<j}$.

2. B is the adjacency matrix for another inhomogeneous Erdős-Rényi graph on n vertices with probability matrix P.

3. The pairs $\{(A_{ij}, B_{ij})\}_{i<j}$ are *mutually independent* bivariate random variables. Here A_{ij} and B_{ij} are the ijth entry of A and B, respectively.

4. For any $i < j$, A_{ij} and B_{ij} are correlated with Pearson correlation R_{ij}.

Remark 1. Definition 4 assumes that A and B have the same marginal distribution. In Section 3.4.3 we will consider a slight extension of this model where we allow for A and B to have different marginal distributions. More specifically, we say that (A,B) are generated from the R-ER(P,Q) model if conditions 1 and 2 in Definition 4 are replaced by the assumption that A is marginally P and B is marginally Q, respectively. The remaining conditions 3 and 4 remain unchanged. Let $\phi(x) = \sqrt{x/(1-x)}$ for $x \in (0,1)$. We then note that for the correlation R_{ij} to be valid in the R-ER(P,Q) model, we assume

$$-\min\Big\{\phi(P_{ij})\phi(Q_{ij}), \frac{1}{\phi(P_{ij})\phi(Q_{ij})}\Big\} \leq R_{ij} \leq \min\Big\{\frac{\phi(P_{ij})}{\phi(Q_{ij})}, \frac{\phi(Q_{ij})}{\phi(P_{ij})}\Big\} \tag{3.2}$$

as a replacement for the condition $R_{ij} \geq -\min\{\frac{1-P_{ij}}{P_{ij}}, \frac{P_{ij}}{1-P_{ij}}\}$ in Definition 4.

Correlated inhomogeneous ER graphs are widely studied in the graph matching literature, see e.g., Fishkind et al. (2019); Sussman et al. (2020); Lyzinski and Sussman (2020) and the references therein. However, as we allude to in the introduction, for graph matching the correlation matrix R is usually assumed to be a constant matrix, i.e., $R_{ij} \equiv c$ for all $\{i,j\}$.

Let $\mathbb{A}_n$ denote the set of $n \times n$ hollow symmetric binary matrices. Given a pair of R-ER(P) graphs and symmetric matrices $[a_{ij}]$, $[b_{ij}] \in \mathbb{A}_n$, the joint likelihood for the adjacency matrices

$A = [A_{ij}]$ and $B = [B_{ij}]$ is

$$\mathbb{P}(A = [a_{ij}], B = [b_{ij}]) = \prod_{i<j} \mathbb{P}(A_{ij} = a_{ij}, B_{ij} = b_{ij})$$

where $\mathbb{P}(A_{ij} = a, B_{ij} = b)$ for $1 \leq i < j \leq n$ is given by

$$\mathbb{P}(A_{ij} = a, B_{ij} = b) = \begin{cases} P_{ij}^2 + P_{ij}(1 - P_{ij})R_{ij}, & a = b = 1 \\ (1 - P_{ij})^2 + P_{ij}(1 - P_{ij})R_{ij}, & a = b = 0, \\ P_{ij}(1 - P_{ij})(1 - R_{ij}), & a \neq b. \end{cases} \tag{3.3}$$

Let (A, B) be a pair of R-ER(P) graphs with unknown correlation matrix R and edge probabilities matrix P. We want to test the hypotheses $\mathbb{H}_0 \colon R = 0$ against $\mathbb{H}_A \colon R \neq 0$. This is equivalent to testing

$$\mathbb{H}_0 : \|R\|_* = 0 \quad \text{against} \quad \mathbb{H}_A : \|R\|_* > 0. \tag{3.4}$$

for any choice of matrix norm $\|\cdot\|_*$. We will use the Frobenius norm as it is one of the simplest and most widely used norms while also yielding useful and interesting theoretical results in the setting of the dissertation; see in particular Theorem 4 below. We expect that other norms, such as the spectral or infinity norms, will lead to results that are related but distinct from those presented here and we leave this for future work.

Let T be any test procedure for the hypothesis in Eq. (3.4). A desirable property for T is consistency, that is, as the number of vertices n approaches infinity, we want T to have *both* vanishing Type-I error and vanishing Type-II error. More specifically, along a sequence of edge probabilities matrices and correlation matrices (P_n, R_n), a test procedure is consistent if its error rate converges to 0 as $n \to \infty$, that is

$$\lim_{n\to\infty} \Big(1 - \mathbb{P}(\text{reject } \mathbb{H}_0 \mid \|R_n\|_F > 0) + \mathbb{P}(\text{reject } \mathbb{H}_0 \mid \|R_n\|_F = 0)\Big) = 0.$$

3.3 Statistically Limit

3.3.1 Detectability Threshold

We first derive a necessary condition for the hypothesis testing problem in Eq.(3.4) to be statistically detectable. Our approach is based on the second moment method for the ratio of the likelihood when $\|R\|_F = 0$ against the likelihood when $\|R\|_F > 0$. See Wu and Xu (2021) for an elegant survey of the second moment method and its application in deriving

detectability thresholds for statistical problems with planted structures.

The second moment method is motivated by the notion of contiguity between distributions. Contiguity is an asymptotic generalization of absolutely continuity characterized as follows.

Definition 5. Let $\{\mathcal{P}_n\}_{n\geq 1}$ and $\{\mathcal{Q}_n\}_{n\geq 1}$ be two sequence of probability measures. We say that $\{\mathcal{P}_n\}$ is contiguous with respect to $\{\mathcal{Q}_n\}$ if $\mathcal{Q}_n(S_n) \to 0$ implies $\mathcal{P}_n(S_n) \to 0$ for every sequence of measurable sets S_n.

If the probability distribution under the alternative hypothesis is contiguous with respect to the distribution under the null hypothesis, then there does not exist a valid and consistent test procedure. Indeed, due to contiguity, any rejection region with vanishing type I error under the null hypothesis will necessarily have vanishing power under the alternative hypothesis. For more on the definition of contiguity and its properties, see Van der Vaart (2000).

Let $\{\mathcal{P}_n\}_{n\geq 1}$ and $\{\mathcal{Q}_n\}_{n\geq 1}$ be the sequences of *joint distributions* for pairs of adjacency matrices $(A_n, B_n)_{n\geq 1}$ from the R-ER(P) random graphs model with $\|R\|_F > 0$ (for $\mathcal{P}_n$) and $\|R\|_F = 0$ (for $\mathcal{Q}_n$), respectively. We emphasize that, for simplicity of notations, we have dropped the index n from the matrices R and P.

It is well known that $\{\mathcal{P}_n\}$ is contiguous with respect to $\{\mathcal{Q}_n\}$ (see e.g., Eq.(13.3) of Wu and Xu (2021)) if the second moment of the likelihood ratio between $\mathcal{P}_n$ and $\mathcal{Q}_n$ is bounded i.e., that

$$\limsup_{n\to\infty} \mathbb{E}_{(A_n,B_n)\sim\mathcal{Q}_n}\Big[\Big(\frac{\mathcal{P}_n(A_n,B_n)}{\mathcal{Q}_n(A_n,B_n)}\Big)^2\Big] < \infty.$$

We then have the following result (see the appendix for a proof).

Theorem 4. Let $\{\mathcal{P}_n\}_{n\geq 1}$ be the sequence of *joint distributions* for pairs of adjacency matrices $(A_n, B_n)_{n\geq 1}$ from the R-ER(P) random graphs model with $\|R\|_F > 0$. Similarly, let $\{\mathcal{Q}_n\}$ be the joint distribution of the $(A_n, B_n)_{n\geq 1}$ when $\|R\|_F = 0$. Then

$$\mathbb{E}_{(A_n,B_n)\sim\mathcal{Q}_n}\Big[\Big(\frac{\mathcal{P}_n(A_n,B_n)}{\mathcal{Q}_n(A_n,B_n)}\Big)^2\Big] = \prod_{i<j}(1+R_{ij}^2).$$

Therefore if $\limsup_{n\to\infty} \|R\|_F < \infty$ then $\mathcal{P}_n$ is contiguous with respect to $\mathcal{Q}_n$.

Theorem 4 showed that the condition $\limsup \|R\|_F = \infty$ as $n \to \infty$ is necessary for the existence of a consistent test procedure for testing Eq. (3.4). We now present an example of a hypothesis testing problem for which $\limsup \|R\|_F = \infty$ is also sufficient. Consider a R-ER(P) model with $P_{ij} \equiv p$, i.e., the marginal distribution for the observed graphs is Erdős-Rényi. Suppose also that $p \geq c_0$ for some constant $c_0 > 0$ not depending on n. Next, suppose that n is even and let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be such that $\sigma_i \in \{-1, 1\}$ for all i and $\sum_{i=1}^n \sigma_i = 0$. Given σ,

the correlation matrix R has entries

$$R_{ij} = \begin{cases} r, & \text{if } \sigma_i = \sigma_j \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

We are interested in testing the hypothesis

$$\mathbb{H}_0 \colon r = 0 \quad \text{against} \quad \mathbb{H}_A \colon r \neq 0$$

and this is a special case of Eq. (3.4) where

$$\frac{(n^2 - 2n)}{4} r^2 \leq \|R\|_F^2 = \sum_{i \neq j} r^2 1_{\{\sigma_i = \sigma_j\}} \leq n^2 r^2.$$

Now consider the matrix $S = A \circ B$ with elements $S_{ij} = A_{ij} B_{ij}$. The S_{ij} are then *independent* Bernoulli random variables with

$$S_{ij} \sim \begin{cases} \text{Bernoulli}(p^2 + rp(1-p)), & \sigma_i = \sigma_j \\ \text{Bernoulli}(p^2), & \sigma_i \neq \sigma_j. \end{cases}$$

In other words S is the adjacency matrix of an Erdős-Rényi graph with edge probabilities p^2 under $\mathbb{H}_0$ and is the adjacency matrix of a 2-blocks stochastic blockmodel graph under $\mathbb{H}_A$. Let $\lambda_1(S)$ be the largest eigenvalue of S. Then by Füredi and Komlós (1981) and Athreya et al. (2022), we have

$$\lambda_1(S) - np^2 \longrightarrow N(1 - p^2, 2p^2(1-p^2)) \quad \text{under } \mathbb{H}_0, \tag{3.6}$$

$$\lambda_1(S) - np^2 - \frac{1}{2} nrp(1-p) \longrightarrow N(\eta, \gamma) \quad \text{under } \mathbb{H}_A. \tag{3.7}$$

Here η and γ are bounded constants. Therefore $\lambda_1(S)$ yields a consistent test procedure for testing $\mathbb{H}_0 \colon r = 0$ against $\mathbb{H}_A \colon r \neq 0$ whenever $\|R\|_F \asymp |nr| \to \infty$. More specifically, as both A and B are *marginally* Erdős-Rényi graphs, first estimate p via

$$\hat{p} = \frac{1}{n(n-1)} \sum_{i<j} (a_{ij} + b_{ij}).$$

Next, define $T(A, B) = |\lambda_1(S) - n\hat{p}^2|$. Then $T(A, B)$ is bounded in probability under $\mathbb{H}_0$ and $T(A, B)$ diverges under $\mathbb{H}_A$ if $\|R\|_F \to \infty$. We summarized the above discussion in the following result.

Proposition 1. Let (A, B) be a pair of R-correlated Erdős-Rényi graphs with marginal edges probability p where $p > 0$ does not depend on n and suppose that R is of the form in Eq. (3.5). Then there exists a consistent test procedure for testing $\mathbb{H}_0 \colon r = 0$ against $\mathbb{H}_A \colon r \neq 0$ if and only if $\|R\|_F \to \infty$ as $n \to \infty$.

Remark 2. We end this subsection with an example of a correlation matrix R for which the condition $\|R\|_F \to \infty$ is *not* sufficient to guarantee the existence of a consistent test procedure. Fix a constant $\epsilon \geq 0$. Let (A, B) be R-correlated Erdős-Rényi graphs with marginal edges probability $p = 0.5$ where R is a symmetric matrix whose (upper triangular) entries are iid random variables with $\mathbb{P}[R_{ij} = \epsilon] = \mathbb{P}[R_{ij} = -\epsilon] = 0.5$.

Given (A, B), detecting correlation between A and B is equivalent to testing $\mathbb{H}_0 \colon \epsilon = 0$ against $\mathbb{H}_A \colon \epsilon > 0$. However, as R is unknown and we only observed (A, B), the probabilities of a given pair (A, B) under either $\mathbb{H}_0$ or $\mathbb{H}_A$ are the same. More specifically, under $\mathbb{H}_0$ we have

$$\mathbb{P}_{\mathbb{H}_0}(A_{ij} = a, B_{ij} = b) = \frac{1}{4}, \quad \text{for all } (a, b) \in \{(0,0), (0,1), (1,0), (1,1)\}. \tag{3.8}$$

Meanwhile under $\mathbb{H}_A$, if we first conditioned on R_{ij} then

$$\mathbb{P}(A_{ij} = a, B_{ij} = b | R_{ij}) = \begin{cases} \frac{1+R_{ij}}{4}, & \text{if } (a, b) \in \{(0,0), (1,1)\} \\ \frac{1-R_{ij}}{4}, & \text{if } (a, b) \in \{(1,0), (0,1)\} \end{cases} \tag{3.9}$$

and hence, as R_{ij} is unobserved,

$$\begin{aligned} \mathbb{P}_{\mathbb{H}_A}(A_{ij} = a, B_{ij} = b) &= \frac{1}{2}\mathbb{P}(A_{ij} = a, B_{ij} = b \mid R_{ij} = \epsilon) + \frac{1}{2}\mathbb{P}(A_{ij} = a, B_{ij} = b \mid R_{ij} = -\epsilon) \\ &= \frac{1}{4}. \end{aligned}$$

There is thus no consistent test procedure for $\mathbb{H}_0 \colon \epsilon = 0$ against $\mathbb{H}_A \colon \epsilon \neq 0$ under this assumed correlation structure for R even though $\|R\|_F = n\epsilon$ for *any* realization of R (so that $\|R\|_F \to \infty$ as $n \to \infty$ for any fixed $\epsilon > 0$). The condition $\|R\|_F \to \infty$ is therefore not sufficient.

3.3.2 Computational Feasibility

The results in Section 3.3.1 provide a necessary condition for statistical detectability, i.e., the existence of a consistent test procedure for testing $\mathbb{H}_0 \colon \|R\|_F$ against $\mathbb{H}_A \colon \|R\|_F > 0$. However, there are numerous problems that are statistically feasible with parameter regimes for which there are no known *computationally efficient* procedures for solving them. Examples include community detection (Banks et al. 2016), sparse PCA (Lesieur et al. 2015) and estimation in

spiked tensor models (Perry et al. 2020). We now present an example of this phenomenon in the context of independence testing. In particular, we show the presence of a statistical vs. computational gap by transforming the independence testing problem to the following well-known planted clique problem in theoretical computer science.

Let A be an Erdős-Rényi graph on n vertices with common edges probability p. Next, given an integer $s_0 \geq 0$, select a subset of s_0 vertices of A and form a clique between these s_0 vertices. Suppose we are now given a graph A generated according to the above planted clique model with *unknown* s_0 and a positive integer $k \geq 1$. The PlantedClique problem seeks to determine if A contains a clique of size at least k. It is conjectured that there is no polynomial time algorithm for the PlantedClique problem that works for *all* values of $k \in \{1, 2, \ldots, n\}$, unless the P = NP hypothesis in computational complexity holds (Braverman et al. 2017). In particular, if $\log n \ll k \ll n^{1/2}$, then all known algorithms require quasi-polynomial time of $n^{O(\log n)}$ (Barak et al. 2019; Alon et al. 1998).

Let (A, B) be R-correlated Erdős-Rényi graphs with marginal edges probability $p = \frac{1}{2}$ and correlation matrix R constructed as follows. First, select a subset $\mathcal{C}$ of vertices with $|\mathcal{C}| = s_0$. Then define

$$R_{ij} = \begin{cases} -1, & \text{if } i \in \mathcal{C} \text{ and } j \in \mathcal{C}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

Note that $\|R\|_F = s_0$ and furthermore, $(A_{ij}, B_{ij}) \in \{(0,1), (1,0)\}$ whenever $i, j \in \mathcal{C}$ and $i \neq j$. Now consider testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$. Assume s_0 is unknown with $\log n \ll s_0 \ll n^{1/2}$ and suppose that there exists a polynomial time consistent test procedure for testing $\mathbb{H}_0$ and $\mathbb{H}_A$. Let $S = |A - B|$ where the absolute value is taken elementwise, i.e., $S_{ij} = |A_{ij} - B_{ij}|$. Note that S is the adjacency matrix of a random graph generated from the planted clique model with clique size s_0 and edges probability $p = \frac{1}{2}$. Then a consistent test procedure for testing $\mathbb{H}_0$ against $\mathbb{H}_A$ will also yield a polynomial time algorithm for deciding whether or not S has a clique of size at least s_0, thereby contradicting the claim of the Planted Clique conjecture; see the appendix for a more formal argument. In summary we have the following example of a statistical versus computational gap for independence testing.

Theorem 5. Let (A, B) be a pair of R-correlated Erdős-Rényi graphs on n vertices where R is of the form in Eq. (3.10) for some $\mathcal{C}$ with $\log n \ll |\mathcal{C}| \ll n^{1/2}$. Then, assuming the Planted Clique conjecture holds, there is no polynomial time consistent test procedure for testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ even though $\|R\|_F \to \infty$ as $n \to \infty$.

3.4 Independence testing in the graphon model

3.4.1 Same Marginal Distribution

We now describe independence testing when A and B are, *marginally*, generated from the class of latent positions models or graphons (Hoff et al. 2002; Bollobás et al. 2007; Lovász 2012). In particular, we will derive an asymptotically valid and consistent test procedure that also runs in time polynomial in n. We first define the notion of a pair of R-correlated latent position graphs where the correlation matrix R is also generated from a collection of latent positions. This is a natural extension of the latent positions model for a single graph to the setting of two graphs sharing a common vertex set with edges that are possibly pairwise correlated.

Definition 6. Let $\{X_1, \ldots, X_n\} \subset U \subset \mathbb{R}^d$, and $\{Y_1, \ldots, Y_n\} \subset V \subset \mathbb{R}^{d'}$ be two collections of *latent positions*. Now let h be a symmetric bivariate function from $\mathbb{R}^d \times \mathbb{R}^d$ to $[0, 1]$ and g be a symmetric bivariate function from $\mathbb{R}^{d'} \times \mathbb{R}^{d'}$ to $[-1, 1]$; assume also that h and g do not depend on n. Let $\rho_n \in [0, 1]$ and $\gamma_n \in [0, 1]$ and define the $n \times n$ matrices P and R by

$$P_{ij} = \rho_n \cdot h(X_i, X_j), \quad R_{ij} = \gamma_n \cdot g(Y_i, Y_j)$$

where we have implicitly assumed that γ_n and g are chosen so that R_{ij} satisfies the constraint in Eq. (3.1) for all i, j. Given P and R, we generate (A, B) as in Definition 4. We then say that (A, B) is a pair of correlated latent position graphs with correlation matrix R and edge probabilities matrix P.

Remark 3. In Definition 6, the factor ρ_n controls the *sparsity* of the observed graphs A and B while the factor γ_n controls the magnitudes of the pairwise correlations $\{R_{ij}\}$. Note that the average degrees for both A and B is $\Theta(n\rho_n)$; in the following presentation, we will allow for the possibility that $\rho_n \to 0$ and $\gamma_n \to 0$ as $n \to \infty$.

Let (A, B) be a pair of graphs generated according to Definition 6 and C be the matrix with entries

$$C_{ij} = \begin{cases} 1 & \text{if } A_{ij} + B_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

Denote by H the matrix whose entries are $H_{ij} = \mathbb{E}[C_{ij}]$, i.e.,

$$\begin{aligned} H_{ij} &= 2P_{ij} - P_{ij}^2 - R_{ij}P_{ij}(1 - P_{ij}) \\ &= \rho_n h(X_i, X_j) + (1 - \gamma_n g(X_i, X_j))\rho_n h(X_i, X_j)(1 - \rho_n h(X_i, X_j)). \end{aligned} \tag{3.12}$$

We now consider testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$. Our test statistic is based on estimating R using singular value thresholding (USVT). More specifically, we first compute an estimate $\hat{P}_1$ (resp. $\hat{P}_2$) of P by truncating the singular value decomposition of A (resp. B) to keep only the k_A (resp. k_B) largest singular values. Here $k_A := \{\max k \colon \sigma_k(A) \geq c_0\sqrt{n\rho_n}\}$, the $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq 0$ are the singular values of A, and $c_0 > 4$ is a universal constant; the value k_B is defined analogously. See Chatterjee (2015); Xu (2018) for more details. We also apply the same operations to C and obtain an estimate $\hat{H}$ of H. Let $\circ$ denote the Hadamard product for matrices. Then under $\mathbb{H}_0$ we have $\|H - 2P + P \circ P\|_F = 0$ and thus we consider a test statistic based on $\|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F$ where $\hat{P} = (\hat{P}_1 + \hat{P}_2)/2$. Leveraging recent results on the estimation error of SVT for graphon estimation (Xu 2018; Chatterjee 2015), we obtain the following consistency guarantee for our test procedure.

Theorem 6. Let (A, B) be a pair of graphs generated according to the model in Definition 6, where g and h are fixed functions and do not vary with n. Assume that both g and h are at least s times differentiable for some $s \geq 1$, where s is assumed known, and that $n\rho_n = \omega(\log n)$ as n increases. Let $\alpha = \frac{s+d+d'}{2s+d+d'}$ and define

$$T(A, B) = \frac{\|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F}{\Delta^\alpha \log^{1/2} n}.$$

where Δ is the average of the maximum degree of A and B. Let $\mathcal{R} = \{T \colon T > 1\}$. The test statistic $T(A, B)$ with rejection region $\mathcal{R}$ yields an asymptotically valid test procedure for testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ and furthermore, T is consistent whenever $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{\alpha'})$ for any $\alpha' > \alpha$ as $n \to \infty$.

We note that the assumption $n\rho_n = \omega(\log n)$ in Theorem 6 guarantees that the average degrees of the observed graphs grow slightly faster than $\log n$. This then allows the singular value thresholding step to yield reasonably accurate estimates $\hat{P}$ and $\hat{H}$. See for example the condition in Eq. 3 of Xu (2018). Furthermore, the consistency regime for T in Theorem 6 is stated in terms of $\|R \circ (P - P \circ P)\|_F$ as opposed to $\|R\|_F$. This is expected as the entries of $R \circ (P - P \circ P)$, which are $R_{ij}P_{ij}(1 - P_{ij})$, correspond to the difference between the edge probabilities matrix of (A, B) under $\mathbb{H}_0$ and $\mathbb{H}_A$. If we assume that the link function h satisfies $h(x, x') > 0$ for all $(x, x') \in U \times U$ then $P_{ij} = \Omega(\rho_n)$ for all i, j, and the condition $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{\alpha'})$ simplifies to $\rho_n\|R\|_F = \Omega((n\rho_n)^{\alpha'})$, or equivalently that $\|R\|_F = \Omega(n^{\alpha'}\rho_n^{\alpha'-1})$. As α decreases when s increases, we see that smoother h and g lead to a sharper consistency threshold for $\|R\|_F$. As a special case of Theorem 6, suppose that both g and h are *infinitely* differentiable. We then have $\alpha \leq 1/2 + \epsilon$ for any $\epsilon > 0$ and our test procedure is consistent whenever $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{1/2+\epsilon})$ for any $\epsilon > 0$. More

Algorithm 1 Bootstrap procedure for graphons

Require: Adjacency matrices A and B, both of size $n \times n$, significance level $\alpha \in (0,1)$, number of bootstrap samples m.

(A) Compute the matrix C whose entries are $C_{ij} = 1$ if $A_{ij} + B_{ij} > 0$ and $C_{ij} = 0$ otherwise.

(B) Compute $\hat{P}_1, \hat{P}_2$ and $\hat{H}$ by applying universal singular value thresholding (USVT) on A, B, and C, respectively.

(C) Let $\hat{P} = \frac{1}{2}(\hat{P}_1 + \hat{P}_2)$ and calculate the test statistic $T = \|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F$.

for $s = 1$ to m **do**

(i) Generate adjacency matrices $(A^{(s)}, B^{(s)})$ according to Definition 4 with $R = 0$ and marginal edge probabilities matrix $\hat{P}$.

(ii) Compute $\hat{P}_1^{(s)}$ and $\hat{P}_2^{(s)}$ as the universal singular value threshold of $A^{(s)}$ and $B^{(s)}$, respectively.

(iii) Calculate $T^{(s)} = \|\hat{H}^{(s)} - 2\hat{P}^{(s)} + \hat{P}^{(s)} \circ \hat{P}^{(s)}\|_F$, where $\hat{P}^{(s)} = \frac{1}{2}(\hat{P}_1^{(s)} + \hat{P}_2^{(s)})$ and $\hat{H}^{(s)}$ is the universal singular value thresholding of $A^{(s)} + B^{(s)}$.

end for

(D) Set c_α to be the $(1-\alpha) \times 100\%$ percentile of the $\{T^{(s)}\}_{s=1}^m$

Output If $T > c_\alpha$ then reject $\mathbb{H}_0$; otherwise fail to reject $\mathbb{H}_0$.

specifically we have the following corollary.

Corollary 1. Consider the setting in Theorem 6 and suppose that g and h are both infinitely differentiable. Now choose an arbitrary $\epsilon > 0$ not depending on n and define

$$T(A, B) = \frac{\|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F}{\Delta^{1/2+\epsilon/2}}.$$

Let the rejection region be given by $\mathcal{R} = \{T \colon T > 1\}$. Then the test statistic $T(A, B)$ with rejection region $\mathcal{R}$ yields an asymptotically valid test procedure for testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ and furthermore, T is consistent whenever $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{1/2+\epsilon})$ as $n \to \infty$.

Remark 4. If we suppose that (1) $h(x, x') > 0$ for all $x, x' \in U$ and (2) $g(y, y') > 0$ for all $y, y' \in V$ then $\|R \circ (P - P \circ P)\|_F = \Theta(n\gamma_n\rho_n)$ and the condition $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{1/2+\epsilon})$ in Corollary 1 is equivalent to the condition $\gamma_n = \Omega((n\rho_n)^{-1/2+\epsilon})$, i.e., the test procedure is consistent whenever the pairwise correlations decay to 0 slower than the reciprocal of the square root of the average degree.

If g and h are not infinitely differentiable then the test statistic in Theorem 6 depends on knowing (1) a lower bound for the smoothness s of the functions g and h and (2) upper bounds for the dimensions d and d' of the latent positions $\{X_i\}$ and $\{Y_i\}$. These values are most likely unknown in practice. Furthermore, even when s, d and d' are known, for finite

sample the rejection region in Theorem 6 is likely to be overly conservative. This is a common issue in many graph testing problems whose test statistics have no known non-degenerate limiting distribution; see for example the test statistics in Tang et al. (2017a), Ghoshdastidar et al. (2020), and Gretton et al. (2012). In light of these limitations, for this dissertation we will instead consider the *unnormalized* test statistic $\tilde{T}(A, B) = \|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F$ and use a bootstrap procedure to determine the rejection region for $\tilde{T}$. More specifically the bootstrap procedure generates additional pairs of graphs $\{A_b, B_b\}_{b=1}^B$ from the estimated edge probabilities matrix $\hat{P}$ with $R = 0$ and then computes the test statistics for these bootstrap pairs to obtain an empirical distribution for the test statistics under the null hypothesis. See Algorithm 1 for more details. Note that $\tilde{T}(A, B)$ differs from $T(A, B)$ only in the term $\Delta^\alpha \log^{1/2} n$. This term is a normalizing factor that guarantees $T(A, B) < 1$ when $\|R\|_F = 0$ and $T(A, B) \to \infty$ when $\|R\|_F = \Omega((n\rho_n)^{\alpha'})$. Hence, by using $\tilde{T}(A, B)$ and bootstrapping, we circumvent the need to know/estimate α (a possibly non-trivial if not impossible task).

3.4.2 Detection thresholds for stochastic blockmodels

We now consider the special case of independence testing when the graphs A and B are, *marginally*, stochastic blockmodel graphs with a common block structure. We begin by formulating an extension of the stochastic blockmodel (Holland et al. 1983) for a single graph to the case of a pair of graphs whose edge correlations also exhibit a block structure. This extension had appeared previously in the literature in the context of graph matching (see e.g., Onaran et al. (2016); Racz and Sridhar (2021); Lyzinski et al. (2014a)).

Definition 7. Let A and B be graphs on n vertices. We say that (A, B) is a pair of K-blocks correlated stochastic blockmodel graphs with common community assignments τ, block probabilities matrices Θ_P and Θ_Q, sparsity factor ρ_n and block correlations matrix Θ_R if (A, B) are generated from the R-ER(P, Q) model where

1. Θ_P and Θ_Q are $K \times K$ *symmetric* matrices with entries in $[0, 1]$.
2. Marginally, the edge probabilities matrix for A and B are $P = \rho_n Z \Theta_P Z^\top$ and $Q = \rho_n Z \Theta_Q Z^\top$ where Z is a $n \times K$ matrix such that, for any $k \in \{1, 2, \ldots, K\}$, we have $Z_{ik} = 1$ if $\tau_i = k$ and $Z_{ik} = 0$ otherwise.
3. The correlation matrix R is $R = Z \Theta_R Z^\top$.

Let (A, B) be generated from the model in Definition 7. We now describe a test statistic T for testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ with the following properties (1) T has a

central χ^2 limiting distribution under the null hypothesis and (2) T is consistent under the alternative hypothesis provided that $\|R\|_F \to \infty$.

First, compute the matrix C as defined in Eq. (3.11) and cluster the vertices of C into K communities using a community detection algorithm that guarantees exact recovery (see e.g., Abbe (2017); Gao et al. (2017); Lyzinski et al. (2014b)) where the value of K can be chosen using model selection procedures such as those described in Li et al. (2020); Wang and Bickel (2017); Lei (2016). Let $\hat{\tau}$ be the resulting estimated community assignment. Next, compute, for $1 \leq k \leq \ell \leq K$, the Pearson sample correlation $\hat{\rho}_{k\ell}$ between the edges $\{A_{ij}, B_{ij}\}$ in the (k,ℓ)th block, i.e.,

$$\hat{\rho}_{k\ell} = \operatorname{cor}\Big(\{A_{ij} \colon i < j, \hat{\tau}_i = k, \hat{\tau}_j = \ell\}, \{B_{ij} \colon i < j, \hat{\tau}_i = k, \hat{\tau}_j = \ell\}\Big).$$

Let $\hat{n}_k = |\{i \colon \hat{\tau}_i = k\}|$ for all k and define the test statistic

$$T(A,B) = \sum_{k \leq \ell} \hat{n}_{k\ell} \hat{\rho}_{k\ell}^2$$

where $\hat{n}_{kk} = \binom{\hat{n}_k}{2}$ if $k = \ell$ and $\hat{n}_{k\ell} = \hat{n}_k \hat{n}_\ell$ if $k \neq \ell$. We then have the following result.

Theorem 7. Let (A,B) be a pair of graphs on n vertices generated according to Definition 7 where Θ_P and Θ_Q are fixed $K \times K$ matrices not depending on n. Suppose that, as n increases, we have $n_k = |\{i \colon \tau_i = k\}| = \Theta(n)$ for all $k \in \{1, 2, \ldots, K\}$ and the sparsity ρ_n satisfies $n\rho_n = \omega(\log n)$. Then under $\mathbb{H}_0 \colon \|R\|_F = 0$ we have $T(A,B) \rightsquigarrow \chi^2_{K(K+1)/2}$ as $n \to \infty$. Furthermore let $\mu > 0$ be a finite constant such that $\|R\|_F > 0$ satisfies

$$\sum_{k \leq \ell} n_{k\ell} \Theta_R^2(k,\ell) \to \mu \tag{3.13}$$

as $n \to \infty$ (here $\Theta_R(k,\ell)$ denote the k,ℓth entry of Θ_R). We then have

$$T(A,B) \rightsquigarrow \chi^2_{K(K+1)/2}(\mu)$$

as $n \to \infty$. Here $\chi^2_{K(K+1)/2}(\mu)$ is a non-central χ^2 with $K(K+1)/2$ degrees of freedom and non-centrality parameter μ.

Theorem 6 (and its associated Corollary 1) to Theorem 7 we see that by assuming a more specialized structure on R we obtain a much sharper detection threshold. Indeed, the local alternative in Theorem 7 implies that our test statistic achieves power converging to 1 whenever Θ_R satisfies the condition $\sum_{k\ell} n_{k\ell} \Theta_R^2(k,\ell) \to \infty$. This is equivalent to the condition that $\|R\|_F \to \infty$ and is thus, by Theorem 4, both necessary and sufficient. Theorem 7 also

assumes that the block structure for R is the same as that for P and Q, and this is done purely for ease of exposition as it allows us to more easily aggregate the edges of A and B into blocks and estimate the pairwise correlations between the edges in each block. Extending Theorem 7 to the case where the SBM structure for R differs from that of P and Q is straightforward but tedious, and we leave it to the interested reader. Finally, Xiong et al. (2019) considered a special case of Definition 7 with $R_{ij} \equiv c$ for all $\{i, j\}$, but they did not derive a non-degenerate limiting distribution for their test statistic.

3.4.3 Different Marginal Distributions

We now discuss how the previous model and results can be extended to the case where the graphs A and B have different *marginal* distributions. We first define a model that generalizes the R-ER(P) model in Definition 4 (see also Remark 1).

Definition 8. Let $n \in \mathbb{N}$. Let $P \in [0,1]^{n \times n}$, $Q \in [0,1]^{n \times n}$, and $R \in [-1,1]^{n \times n}$ be symmetric matrices where R satisfies the constraint

$$-\min\left\{\frac{P_{ij}Q_{ij}}{(1-P_{ij})(1-Q_{ij})}, \frac{(1-P_{ij})(1-Q_{ij})}{P_{ij}Q_{ij}}\right\}^{1/2} \leq R_{ij} \leq \min\left\{\frac{P_{ij}(1-Q_{ij})}{Q_{ij}(1-P_{ij})}, \frac{Q_{ij}(1-P_{ij})}{P_{ij}(1-Q_{ij})}\right\}^{1/2}, \tag{3.14}$$

for all i, j. We say that (A, B) are R-correlated heterogeneous Erdős-Rényi graphs on n vertices with *marginal* edge probabilities (P, Q) and correlations R, denoted by $(A, B) \sim R\text{-ER}(P, Q)$, if

1. A is the adjacency matrix for an inhomogeneous Erdős-Rényi graph on n vertices with probability matrix P.
2. B is the adjacency matrix for another inhomogeneous Erdős-Rényi graph on n vertices with probability matrix Q.
3. The pairs of entries $\{(A_{ij}, B_{ij})\}_{1 \leq i < j \leq n}$ are independent bivariate random vectors.
4. For $i < j$, A_{ij} and B_{ij} are correlated with Pearson correlation R_{ij}.

Following the proof of Theorem 4, we can show that the condition $\limsup \|R\|_F < \infty$ as$n \to \infty$ is also necessary for the existence of a consistent test procedure for detecting the correlation between R-ER(P, Q) graphs. A simple generative model for R-ER(P, Q) graphs is the following variant of the graphon model described in Definition 6 where, instead of only having a single h, we have two link functions h_1 and h_2 from $U \times U \mapsto [0,1]$ and define P and Q via $P_{ij} = \rho_n h_1(X_i, X_j)$ and $Q_{ij} = \rho_n h_2(X_i, X_j)$ for all $i \leq j$. The function g is also chosen so that the correlation matrix R with entries $R_{ij} = \gamma_n g(Y_i, Y_j)$ satisfies the constraint in

Eq. (3.14). Note that for ease of exposition we had assumed that the sparsity factor for both P and Q are the same. The case where P and Q have different sparsity factors involves more tedious book-keeping but is, otherwise, conceptually identical and leads to similar results as those described below.

Given (A, B) sampled from the above variant of the R-ER(P, Q) model we once again let C be the matrix with entries $C_{ij} = 1$ if $A_{ij} + B_{ij} > 0$ and $C_{ij} = 0$ otherwise. Denote $H = \mathbb{E}[C]$. We then have

$$H_{ij} = P_{ij} + Q_{ij} - P_{ij}Q_{ij} - R_{ij}\sqrt{P_{ij}(1 - P_{ij})Q_{ij}(1 - Q_{ij})} \tag{3.15}$$

and thus we can construct a test statistic for $\mathbb{H}_0 \colon \|R\|_F$ against $\mathbb{H}_A \colon \|R\|_F > 0$ by first applying USVT to A and B to obtain estimates $\hat{P}$ of P and $\hat{Q}$ of Q, then applying USVT to C to obtain an estimate $\hat{H}$ of H, and finally compute $\tilde{T}(A, B) = \|\hat{H} - \hat{P} - \hat{Q} + \hat{P} \circ \hat{Q}\|_F$. The following result is derived using an identical argument to that for Theorem 6 and shows that rejecting $\mathbb{H}_0$ for large values of $\tilde{T}$ yields a consistent test procedure.

Theorem 8. Let (A, B) be a pair of graphs generated from the above R-ER(P, Q) model where g, h_1 and h_2 are fixed functions and do not vary with n. Assume that all of g, h_1, h_2 are at least s times continuously differentiable for some $s \geq 1$, where s is assumed known, and that $n\rho_n = \omega(\log n)$ as n increases. Let $\alpha = \frac{s+d+d'}{2s+d+d'}$ and define

$$T(A, B) = \frac{\|\hat{H} - \hat{P} - \hat{Q} + \hat{P} \circ \hat{Q}\|_F}{\Delta^{\alpha} \log^{1/2} n}.$$

where Δ is the average of the maximum degree of A and B. Let $\mathcal{R} = \{T \colon T > 1\}$. The test statistic $T(A, B)$ with rejection region $\mathcal{R}$ yields an asymptotically valid test procedure for testing $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ and furthermore, T is consistent whenever $\|R \circ \Xi\|_F = \Omega((n\rho_n)^{\alpha'})$ for any $\alpha' > \alpha$ as $n \to \infty$. Here Ξ is the $n \times n$ matrix with entries $\Xi_{ij} = \big(P_{ij}(1 - P_{ij})Q_{ij}(1 - Q_{ij})\big)^{1/2}$.

Similar to our discussion after Corollary 1, if the link functions g, h_1 and h_2 are not infinitely differentiable then the test statistic in Theorem 8 depends on knowing a lower bound for the smoothness s and upper bounds for the dimensions d and d' of the latent positions $\{X_i\}$ and $\{Y_i\}$. These values are once again unknown or, even if known, the rejection region in Theorem 8 is still overly conservative for finite sample inference. We will thus also use bootstrap resampling to calibrate the rejection region for the above test statistic $T(A, B)$; see Algorithm 2 in Section 3.6 for more details.

Table 3.1: Empirical estimates (based on $m = 100$ Monte Carlo replicates) for the Type I and type II error when P_{ij} is from the cosine similarity. Given values correspond to the Type-I error when $r = 0$ and to the power (i.e., one minus the Type II error) when $r > 0$.

$n \setminus r$	$r = 0$	$r = 0.1$	$r = 0.3$	$r = 0.5$
$n = 100$	0.06	0.17	0.85	1
$n = 200$	0.07	0.98	1	1
$n = 500$	0.04	1	1	1
$n = 1000$	0.02	1	1	1
$n = 2000$	0.02	1	1	1

3.5 Simulation Results

We now conduct simulation experiments to evaluate the performance of our test procedures for testing the hypothesis in Section 3.4. For our first experiment, we generate $\{X_i\}_{i=1}^n$ as iid sample from a bivariate normal with mean 0 and identity covariance matrix. We then consider two different choices of link functions for P. The first is the cosine similarity

$$P_{ij} = \frac{|X_i^\top X_j|}{2\|X_i\|\|X_j\|} \tag{3.16}$$

and the second is the Gaussian kernel $P_{ij} = \exp\big(-\|X_i - X_j\|^2\big)/2$. Note that the rank of P is 2 and n for the cosine and Gaussian similarity, respectively. Given P we set $R_{ij} \equiv r$ for some value r to be specified later. We then generate a pair of graphs $(A, B) \sim R\text{-ER}(P)$ on n vertices and apply the test statistic T in Corollary 1 to test $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$ where the rejection region is calibrated via bootstrapping with significance level 0.05 (see Algorithm 1). We repeat the above steps for $m = 100$ Monte Carlo replicates to obtain an empirical estimate of the type I error (when $r = 0$) and type II error (when $r > 0$) for T. The results are presented in Tables 3.1 and 3.2 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $r \in \{0, 0.1, 0.3, 0.5\}$. We observe that the type I error of the test statistic is well-controlled and that the test statistic also exhibits power even for small values of r and moderate values of n. Note that we fixed $k = 2$ when P is the cosine similarity; here k is the number of singular values used to construct T. In contrast, we chose k via USVT (see Section 3.4) when P is the Gaussian similarity.

For our second experiment we consider the case where the graphs A and B have different marginal distributions (see Section 3.4.3). In particular we generate $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ as iid samples from a bivariate normal with mean 0 and identity covariance matrix and then set

Table 3.2: Empirical estimates (based on $m = 100$ Monte Carlo replicates) for the type I and type II error when P_{ij} is from the Gaussian kernel. The given values correspond to the Type-I error when $r = 0$ and to the power (i.e., one minus the type II error) when $r > 0$.

$n \setminus r$	$r = 0$	$r = 0.1$	$r = 0.3$	$r = 0.5$
$n = 100$	0	0	0.06	0.74
$n = 200$	0	0	0.26	1
$n = 500$	0	0	1	1
$n = 1000$	0	0	1	1
$n = 2000$	0	0.02	1	1

Table 3.3: Empirical estimates (based on $m = 100$ Monte Carlo replicates) for the type I and type II error when P_{ij} and Q_{ij} are from the cosine similarity with $P \neq Q$. The given values correspond to the type I error when $r = 0$ and to the power (i.e., one minus the type II error) when $r > 0$.

$n \setminus r$	$r = 0$	$r = 0.1$	$r = 0.3$	$r = 0.5$
$n = 100$	0	0.35	1	1
$n = 200$	0	0.83	1	1
$n = 500$	0.11	1	1	1
$n = 1000$	0.01	1	1	1
$n = 2000$	0.02	1	1	1

P and Q to have entries

$$P_{ij} = \frac{|X_i^\top X_j|}{2\|X_i\|\|X_j\|}, \quad \text{and} \quad Q_{ij} = \frac{|Y_i^\top Y_j|}{4\|Y_i\|\|Y_j\|}.$$

Given P and Q we once again set $R_{ij} \equiv r$ and generate pair of graphs (A, B) on n vertices from the R-ER(P, Q) model. We apply the test statistic in Theorem 8 to test $\mathbb{H}_0 \colon \|R\|_F = 0$ against $\mathbb{H}_A \colon \|R\|_F > 0$; the rejection region is also calibrated via bootstrapping with significance level 0.05 using Algorithm 2. Empirical estimates of the type I and type II errors (based on $m = 100$ Monte Carlo replicates) are presented in Table 3.3 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $r \in \{0, 0.1, 0.3, 0.5\}$. Once again, the type-I error is well-controlled and the test also exhibits power even for small values of r and moderate values of n.

For our last experiment, we consider correlation testing when A and B are *marginally* stochastic blockmodel graphs (c.f. Section 3.4.2). More specifically, we assume that A is a K-block SBM where each vertex of A is assigned to some a block $k \in \{1, \ldots, K\}$ with probability $1/K$ and the marginal edge probabilities between vertices in block k and vertices in block ℓ are given by $0.45 - |k - \ell|/(2K)$ for any $k, \ell \in \{1, 2, \ldots, K\}$. Similarly, B is a

K-block SBM with the same membership assignment as A and marginal edge probabilities $0.4-|k-\ell|/(2K+2)$. We set R to have the same block structure as both A and B and entries of the form $r(1-|k-\ell|/K)$ for values of r that are specified later. We then sample a pair (A, B) from the model in Definition 7 with parameters given above and test the hypothesis that $\|R\|_F = 0$ against $\|R\|_F > 0$ using the test statistic in Theorem 7 where the rejection region is based on the 95% percentile of the $\chi^2_{K(K+1)/2}$ distribution. We repeat these steps for $m = 1000$ Monte Carlo replicates to obtain empirical estimates of the type-I and type-II error of our test statistic. The results are presented in Table 3.4 for various combinations of $n \in \{200, 500, \ldots, 3000\}$, $K \in \{2, 5, 7\}$ and $r \in \{0, 0.001, 0.005, 0.01\}$. For comparisons we also include the limiting theoretical power given by $F_\mu^{-1}(c_*)$ where c_* is the 95% percentile of the (central) $\chi^2_{K(K+1)/2}$ distribution and F_μ^{-1} is the quantile function for a *non-central* $\chi^2_{K(K+1)/2}$ with non-centrality parameter $\mu = r^2(\frac{K^2+1}{4K^2}n^2 - \frac{n}{2})$. We observe that the empirical type-I errors (when $r = 0$) and empirical power (when $r > 0$) are close to their limiting theoretical counterparts provided that n is sufficiently large compared to K; indeed, as K increases we generally need larger values of n to achieve accurate recovery of the latent community assignments.

3.6 Real Data Experiments

3.6.1 Analysis of C. elegans Data

We now apply the test statistics in Section 3.4 to the connectomes of the *C. elegans* roundworm. More specifically, we used the wiring diagram formed by the somatic nervous system, which consists of 279 neurons; these neurons are classified into one of three categories namely motor neurons, sensory neurons, and inter-neuron. There are two types of connections between the neurons, i.e., either via chemical synapses or electrical gap junctions. This result in two related but distinct networks, namely a chemical synapse network A_c with 6394 edges and a gap junction network A_g with 1777 edges. See Varshney et al. (2011) for a more detailed description of the construction of these connectomes.

We first consider testing the null hypothesis that A_c and A_g are independent against the alternative hypothesis that they are correlated. As the two graphs have a quite large difference in the edge densities, we suppose that A_c and A_g are generated from the R-ER(P, Q) model (see Section 3.4.3) and use the test statistic T in Theorem 8 with $k = 3$ as the rank for the estimates $\hat{P}$ and $\hat{Q}$; the choice $k = 3$ is motivated by the fact that there are three categories of neurons. This result in an observed value of $T = 8.313$. We calibrate our test statistic using the bootstrapping procedure in Algorithm 2 with $m = 10000$ Monte Carlo replicates

Table 3.4: Empirical estimates (based on $m = 100$ Monte Carlo replicates) for the type I and type II error compared to the theoretical (limiting) value. Here A and B are R-correlated SBM graphs. The first (resp. second) entry in each cell correspond to the empirical estimate (resp. theoretical value) of the type I error when $r = 0$ and to the power (i.e., one minus the type II error) when $r > 0$. The theoretical values are based on the non-central χ^2 distribution with non-centrality parameter $\mu = r^2(\frac{K^2+1}{4K^2}n^2 - \frac{n}{2})$.

$K = 2$	$r = 0$	$r = 0.001$	$r = 0.005$	$r = 0.01$
$n = 200$	0.044/0.050	0.045/0.051	0.056/0.069	0.147/0.133
$n = 500$	0.047/0.050	0.055/0.055	0.177/0.188	0.621/0.641
$n = 1000$	0.056/0.050	0.070/0.069	0.638/0.642	1/0.999
$n = 2000$	0.048/0.050	0.139/0.134	0.998/0.999	1/1
$n = 3000$	0.040/0.050	0.236/0.259	1/1	1/1
$K = 5$	$r = 0$	$r = 0.001$	$r = 0.005$	$r = 0.01$
$n = 200$	0.864/0.050	0.755/0.050	0.795/0.056	0.888/0.076
$n = 500$	0.055/0.050	0.053/0.051	0.103/0.093	0.311/0.289
$n = 1000$	0.051/0.050	0.067/0.056	0.280/0.290	0.940/0.931
$n = 2000$	0.054/0.050	0.070/0.076	0.936/0.932	1/1
$n = 3000$	0.041/0.050	0.111/0.116	1/1	1/1
$K = 7$	$r = 0$	$r = 0.001$	$r = 0.005$	$r = 0.01$
$n = 200$	0.954/0.050	0.912/0.050	0.955/0.054	0.973/0.067
$n = 500$	0.617/0.050	0.560/0.051	0.656/0.079	0.826/0.208
$n = 1000$	0.058/0.050	0.062/0.054	0.218/0.209	0.859/0.833
$n = 2000$	0.056/0.050	0.074/0.068	0.823/0.834	1/1
$n = 3000$	0.049/0.050	0.104/0.094	1/0.999	1/1

Table 3.5: Sample means of the estimated correlations $\{\hat{R}_{ij}\}$ for different combinations of neuron types for i and j.

	motor	inter	sensory
motor	0.144	0.111	0.153
inter	0.111	0.088	0.137
sensory	0.153	0.137	0.193

and obtain an approximate p-value of 5×10^{-5}. There is thus strong evidence to reject the null hypothesis in favor of the alternative hypothesis that the two connectomes are correlated. This conclusion, while biologically relevant, is also certainly expected.

We now quantify the degree of correlation between the edges of the two graphs. Recalling Eq. (3.15) we first compute an estimate of the correlation R_{ij} between the edges of A_c and A_g via

$$\hat{R}_{ij} := \begin{cases} 0, & \text{if } \hat{P}_{ij} \text{ or } \hat{Q}_{ij} \in \{0, 1\} \\ \max\left\{ \min\left\{ \frac{\hat{P}_{ij} + \hat{Q}_{ij} - \hat{P}_{ij}\hat{Q}_{ij} - \hat{H}_{ij}}{\left(\hat{P}_{ij}(1-\hat{P}_{ij})\hat{Q}_{ij}(1-\hat{Q}_{ij})\right)^{1/2}}, 1 \right\}, -1 \right\}, & \text{otherwise.} \end{cases} \tag{3.17}$$

where $\hat{P}, \hat{Q}$ and $\hat{H}$ are obtained by applying USVT to the adjacency matrices A, B, and C respectively; recall that $C_{ij} = 1$ if $A_{ij} + B_{ij} > 0$ and $C_{ij} = 0$ otherwise. We then compute the average correlations for edges connecting vertices from the same category as well as edges connecting vertices from different categories, e.g., we calculate the sample mean of the $\{\hat{R}_{ij}\}$ when i and j are both motor neurons as well as the sample mean of the $\{\hat{R}_{ij}\}$ when i is a motor neuron and j is a sensory neuron. The results are presented in Table 3.5 for all possible pairs of neuron categories; these correlations are all positive and quite large. As a sanity check, we also compute the sample Pearson correlation based on the binary entries of A_c and A_g directly, i.e., we compute $\mathrm{Cor}(\{A_c(i,j), A_g(i,j)\})$ where i ranges over all neurons of type k and j ranges over all neurons of type ℓ (with k possibly being the same as ℓ). The results are presented in Table 3.6. We see that these sample Pearson correlations exhibit the same general pattern as that for the $\{\hat{R}_{ij}\}$ in Table 3.5. Indeed, the difference between the entries in Table 3.5 and Table 3.6 are all less than 0.1 and can be as small as 0.01 or 0.03.

Finally, we discuss the use of the $\{\hat{R}_{ij}\}$ to help improve link predictions for the edges of A_g. More specifically, we evaluate the accuracy for link prediction using only the estimated edge probabilities matrix $\hat{P}$ against the accuracy when using $\hat{P}$ in conjunction with $\hat{R}$. For both approaches, we first sub-sample a $A_g^{(\mathrm{sub})}$ from A_g by setting 10% of the entries of A_g to 0. We next apply USVT to $A_g^{(\mathrm{sub})}$ to obtain an estimate $\hat{P}^{(\mathrm{sub})}$ of P. Now let $\mathcal{E}$ be the

Table 3.6: Pearson correlations between the edges in A_c and A_g for different combinations of neuron types.

	motor	inter	sensory
motor	0.153	0.203	0.129
inter	0.203	0.145	0.147
sensory	0.129	0.147	0.171

set of entries in A_g that are set to 0 in A_g^{sub}. We then threshold the entries $\hat{P}_{ij}^{(\text{sub})}$ for all $(i,j) \in \mathcal{E}$, i.e., for $(i,j) \in \mathcal{E}$ we predict the presence of a link if $\hat{P}_{ij}^{(\text{sub})} > t$ and an absence of a link otherwise. By varying $t \in [0,1]$ we obtain a ROC curve and an associated AUC for link prediction using only the estimated $\hat{P}^{(\text{sub})}$. A similar approach had also been used in Zhang et al. (2017); Gao et al. (2015); Rubin-Delanchy et al. (2022) when the graphs are assumed to be generated from a latent space or graphon model. Link prediction using both A_c and A_g also follows a similar procedure, but this time we use both $A_g^{(\text{sub})}$ and $A_c^{(\text{sub})}$ to estimate the marginal edge probabilities $\hat{P}^{(\text{sub})}$ for A_g and $\hat{Q}^{(\text{sub})}$ for A_c as well as the estimated correlations $\hat{R}^{(\text{sub})}$; here $A_c^{(\text{sub})}$ is obtained by setting the entries of A_c indexed by $\mathcal{E}$ to 0 and $\hat{R}^{(\text{sub})}$ is calculated using a similar expression as that in Eq. (3.17) but with $\hat{P}$ and $\hat{Q}$ replaced by $\hat{P}^{(\text{sub})}$ and $\hat{Q}^{(\text{sub})}$, respectively. Given the $\hat{P}^{(\text{sub})}$ and $\hat{R}^{(\text{sub})}$ we then threshold the entries of $\hat{P}_{ij}^{(\text{sub})} + \hat{R}_{ij}^{(\text{sub})}(A_c(i,j) - \hat{P}_{ij}^{(\text{sub})})$ for all $(i,j) \in \mathcal{E}$; note that these choices of quantities is motivated from the fact that if $(A,B) \sim R\text{-}ER(P,Q)$ then $\mathbb{P}[A_{ij} = 1 \mid B_{ij}] = P_{ij} + R_{ij}(B_{ij} - P_{ij})$. By varying the threshold $t \in [0,1]$ we also obtain a ROC curve and associated AUC for link prediction using both $\hat{P}^{(\text{sub})}$ and $\hat{R}^{(\text{sub})}$.

We perform the above AUC calculations 100 times, each time choosing a random subset of entries $\mathcal{E}$ to set to 0. ROC curves for a random realization of $\mathcal{E}$ are shown in Figure 3.1. The average AUC when using only $\hat{P}^{(\text{sub})}$ is 0.575 with a standard error of 0.002; in contrast, the average AUC when using both $\hat{P}^{(\text{sub})}$ and $\hat{R}^{(\text{sub})}$ is 0.705 with a standard error of 0.003. The use of $\hat{R}^{(\text{sub})}$ thus leads to a significant increase in accuracy.

3.6.2 Wikipedia Data

We now analyze two networks formed by a collection of Wikipedia articles. The first network, denoted as A_e, consists of 1382 vertices and 37714 edges. Each vertex in A_e represents an article in the English Wikipedia on topics related to Algebraic Geometry, and two given vertices are connected if there is a hyperlink between them in the English Wikipedia. The second network, denoted as A_f, consists of 1382 vertices and 29946 edges corresponding to the same Wikipedia articles as that in A_e but the hyperlinks are now for the French

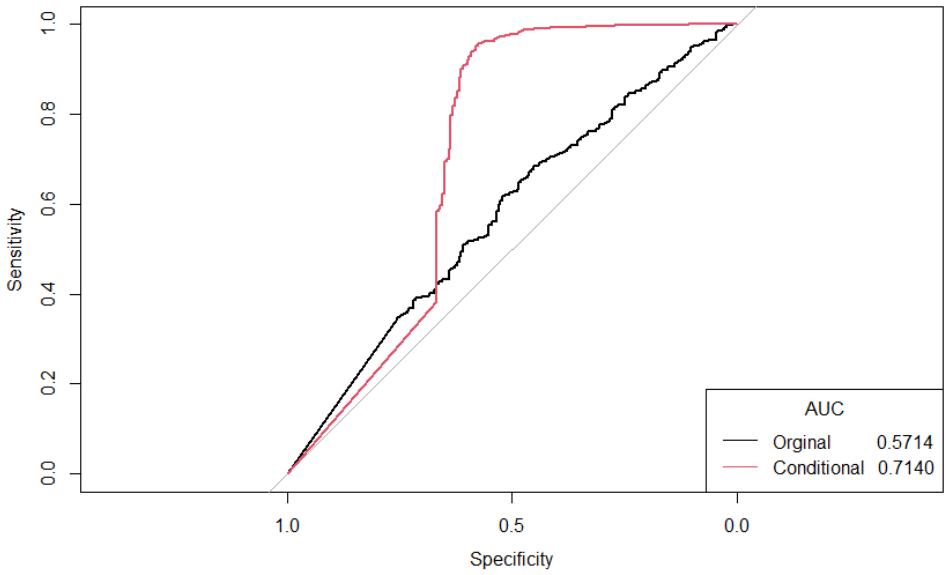


Figure 3.1: ROC curves for link prediction on a randomly selected set of entries $\mathcal{E}$ of the *C. elegans* gap junction network A_g. The black curve is for using A_g only while the red curve is for using both A_g and A_c.

Wikipedia. See Ma et al. (2012) for a more detailed description of these networks. We now follow the same analysis as that described in Section 2 for the *C. elegans data*. In particular, we first test the null hypothesis that A_e and A_f are independent. As A_e and A_f are both quite sparse (their edge densities are 0.02 and 0.016, respectively), we apply the test statistic in Theorem 6 to their complements $\bar{A}_e$ and $\bar{A}_f$, i.e., $\bar{A}_e = 11^\top - A_e$ and $\bar{A}_f = 11^\top - A_f$ where $11^\top$ is the 1382×1382 matrix of all ones. Note that, by Eq. (3.3), if $(A, B) \sim R\text{-}ER(P)$ then $(\bar{A}, \bar{B}) \sim R - ER(11^\top - P)$ and hence, assuming the model in Section 3.4 is appropriate, inference based on $T(A, B)$ and $T(\bar{A}, \bar{B})$ are theoretically equivalent. This yields an observed test statistic of $T(\bar{A}_e, \bar{A}_f) = 21.514$ and, using the bootstrapping procedure in Algorithm 2 with $m = 10000$, an approximate p-value of 5×10^{-5}. We thus reject the null hypothesis and are in favor of the alternative hypothesis that the English and French Wikipedia networks are correlated.

We next quantify the degree of correlations between the edges of A_e and A_f. The articles in A_e and A_f can be grouped into six classes, namely (1) people, (2) places, (3) dates, (4) math things (articles about math topics that are neither people, places, nor dates) (5) things (article about non-math topics that are neither people, places nor dates) and (6) categories (a special type of Wikipedia article). We then calculate the sample means of the estimated correlations $\hat{R}$ for edges within the same categories and between different categories (see the

Algorithm 2 Bootstrap procedure for graphons with possibly $P \neq Q$.

Require: Adjacency matrices A and B, both of size $n \times n$, significance level $\alpha \in (0, 1)$, number of bootstrap samples m.

(A) Compute the matrix C whose entries are $C_{ij} = 1$ if $A_{ij} + B_{ij} > 0$ and $C_{ij} = 0$ otherwise.

(B) Compute $\hat{P}, \hat{Q}$ and $\hat{H}$ by applying universal singular value thresholding (USVT) on A, B, and C, respectively.

(C) Calculate the test statistic $T = \|\hat{H} - \hat{P} - \hat{Q} + \hat{P} \circ \hat{Q}\|_F$.

for $s = 1$ to m **do**

(i) Generate adjacency matrices $(A^{(s)}, B^{(s)})$ according to Definition 5 with $R = 0$ and marginal edge probabilities matrices $\hat{P}$ and $\hat{Q}$.

(ii) Compute $\hat{P}$ and $\hat{P}$ as the universal singular value threshold of $A^{(s)}$ and $B^{(s)}$, respectively.

(iii) Calculate $T^{(s)} = \|\hat{H}^{(s)} - \hat{P}^{(s)} - \hat{Q}^{(s)} + \hat{P}^{(s)} \circ \hat{Q}^{(s)}\|_F$, where $\hat{H}^{(s)}$ is the universal singular value thresholding of $A^{(s)} + B^{(s)} - A^{(s)} \circ B^{(s)}$.

end for

(D) Find the smallest number t such that $T > T_t$, where T_t is the t-th largest element in $\{T^{(s)}\}_{s=1}^m$,

(D) p-value$= (t - 0.5)/m$

Output p-value.

description on Table 3.5 and Table 3.6 in Section 3.6.1 for more details). The results are presented in Table 3.7 and Table 3.8; once again we see that the entries in the two tables are highly similar and they both indicate that the correlations between the edges of A_e and A_f are positive and quite large.

Table 3.7: Sample means of the estimated correlations $\{\hat{R}_{ij}\}$ for different combinations of Wikipedia article types for i and j.

	people	places	dates	things	math things	categories
people	.501	.419	.336	.381	.375	.417
places	.419	.318	.269	.296	.272	.307
dates	.336	.269	.278	.227	.148	.159
things	.381	.296	.227	.267	.238	.282
math things	.375	.272	.148	.238	.192	.231
categories	.417	.307	.159	.282	.231	.273

Finally, we consider link prediction for the English Wikipedia network A_e. We follow the procedure described in Section 2 wherein we set 10% of the entries of A_e to 0 and then compare the AUC for link prediction from $\hat{P}^{(\text{sub})}$ alone against that of $\hat{P}^{(\text{sub})}$ and $\hat{R}^{\text{sub}}$. The

Table 3.8: Pearson correlations between the edges of A_e and A_f for different combinations of Wikipedia article types. The value `NA` for the pair `dates` and `categories` is because there are no edges between any vertices in `dates` and any vertices in `categories` for both graphs.

	people	places	dates	things	math things	categories
people	.570	.501	.460	.399	.499	.751
places	.501	.489	.376	.412	.295	.607
dates	.460	.376	.589	.292	.100	NA
things	.399	.412	.292	.348	.233	.452
math things	.499	.295	.100	.233	.283	.507
categories	.751	.607	NA	.452	.507	.462

sample mean of the AUCs based on 100 randomly selected $\mathcal{E}$ are 0.863 (standard error = 0.0006) for $\hat{P}^{(\text{sub})}$ only and improve to 0.956 (standard error = 0.0005) when we also include $\hat{R}^{(\text{sub})}$. ROC curves for a random realization of $\mathcal{E}$ are shown in Figure 3.2.

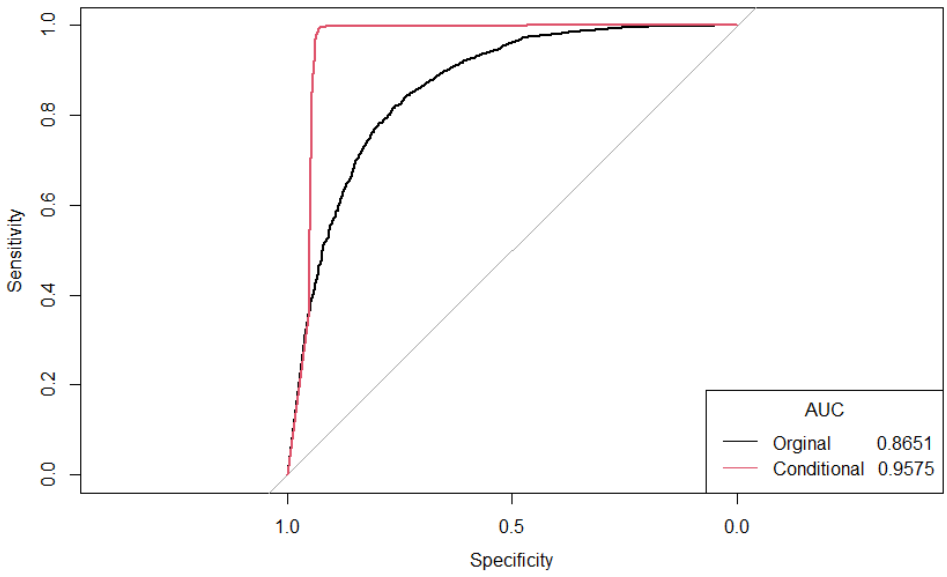


Figure 3.2: ROC curves for link prediction on a randomly selected set of entries $\mathcal{E}$ of the English Wikpedia network A_e. The black curve is for using A_e only while the red curve is for using both A_e and A_f.

3.7 Discussion

In this dissertation, we formulated independence testing between graphs as, given a pair of inhomogeneous Erdős-Rényi graphs with edge-correlation R, deciding between $\mathbb{H}_0 \colon \|R\|_F = 0$ and $\mathbb{H}_A \colon \|R\|_F > 0$. We show that there exists an asymptotically valid and consistent test procedure only if $\|R\|_F \to \infty$ as the number of vertices n diverges. When the graphs and their pairwise correlations are generated from a latent position model, we propose an asymptotically valid and consistent test procedure that also runs in time polynomial in n. We now mention two directions for future research.

Comparing the theoretical results in Theorem 6 and Theorem 7 against either Remark 2 or Theorem 5, we see that while $\|R\|_F \to \infty$ for all of these examples, it is nevertheless easier, both statistically and computationally, to detect $R \neq 0$ when it has some structure. Indeed, for both Theorem 6 and Theorem 7 we have $R_{ij} = f(m_{ij})$ where f is a smooth function and the matrix $M = (m_{ij})$ is low-rank. In contrast, the matrix R in Remark 2 and Theorem 5 is either completely random or has no low-rank structure. Therefore, while $R \neq 0$ if and only if $\|R\|_F > 0$, the magnitude of $\|R\|_F$ itself is not sufficiently refined to distinguish between the simple and more difficult settings for $R \neq 0$. Determining the right measure of the correlation between graphs is thus of both theoretical and practical interest, especially if this measure also leads to thresholds that are both necessary and sufficient for our independence testing problem.

Continuing on the above theme, the critical region for our test statistics in Section 3.4.1 and Section 3.4.3 are based on bootstrapping graphs from the estimated edge probabilities matrices (see e.g., Algorithm 1). The validity of these resampling techniques is justified by the empirical simulation studies as well as real data analysis. However, bootstrap sampling of a graph on n vertices generally requires $O(n^2)$ time and $O(n^2)$ memory, which can be prohibitive if n is large. Therefore our test procedures could be more robust and computationally efficient if we are able to derive the limiting distribution of the test statistics in Theorem 6 and Theorem 8 and thereby obtain approximate critical values. We surmise, however, that this will be a quite technical and challenging problem as it requires substantial refinement of all existing results for USVT as these exclusively focus on upper bounds for the estimation error in Frobenius norm.

Finally, it will also be useful to study other formulations of independence testing for graphs, e.g., by not assuming that they are marginally inhomogeneous Erdős-Rényi graphs, or by considering more complex correlation structures. A natural and interesting example of this latter type of problem is when we have three or more graphs as their joint distributions cannot be specified using only the marginal distributions and pairwise edges correlations.

CHAPTER

4

INFERENCE ON MULTIPLE RANDOM GRAPHS

4.1 Introduction

In the previous chapters, we explored two-sample hypothesis testing and independence testing between two graphs. This chapter will extend our focus to statistical inference on multiple random graphs. Specifically, we will propose a new joint inference method and validate its effectiveness across various scenarios. This approach allows for the simultaneous analysis of multiple graphs, providing deeper insights into their underlying structures and relationships.

Graphs are ideal for representing complex systems, where there are different objects represented by vertices and their pairwise relationships represented by edges. Statistical inference across multiple graphs is extensively studied as it holds vital interdisciplinary interest in diverse fields such as machine learning, neuroscience, transportation systems, social sciences, and epidemiology (Athreya et al. 2018; Kong et al. 2021; Cardillo et al. 2013; Kim et al. 2021).

Inference typically relies on effective low-dimensional representations of graphs, often achieved through spectral decompositions (Belkin and Niyogi 2003; Sussman et al. 2012). In the context of stochastic block models, multilayer networks facilitate the clustering of vertices,

enhancing our understanding of graph structures across different layers (Bhattacharyya and Chatterjee 2018; Lei and Lin 2023; Huang et al. 2023). Our focus is also on node-aligned graphs that share a common vertex set. Tang et al. (2017a, 2014) have developed tests to assess the similarity between two such graphs with latent models. As the number of graphs increases, the complexity of inference grows, necessitating research that extends beyond simple pairwise comparisons.

In this regard, Levin et al. (2017); Draves and Sussman (2020) have broadened the scope of these tests to include generalized multiple graphs through joint low-dimensional embedding with the Omnibus matrix. The approach enables them to access latent positions across all graphs and simultaneously infer issues related to these positions. Additionally, the embedding is particularly useful for various inferences, such as community detection, vertex classification, hypothesis testing, and anomaly detection (Pantazis et al. 2022; Jones and Rubin-Delanchy 2020; Chen et al. 2020).

Despite these advances, most existing approaches presume that all graphs share the same distribution or have significantly similar distributions. Specifically, the theoretical properties of the Omnibus embedding are established for each random adjacency matrix marginally distributed according to a random dot product graph model with the same latent positions or for a group of random dot product graph models with a specific structure, such as eigen-scaling random dot product graph models. Under these conditions, the column spaces of the latent positions or the probability matrices for all random graphs are identical. Additionally, the rank of combination of latent positions for all random graphs is kept as the dimension of the latent position, so it is sufficient to use the same dimension for the Omnibus embedding.

Our work diverges from this norm by not imposing any distributional constraints across the graphs. Without such constraints among graphs, the column spaces can differ, and the rank can be larger. Therefore, we need an Omnibus embedding with a possible larger dimension to include the information for all graphs. We adopt a generalized random dot product graph model, presenting a more flexible and encompassing framework compared to those typically explored in the literature. This approach allows for a more nuanced exploration of the underlying structures and relationships within and between multiple graph datasets.

In Section 4.2, we introduce the Omnibus embedding, discuss the limitations of existing results, and present our approach to overcoming these limitations. In Section 4.3, we provide theoretical results on the convergence and asymptotic normality of the Omnibus embedding and the estimated latent positions when all graphs are generalized random dot product graph models. In Section 4.4, we state similar results for graphs with Gaussian errors. In Section 4.5, we conduct simulations to demonstrate behavior of latent position estimation, hypothesis testing, and community detection from our Omnibus embedding. In Section 4.6, we explore

real data with our Omnibus embedding.

4.2 Preliminary

The Random Dot Product Graph (RDPG) is a common random graph model that can successfully approximate a wide range of random graphs, from simple to complex random graphs. Recall Definition 1.

The adjacency spectral embedding (ASE) of Sussman et al. (2013) helps to estimate the latent positions of a single RDPG. Consider a collection of RDPGs on the same vertex set with known correspondence, Levin et al. (2017) provide an Omnibus embedding for all the RDPGs. This embedding is from a large matrix combining those graphs named Omnibus matrix, defined as below. Let $\{A^{(k)}\}_{k=1}^{m} \in \mathbb{R}^{n\times n}$ be a set of undirected, vertex-aligned, adjacency matrices. Let $M_A \in \mathbb{R}^{mn\times mn}$ be the omnibus matrix of $\{A^{(k)}\}_{k=1}^{m}$ given by

$$M_A = \begin{pmatrix} A^{(1)} & \frac{1}{2}(A^{(1)}+A^{(2)}) & \dots & \frac{1}{2}(A^{(1)}+A^{(m)}) \\ \frac{1}{2}(A^{(2)}+A^{(1)}) & A^{(2)} & \dots & \frac{1}{2}(A^{(2)}+A^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}(A^{(m)}+A^{(1)}) & \frac{1}{2}(A^{(m)}+A^{(2)}) & \dots & A^{(m)} \end{pmatrix}.$$

We can consider the expected value of the Omnibus matrix and extend it to probability matrices. Let $\{P^{(k)}\}_{k=1}^{m} \in \mathbb{R}^{n\times n}$ be a set of probability matrices. Let $M_P \in \mathbb{R}^{mn\times mn}$ be the omnibus matrix of $\{P^{(k)}\}_{k=1}^{m}$ given by

$$M_P = \begin{pmatrix} P^{(1)} & \frac{1}{2}(P^{(1)}+P^{(2)}) & \dots & \frac{1}{2}(P^{(1)}+P^{(m)}) \\ \frac{1}{2}(P^{(2)}+P^{(1)}) & P^{(2)} & \dots & \frac{1}{2}(P^{(2)}+P^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}(P^{(m)}+P^{(1)}) & \frac{1}{2}(P^{(m)}+P^{(2)}) & \dots & P^{(m)} \end{pmatrix}.$$

We use one Omnibus matrix M_A or M_P to show the information of all the adjacency matrices or all the probability matrices. Levin et al. (2017) provide the convergence about the Omnibus embedding on the Joint Random Dot Product Graph (JRDPG) defined as below.

Definition 9. We say that random graphs $A^{(1)}, \dots, A^{(m)}$ are distributed as a joint random dot product graph (JRDPG) if $X = [X_1|\dots|X_n]^T \in \chi_d^n$ and we have the distributions $A^{(k)} \sim \text{RDPG}(X)$ independently for $k = 1, 2, \dots, m$.

For a JRDPG, all graphs are generated from the same latent positions X. Levin et al. (2017) show that the estimates $\hat{X}$ generated from the ASE of the Omnibus matrix M_A converge

to X, up to some orthogonal transformation W and furthermore the rows of $\hat{X}W - X$ are asymptotically (multivariate) normals.

Draves and Sussman (2020) provide the convergence on the Eigen-Scaling Random Dot Product Graph (ESRDPG) defined as below.

Definition 10. Let $C^{(1)}, \ldots, C^{(m)} \in \{C \in \mathbb{R}^{d\times d} : C \text{ is diagonal and non-negative}, X_i^\top C X_j \in [0,1], \forall X \in \chi_d^n, 1 \le i, j \le n\}$ with property that $\min_{1\le i\le d} \max_{1\le k\le m} C_{ii}^{(k)} > 0$. We say that random graphs $A^{(1)}, \ldots, A^{(m)}$ are distributed according to the Eigen-Scaling Random Dot Product Graph (ESRDPG) if $X = [X_1|\ldots|X_n]^T \in \chi_d^n$ and $A^{(k)} \sim \text{RDPG}(X\sqrt{C^{(k)}})$ independently for $k = 1, 2, \ldots, m$.

This model is more general than JRDPG. The latent positions for graphs can be different, but still with some limitations. For example, the column space of any $X\sqrt{C^{(k)}}$ is the same as that for X. In this chapter we consider more general models where the latent positions of all graphs are arbitrary and can be totally different. We provide a method to recover the latent positions for every graph, even though graphs have different marginal distributions. In particular the spectral embedding of our Omnisbus matrix M_A might require up to md dimensions, which is considerably larger than the d dimensions for the models in Levin et al. (2017) and Draves and Sussman (2020).

We now introduce the Generalized Random Dot Product Graph (GRDPG) model to allow for edge probability matrices that are symmetric but not positive semidefinite.

Definition 11 (Generalized Random Dot Product Graph)**.** Let χ_d^n be defined by

$$\chi_d^n = \{M \in \mathbb{R}^{n\times d} : MM^T \in [0,1]^{n\times n} \text{ and } \text{rank}(M) = d\},$$

and let $X = [X_1|\ldots|X_n]^T \in \chi_d^n$. Suppose A is a random adjacency matrix given by

$$\mathbb{P}[A|X] = \prod_{i<j} (X_i^T I_{p,q} X_j)^{A_{ij}} (1 - X_i^T I_{p,q} X_j)^{1-A_{ij}},$$

where $p + q = d$, then it is denoted by $A \sim \text{GRDPG}(X)$ and say that A is the adjacency matrix of a generalized random dot product graph with latent position X of rank d with signature (p, q). Also, A can be presented as $A \sim \text{Bernoulli}(P)$, where $P = XI_{p,q}X^T$. The matrix X represents the latent position.

4.3 Results and Implementation

In this dissertation, we provide a method using the Omnibus matrix with GRDPG blocks to obtain latent positions of m generalized random dot product graphs simultaneously. We assume that all graphs are on the same vertex set with known vertex correspondence.

Let $X^{(k)} \in \chi_d^n$, and $A^{(k)} \sim \text{GRDPG}(X^{(k)})$ with signature (p_k, q_k) for $1 \leq k \leq m$. Also, we have $P^{(k)} = X^{(k)} I_{p_k,q_k} (X^{(k)})^\top$. Set Omnibus matrices $M_P = \text{Omni}(P^{(1)}, P^{(2)}, \ldots, P^{(k)})$ and $M_A = \text{Omni}(A^{(1)}, A^{(2)}, \ldots, A^{(k)})$. Let $d' := \text{rank}(\begin{pmatrix} X^{(1)} & X^{(2)} & \ldots & X^{(m)} \end{pmatrix})$ and $d'' := \text{rank}(M_P)$. We now list several assumptions on the latent positions $\{X^{(i)}\}_{i=1}^m$ and the associated omnibus matrix M_P.

Assumption 1. Let $VSV^\top$ be the eigen-decomposition of M_P.

A) $C_1\sqrt{n\rho} \leq \|X^{(k)}\| \leq C_2\sqrt{n\rho}$ for some constant $C_1, C_2 > 0$, where $(0, 1] \ni \rho \to c \in [0, 1]$ as $n \to \infty$, with $n\rho \geq C_3(\log n)^{2\delta^*}$ for some constants $C_3 > 0, \delta^* > 1$.

B) $\sqrt{n}\|V\|_{2\to\infty} \leq C_4$ for some constant $C_4 > 0$.

C)

$$\frac{\sigma_1(\begin{pmatrix} X^{(1)} & X^{(2)} & \ldots & X^{(m)} \end{pmatrix})}{\sigma_{d'}(\begin{pmatrix} X^{(1)} & X^{(2)} & \ldots & X^{(m)} \end{pmatrix})} \leq C_5,$$

for some constant $C_5 > 0$.

D)

$$\frac{\sigma_1(M_P)}{\sigma_{d''}(M_P)} \leq C_6,$$

for some constant $C_6 > 0$.

Assumption A introduces a sparsity scaling for latent positions on a global scale rather than on individual ones. This is weaker than assuming every single individual latent position has a sparsity scaling.

Assumption B shows that the eigenvectors of the Omnibus matrix have bounded coherence, which is a common mild assumption in many high-dimensional statistics problems including matrix completion, covariance estimation and random graph inference (Abbe et al. 2020; Candes and Recht 2012; Cape et al. 2019b; Fan et al. 2018; Lei 2019).

Assumptions C and D specify the magnitude of the leading eigenvalues of interest for the sum of all graphs and the Omnibus matrix.

Remark 5. If $\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix}$ has full column rank, then

$$\frac{\sigma_1(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})}{\sigma_{d'}(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})} \leq C_5 \Rightarrow \frac{\sigma_1(M_P)}{\sigma_{d''}(M_P)} \leq C_6 \Rightarrow \frac{\sigma_1(X_i^{(k)})}{\sigma_d(X_i^{(k)})} \leq C_7^{(k)}$$

so that condition (C) implies condition (D). However, if $(X^{(1)}\ X^{(2)}\ \dots\ X^{(m)})$ is not of full column rank then conditions (C) and (D) are distinct and neither one implies the other.

With the above assumptions in place, we now state several results for the estimate of $X^{(i)}$ generated from the adjacency spectral embedding of M_A. Set $\begin{pmatrix} V_1^\top & V_2^\top & \dots & V_m^\top \end{pmatrix} := V^\top$, where V_k's have the same size for $1 \leq k \leq m$. Let $U\Sigma U^\top$ be the eigen-decomposition of $\frac{1}{m}\sum_{k=1}^m V_k V_k^\top$. Let $\hat{V}\hat{S}\hat{V}^\top$ be the truncated eigendecomposition of M_A to keep only the d'' largest eigenvalues. Set $\begin{pmatrix} \hat{V}_1^\top & \hat{V}_2^\top & \dots & \hat{V}_m^\top \end{pmatrix} := \hat{V}^\top$, where $\hat{V}_k$'s have the same size for $1 \leq k \leq m$. Let $\hat{U}\hat{\Sigma}\hat{U}^\top$ be the top-d' eigen-decomposition of $\frac{1}{m}\sum_{k=1}^m \hat{V}_k\hat{V}_k^\top$. Let $\Xi_i^\circ \in \mathbb{R}^{mn\times mn}$ be a diagonal matrix with diagonal elements

$$\text{vec}[\begin{pmatrix} X^{(1)}(X^{(1)})_i & X^{(2)}(X^{(2)})_i & \dots & X^{(m)}(X^{(m)})_i \end{pmatrix}],$$

and $(\Xi_i^*)^2 = \Xi_i^\circ - (\Xi_i^\circ)^2$. Note that $\Xi_i^2 = (\Xi_i^*)^2[I - \text{diag}(e_i + e_{i+n} + \dots + e_{i+(m-1)n})]$ is the covariance matrix of $\begin{pmatrix} A^{(1)} & A^{(2)} & \dots & A^{(m)} \end{pmatrix}_i$.

Lemma 1. Suppose $X^{(k)} \in \chi_d^n$.Then we have $\mathcal{C}(U) = \mathcal{C}(V_k) = \mathcal{C}(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})$, for $1 \leq k \leq m$.

We first state some results about the convergence of the estimated latent positions.

Lemma 2. Suppose that Assumption 1 holds. Then there exists a $d' \times d'$ orthogonal matrix W such that

$$\|\hat{U} - UW\| = O_{\mathbb{P}}(\frac{1}{\sqrt{n\rho}}), \quad \|\hat{U} - UW\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}).$$

Recall Section 2.2.1, we use $f(n) = O_{\mathbb{P}}(g(n))$ if $f(n) = O(g(n))$ with the probability at least $1 - n^{-C}$ for any $C > 0$. The analogous notation applies for $o_{\mathbb{P}}(\cdot), \omega_{\mathbb{P}}(\cdot)$.

Theorem 9. Suppose that Assumption 1 holds. Let $\hat{R}_{X^{(1)}}$ be the d-dimensional adjacency spectral embedding of $\hat{U}^\top A^{(1)}\hat{U}$. Let $\hat{X}^{(1)} := \hat{U}\hat{R}_{X^{(1)}}$. Then there exist matrices $W_{X^{(1)}} \in \mathcal{O}(d)$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\|\hat{X}^{(1)} - X^{(1)}QW_{X^{(1)}}\| = O_{\mathbb{P}}(1), \quad \|\hat{X}^{(1)} - X^{(1)}QW_{X^{(1)}}\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{\sqrt{n}}).$$

Theorem 10. Suppose that Assumption 1 holds. Let $\hat{R}^{(4)}_{X^{(1)}}$ be the rank-d truncated SVD of $\hat{U}^\top (A^{(1)})^2 \hat{U}$. Let $\hat{R}^{(2)}_{X^{(1)}}$ be the square root of $\hat{R}^{(4)}_{X^{(1)}}$ and $\hat{R}^{(1)}_{X^{(1)}}$ be the square root of $\hat{R}^{(2)}_{X^{(1)}}$. Let $\tilde{X}^{(1)} := \hat{U}\hat{R}^{(1)}_{X^{(1)}}$. Then there exist matrices $U^* \in \mathbb{R}^{d' \times d}$ satisfying $(U^*)^\top U^* = I_d$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\|\tilde{X}^{(1)} - X^{(1)} Q (U^*)^\top\| = O_{\mathbb{P}}(1), \quad \|\tilde{X}^{(1)} - X^{(1)} Q (U^*)^\top\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{\sqrt{n}}).$$

We next state some results about the row-wise asymptotic normality of V. These results depend on the following assumption on the ℓ_2 norms of the rows of M_P.

Assumption 2. $\frac{\max_i \|(M_P)_i\|}{\min_i \|(M_P)_i\|} \le C_{10}$ for some constant $C_{10} > 0$.

Lemma 3. Suppose that Assumptions 1 and 2 hold. Then there exists a $d' \times d'$ orthogonal matrix W such that

$$\sqrt{mn}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(mn) \to \mathcal{N}(0, I_{d'}),$$

where $\bar{V}^* = (mn)^{-1}(\Xi_i^* V^*)^\top (\Xi_i^* V^*)$,

$$V^* = \frac{1}{2}\begin{pmatrix}(M_1 + I_m)\otimes I_n & (M_2 + I_m)\otimes I_n & \dots & (M_m + I_m)\otimes I_n\end{pmatrix}\begin{pmatrix} VS^{-1}V_1^\top U\Sigma^{-1} \\ VS^{-1}V_2^\top U\Sigma^{-1} \\ \vdots \\ VS^{-1}V_m^\top U\Sigma^{-1}\end{pmatrix},$$

$M_k \in \mathbb{R}^{m\times m}$ is the matrix whose elements in k-th row are 1 and 0 otherwise for $1 \le k \le m$, U_i and $\hat{U}_i$ for $1 \le i \le n$ are the rows of U and $\hat{U}$.

Remark 6. The matrix $\frac{1}{2}\begin{pmatrix}(M_1 + I_m)\otimes I_n & (M_2 + I_m)\otimes I_n & \dots & (M_m + I_m)\otimes I_n\end{pmatrix}$ for $m = 2$ and $m = 3$ are given by

$$\begin{pmatrix} I_n & I_n/2 & I_n/2 & 0 \\ 0 & I_n/2 & I_n/2 & I_n\end{pmatrix}$$

$$\begin{pmatrix} I_n & I_n/2 & I_n/2 & I_n/2 & 0 & 0 & I_n/2 & 0 & 0 \\ 0 & I_n/2 & 0 & I_n/2 & I_n & I_n/2 & 0 & I_n/2 & 0 \\ 0 & 0 & I_n/2 & 0 & 0 & I_n/2 & I_n/2 & I_n/2 & I_n\end{pmatrix}.$$

Similarly, for $m = 2$ we have,

$$V^* = \begin{pmatrix} V_1 S^{-1} V_1^\top + \frac{1}{2}(V_1 S^{-1} V_2^\top + V_2 S^{-1} V_1^\top) \\ V_2 S^{-1} V_2^\top + \frac{1}{2}(V_1 S^{-1} V_2^\top + V_2 S^{-1} V_1^\top) \end{pmatrix} U \Sigma^{-1}.$$

Remark 7. Lemma 3 implies that $n\rho^{1/2}(W\hat{U}_i - U_i)$ can be approximated by $\mathcal{N}\Big(0, n^2\rho (V^*)^\top (\Xi_i^*)^2 V^*\Big)$ where the covariance matrix $(V^*)^\top (\Xi_i^*)^2 V^*$ is of full-rank with eigenvalues of order $\Theta(1)$.

Theorem 11. Suppose that Assumptions 1 and 2 hold. Let $\hat{R}_{X^{(1)}}$ be the d-dimensional adjacency spectral embedding of $\hat{U}^\top A^{(1)} \hat{U}$. Let $\hat{X}^{(1)} := \hat{U}\hat{R}_{X^{(1)}}$. Then there exist matrices $W_{X^{(1)}} \in \mathcal{O}(d)$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\sqrt{mn}((X^{(1)})^\top U \bar{V}^* U^\top X^{(1)})^{-1/2}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)})/(mn) \to \mathcal{N}(0, I_d),$$

where $\bar{V}^*$ is defined in Lemma 3 and $\hat{X}_i^{(1)}, X_i^{(1)}$ for $1 \le i \le n$ are the rows of $\hat{X}^{(1)}$ and $X^{(1)}$.

Remark 8. We have $\sqrt{n}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)})$ approximates

$$\mathcal{N}\Big(0, n(X^{(1)})^\top U (V^*)^\top (\Xi_i^*)^2 V^* U^\top X^{(1)}\Big).$$

The covariance matrix has eigenvalues of $\theta(1)$ with full rank. Also, if

$$n\rho^{1/2}(W\hat{U}_i - U_i) \approx \mathcal{N}(0, \Gamma_i),$$

then

$$\sqrt{n}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)}) \approx \mathcal{N}(0, (X^{(1)})^\top U \Gamma_i U^\top X^{(1)}/(n\rho))$$

for $QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)}$ approximating $(X^{(1)})^\top U(W\hat{U}_i - U_i)$.

Theorem 12. Suppose that Assumptions 1 and 2 hold. Let $\hat{R}^{(4)}_{X^{(1)}}$ be the d-TSVD about $\hat{U}^\top (A^{(1)})^2 \hat{U}$. Let $\hat{R}^{(2)}_{X^{(1)}}$ be the square root of $\hat{R}^{(4)}_{X^{(1)}}$ and $\hat{R}^{(1)}_{X^{(1)}}$ be the square root of $\hat{R}^{(2)}_{X^{(1)}}$. Let $\tilde{X}^{(1)} := \hat{U}\hat{R}^{(1)}_{X^{(1)}}$. Then there exist matrices $W \in \mathcal{O}(d)$, $U^* \in \mathbb{R}^{d' \times d}$ satisfying $(U^*)^\top U^* = I_d$ and $Q \in \mathcal{O}(p_1, q_1)$ such that $W\tilde{X}_i^{(1)} - U^* Q X_i^{(1)} = R^{(1)}_{X^{(1)}}(W\hat{U}_i - U_i) + o_{\mathbb{P}}(\frac{1}{\sqrt{n}})$, where $\tilde{X}_i^{(1)}$ and $X_i^{(1)}$ for $1 \le i \le n$ are the rows of $\tilde{X}^{(1)}$ and $X^{(1)}$, and $R^{(1)}_{X^{(1)}}$ is the non-negative square root of $U^\top P^{(1)} U$.

Remark 9. Note that we do not have asymptotic normality in Theorem 12 because the $R^{(1)}_{X^{(1)}}$ are not full rank. Nevertheless, if $n\rho^{1/2}(W\hat{U}_i - U_i)$ is distributed approximately as $\mathcal{N}(0, \Gamma_i)$, then $\sqrt{n}(W\tilde{X}_i - U^* Q X_i)$ is distributed approximately as $\mathcal{N}(0, R_X^{(1)} \Gamma_i R_X^{(1)}/(n\rho))$. The covariance matrix $R_X^{(1)} \Gamma_i R_X^{(1)}$ is now of rank d with non-zero eigenvalues of order $\Theta(1)$.

Based on Theorem 9, we know the estimated latent positions can be close to the true latent positions through an orthogonal transformation. Furthermore, for the corresponding estimated and true latent positions, the difference between them follows an asymptotically normal distribution from Theorem 11. So we provide a method to recover the latent positions. With the information about the latent positions, we are able to conduct many inferences, such as sample tests on the graphs and community detection for the vertices. They are based on arbitrary m d-dimensional GRDPGs without specific limitations, compared with JRDGP and ESRDPG.

Some advantages of this method to obtain estimated latent positions are robustness and not requiring the relation among graphs. The convergence rate of this method has the same order with the convergence rate of using the adjacency spectral embedding directly on a single graph. However, in practice, if all graphs have the same distribution, estimators with more samples have less standard error. Hence, our estimators behave better than the adjacency spectral embedding from a single graph under the situations where all graphs have the same distribution. Under this situation, one of common ways is to get the adjacency spectral embedding from a linear combination of all adjacency matrices for all graphs, and usually, it is average of adjacency matrices. But if at least one graph has different distributions, the combination leads to a wrong estimator. As we can see, our estimator converges to the true latent positions up to an orthogonal transformation, no matter whether the graphs have the same distribution. This makes our method have more application situations.

Based on Theorems 10, 12, there exists an estimated latent positions that might be in a higher dimensional Euclidean space. There is a linear isometry from the original Euclidean space to the higher dimensional Euclidean space, such that the difference between the updated estimated latent positions and the true latent positions up to the linear isometry is close and the difference has an asymptotically normal distribution. So we know $\tilde{X}_i - \tilde{X}_j \approx U^*(Z_i - Z_j)$, which leads to that $\|\tilde{X}_i - \tilde{X}_j\| \approx \|Z_i - Z_j\|$, where $(U^*)^\top U^* = I$ and Z is the adjacency spectral embedding of P. It shows that this updated higher dimensional estimator preserves the pair-wise distance between latent positions as well. We can use the updated estimates to express probability matrices directly and conduct the testing procedure, omitting the orthogonal Procrustes process.

Besides, we can obtain that, if $P^{(1)} = P^{(2)}$, we know $\tilde{X}^{(1)} \approx \tilde{X}^{(2)}$. On the other hand, we consider the difference between $P^{(1)}$ and $P^{(1)}$ assuming $p_1 = p_2$. If $\mathcal{C}(Z^{(1)}) = \mathcal{C}(Z^{(2)})$, we have $\|\tilde{X}^{(1)} - \tilde{X}^{(2)}\|_F \approx \|Z^{(1)}W_1U_1^\top - Z^{(2)}W_2U_2^\top\|_F \geq \min_{W^* \in \mathcal{W}(p_1,q_1)} \|Z^{(1)} - Z^{(2)}W^*\|_F$, where $Z^{(1)}$ and $Z^{(1)}$ are adjacency spectral embedding of $P^{(1)}$ and $P^{(2)}$, $U_1^\top U_1 = U_2^\top U_2 = I_n$,

$W_1, W_2 \in \mathcal{W}(p_1, q_1)$, and

$$\mathcal{W}(a,b) = \{W = \begin{pmatrix} W_{11} & 0 \\ 0 & W_{22} \end{pmatrix} : W_{11} \in \mathcal{O}(a), W_{22} \in \mathcal{O}(b)\}.$$

If $\mathcal{C}(Z^{(1)}) \neq \mathcal{C}(Z^{(2)})$, we have $\|\tilde{X}^{(1)} - \tilde{X}^{(2)}\|_F \geq \|P_{Z^{(1)}} Z^{(2)} - Z^{(2)}\|_F$ and $\mathcal{C}(\tilde{X}^{(1)}) \approx \mathcal{C}(P^{(1)})$, $\mathcal{C}(\tilde{X}^{(2)}) \approx \mathcal{C}(P^{(2)})$. Therefore, if $P^{(1)} \neq P^{(2)}$, it is likely to observe that $\tilde{X}^{(1)} \neq \tilde{X}^{(2)}$. Furthermore, if $P^{(1)} \approx P^{(2)}$, then $\tilde{X}^{(1)} \approx \tilde{X}^{(2)}$. However, this approximation does not hold for $\hat{X}^{(1)}$ and $\hat{X}^{(2)}$ in certain cases, such as when $X^{(1)}$ has some equal singular values.

Hence, we can use estimates $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$ to express $P^{(1)}$ and $P^{(2)}$ directly. Generally, we can infer that the closer $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$ are, the closer $P^{(1)}$ and $P^{(2)}$ are, without any orthogonal transformation. We can use the estimators in this way to conduct testing procedure, without the orthogonal Procrustes process. For example, we are able to directly conduct multiple sample testing for the distributions of all graphs, like $T = \sum_{i=1}^{m} \|\tilde{X}^{(i)} - \bar{\tilde{X}}\|_F^2$, with less computing complexity.

4.4 Extension

Besides the Bernoulli error, we can also recover the latent positions based on Gaussian errors under some certain conditions.

Compered with the setting in 4.3, $P^{(k)} = f(n) X^{(k)} I_{p_k,q_k} (X^{(k)})^\top$, $A^{(k)} = P^{(k)} + E_{ij}^{(k)}$, where E is a symmetric, and $E_{ij}^{(k)} \overset{\text{i.i.d}}{\sim} \mathcal{N}\big(0, g(n)^2 (\sigma_{ij}^{(k)})^2\big)$ for $i < j$, $E_{ii}^{(k)} = 0$ for $1 \leq i \leq n$.

Consider the following assumption.

Assumption 3. Let $M_P = VSV^\top$ be the eigen-decomposition.

1. $C_1\sqrt{n} \leq \|X^{(k)}\| \leq C_2\sqrt{n}$ for some constant $C_1, C_2 > 0$.
2. $\sqrt{n}\|V\|_{2\to\infty} = C_3$ for some constant $C_3 > 0$.
3. $$\frac{\sigma_1(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})}{\sigma_{d'}(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})} \leq C_4,$$

   for some constant $C_4 > 0$.

4.

$$\frac{\sigma_1(M_P)}{\sigma_{d''}(M_P)} \leq C_5,$$

for some constant $C_5 > 0$,

5. $(0, \infty) \ni f(n) \to c \in [0, \infty]$ as $n \to \infty$, $\frac{g(n)}{f(n)} = \omega(\frac{1}{\sqrt{n}})$,

6. Matrices $\Sigma^{(k)}$ satisfy $C_6\sqrt{n} \leq \|\Sigma^{(k)}\|_{2\to\infty} \leq C_7\sqrt{n}$, $C_8 \leq \max_{ij} |\sigma_{ij}^{(k)}| \leq C_9$ for some constants $C_6, C_7, C_8, C_9 > 0$.

We can consider $(P^{(k)})^* = P^{(k)}/f(n) = X^{(k)}(X^{(k)})^\top$, $(A^{(k)})^* = A^{(k)}/f(n) = P^{(k)}/f(n) + E^{(k)}/f(n) = (P^{(k)})^* + (E^{(k)})^*$, where $(E^{(k)})^* = \mathcal{N}\big(0, \frac{g(n)^2}{f(n)^2}(\sigma_{ij}^{(k)})^2\big)$. Hence, it is reasonable to change the fifth item in Assumption 3 to $(0, \infty) \ni g(n) \to c \in [0, \infty]$ as $n \to \infty$, $g(n) = \omega(\frac{1}{\sqrt{n}})$ with $P^{(k)} = X^{(k)} I_{p_k,q_k} (X^{(k)})^\top$. Then we can have the below modified model and assumptions for analysis.

We set $P^{(k)} = X^{(k)} I_{p_k,q_k} (X^{(k)})^\top$, $A^{(k)} = P^{(k)} + E_{ij}^{(k)}$, where E is a symmetric, and $E_{ij}^{(k)} \overset{\text{i.i.d}}{\sim} \mathcal{N}\big(0, g(n)^2(\sigma_{ij}^{(k)})^2\big)$ for $i < j$, $E_{ii}^{(k)} = 0$ for $1 \leq i \leq n$.

Assumption 4. Let $M_P = VSV^\top$ be the eigen-decomposition.

1. $C_1\sqrt{n} \leq \|X^{(k)}\| \leq C_2\sqrt{n}$ for some constant $C_1, C_2 > 0$.

2. $\sqrt{n}\|V\|_{2\to\infty} = C_3$ for some constant $C_3 > 0$.

3.

$$\frac{\sigma_1(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})}{\sigma_{d'}(\begin{pmatrix} X^{(1)} & X^{(2)} & \dots & X^{(m)} \end{pmatrix})} \leq C_4,$$

for some constant $C_4 > 0$.

4.

$$\frac{\sigma_1(M_P)}{\sigma_{d''}(M_P)} \leq C_5,$$

for some constant $C_5 > 0$,

5. $g(n) = \omega(\frac{1}{\sqrt{n}})$,

6. Matrices $\Sigma^{(k)}$ satisfy $C_6\sqrt{n} \leq \|\Sigma^{(k)}\|_{2\to\infty} \leq C_7\sqrt{n}$, $C_8 \leq \max_{ij} |\sigma_{ij}^{(k)}| \leq C_9$ for some constants $C_6, C_7, C_8, C_9 > 0$.

We use $\rho := g^2(n)$ because it expresses the magnitude of the error. The reason we have a lower bound for ρ is that when ρ is too small, the off-diagonal random errors cannot eliminate the diagonal fixed error to maintain the centered normality asymptotically. We can also use the estimated latent positions to approximate the true latent positions, but the difference between them does not follow the normality for extreme small ρ. If we assume that the diagonal errors can be small, like $\max_i |A_{ii} - P_{ii}| = O(\rho)$, there still exists the normality and ρ can be arbitrary small.

The notations are the same as in 4.3, except that $\Xi_i^* \in \mathbb{R}^{mn \times mn}$ be a diagonal matrix with diagonal elements

$$\text{vec}[\sqrt{\rho}\begin{pmatrix}\Sigma_i^{(1)} & \Sigma_i^{(2)} & \dots & \Sigma_i^{(m)}\end{pmatrix}].$$

We have the following similar results as for GRDPG.

Lemma 4. Suppose that Assumption 4 holds. Then there exists an orthogonal square matrix $W \in \mathbb{R}^{d' \times d'}$ such that

$$\|\hat{U} - UW\| = O_{\mathbb{P}}(\sqrt{\frac{\rho}{n}}), \quad \|\hat{U} - UW\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n}).$$

Theorem 13. Suppose that Assumption 4 holds. Let $\hat{R}_{X^{(1)}}$ be the d-dimensional adjacency spectral embedding of $U^\top A^{(1)} U$. Let $\hat{X}^{(1)} := \hat{U}\hat{R}_{X^{(1)}}$. Then there exist matrices $W_{X^{(1)}} \in \mathcal{O}(d)$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\|\hat{X} - XQW_{X^{(1)}}\| = O_{\mathbb{P}}(\sqrt{\rho}), \quad \|\hat{X} - XQW_{X^{(1)}}\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{\sqrt{n}}).$$

Theorem 14. Suppose that Assumption 4 holds. Let $\hat{R}^{(4)}_{X^{(1)}}$ be the d-TSVD about $\hat{U}^\top (A^{(1)})^2 \hat{U}$. Let $\hat{R}^{(2)}_{X^{(1)}}$ be the square root of $\hat{R}^{(4)}_{X^{(1)}}$ and $\hat{R}^{(1)}_{X^{(1)}}$ be the square root of $\hat{R}^{(2)}_{X^{(1)}}$. Let $\tilde{X}^{(1)} := \hat{U}\hat{R}^{(1)}_{X^{(1)}}$. Then there exist matrices $U^* \in \mathbb{R}^{d' \times d}$ satisfying $(U^*)^\top U^* = I_d$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\|\tilde{X} - XQ(U^*)^\top\| = O_{\mathbb{P}}(\sqrt{\rho}), \quad \|\tilde{X} - XQ(U^*)^\top\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{\sqrt{n}}).$$

Lemma 5. Suppose that Assumption 4 holds. Then there exists an orthogonal square matrix $W \in \mathbb{R}^{d' \times d'}$ such that

$$\sqrt{mn}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(mn) \to \mathcal{N}(0, I_{d'}),$$

where $\bar{V}^* = (mn)^{-1}(\Xi_i^* V^*)^\top(\Xi_i^* V^*)$,

$$V^* = \frac{1}{2}\begin{pmatrix}(M_1 + I_m) \otimes I_n & (M_2 + I_m) \otimes I_n & \dots & (M_m + I_m) \otimes I_n\end{pmatrix}\begin{pmatrix}VS^{-1}V_1^\top U\Sigma^{-1} \\ VS^{-1}V_2^\top U\Sigma^{-1} \\ \vdots \\ VS^{-1}V_m^\top U\Sigma^{-1}\end{pmatrix},$$

$M_k \in \mathbb{R}^{m\times m}$ is the matrix whose elements in k-th row are 1 and 0 otherwise for $1 \leq k \leq m$, U_i and $\hat{U}_i$ for $1 \leq i \leq n$ are the rows of U and $\hat{U}$.

Remark 10. We have $n\rho^{-1/2}(W\hat{U}_i - U_i)$ approximates

$$\mathcal{N}\Big(0, n^2\rho^{-1}(V^*)^\top(\Xi_i^*)^2 V^*\Big).$$

The covariance matrix has eigenvalues of $\theta(1)$ with full rank.

Theorem 15. Assumption 4 holds. Let $\hat{R}_{X^{(1)}}$ be the d-dimensional adjacency spectral embedding of $U^\top A^{(1)} U$. Let $\hat{X}^{(1)} := \hat{U}\hat{R}_{X^{(1)}}$. Then there exist matrices $W_{X^{(1)}} \in \mathcal{O}(d)$ and $Q \in \mathcal{O}(p_1, q_1)$ such that

$$\sqrt{mn}((X^{(1)})^\top U\bar{V}^* U^\top X^{(1)})^{-1/2}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)})/(mn) \to I_d,$$

where $\bar{V}^*$ is defined in Lemma 3 and $\hat{X}_i^{(1)}, X_i^{(1)}$ for $1 \leq i \leq n$ are the rows of $\hat{X}^{(1)}$ and $X^{(1)}$.

Remark 11. We have $\sqrt{\frac{n}{\rho}}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)})$ approximates

$$\mathcal{N}\Big(0, n\rho^{-1}(X^{(1)})^\top U(V^*)^\top(\Xi_i^*)^2 V^* U^\top X^{(1)}\Big).$$

The covariance matrix has eigenvalues of $\theta(1)$ with full rank. Also, if

$$n\rho^{-1/2}(W\hat{U}_i - U_i) \approx \mathcal{N}(0, \Gamma_i),$$

then

$$\sqrt{\frac{n}{\rho}}(QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)}) \approx \mathcal{N}(0, (X^{(1)})^\top U\Gamma_i U^\top X^{(1)}/n)$$

for $QW_{X^{(1)}}\hat{X}_i^{(1)} - X_i^{(1)}$ approximating $(X^{(1)})^\top U(W\hat{U}_i - U_i)$.

Theorem 16. Assumption 4 holds. Let $\hat{R}^{(4)}_{X^{(1)}}$ be the d-TSVD about $\hat{U}^\top(A^{(1)})^2\hat{U}$. Let $\hat{R}^{(2)}_{X^{(1)}}$ be the square root of $\hat{R}^{(4)}_{X^{(1)}}$ and $\hat{R}^{(1)}_{X^{(1)}}$ be the square root of $\hat{R}^{(2)}_{X^{(1)}}$. Let $\tilde{X}^{(1)} := \hat{U}\hat{R}^{(1)}_{X^{(1)}}$. Then there exist matrices $W \in \mathcal{O}(d)$, $U^* \in \mathbb{R}^{d'\times d}$ satisfying $(U^*)^\top U^* = I_d$ and $Q \in \mathcal{O}(p_1, q_1)$ such

that $W\tilde{X}_i^{(1)} - U^*QX_i^{(1)} = R_{X^{(1)}}^{(1)}(W\hat{U}_i - U_i) + o_{\mathbb{P}}(\frac{1}{\sqrt{n}})$, where $\tilde{X}_i^{(1)}$ and $X_i^{(1)}$ for $1 \leq i \leq n$ are the rows of $\tilde{X}^{(1)}$ and $X^{(1)}$, and $R_{X^{(1)}}^{(1)}$ is the non-negative square root of $U^\top P^{(1)} U$.

Remark 12. We do not have the asymptotic normality because that $R_{X^{(1)}}^{(1)}$ is not full rank.

W can be W in Lemma 3. If $n\rho^{-1/2}(W\hat{U}_i - U_i)$ approximates $\mathcal{N}(0, \Gamma_i)$, then $\sqrt{\frac{n}{\rho}}(W\tilde{X}_i - U^*QX_i)$ approximates $\mathcal{N}(0, R_X^{(1)}\Gamma_i R_X^{(1)}/n)$. The covariance matrix has eigenvalues of $\theta(1)$ and 0 with the rank of d.

So for the Gaussian error, we can also use the spectral embedding to express the latent positions.

4.5 Simulations

4.5.1 Latent Positions Recovery

In this subsection, we estimate the performance of our estimated latent positions. We follow the recovery method ($\hat{X}, \tilde{X}$ defined in Theorems 9 and 10) to perform simulations to estimate the distance between estimated estimators and true latent positions. We generate $\{X_i^{(0)}\}_{i=1}^n$ as an i.i.d. sample from a multivariate normal with mean 0 and identity covariance matrix, and let $X_i = \frac{X_i^{(0)}}{\sqrt{2}\|X_i^{(0)}\|}$ for a standardization. In addition, we generate latent positions Y in the same way independently for different groups of latent positions ($P \neq Q$) or $Y = X$ for the same group of latent positions $P = Q$. Recall that $P = XX^\top$ and $Q = YY^\top$.

We observe a pair of graphs $A \sim \text{Bernoulli}(P)$ and $B \sim \text{Bernoulli}(Q)$, and use our method to obtain latent position estimates $\hat{X}, \hat{Y} \in \mathbb{R}^d$ and ones $\tilde{X}, \tilde{Y}$ in a higher-dimensional space $\mathbb{R}^{d'}$ to compare with the true latent positions. For $\hat{X}, \hat{Y}$, we focus on $\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F$ to evaluate the performance of approximating the true latent positions by estimated latent positions. For $\tilde{X}, \tilde{Y}$, we focus on $\min_{U^*U^{*\top}=I} \|\tilde{X} - XU^*\|_F$ to evaluate the performance of approximating the true latent positions by estimated latent positions. We repeat the recovery for 100 times to obtain the distance between the true latent positions and the estimated latent positions. The results are presented in Table 4.1 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$. We observe that for both ways, the estimated latent positions converge to the true latent positions row-wisely. The Frobenius norms should converge to a constant based on theoretical results, and $\hat{X}$ satisfy it. For $\tilde{X}$, the Frobenius norms grows slowly because n is not sufficiently large. Despite that, the convergence of the two-to-infinity norms support that we can use $\tilde{X}$ to express X. Also, the performances are not significantly different between whether two graphs are form the same distribution or not.

Table 4.1: Performance for Recovery: Outputs are the corresponding distances between estimated and true latent positions up to orthogonal transformations.

$d=2$	$P=Q$				$P\neq Q$			
	$\hat{X}$		$\tilde{X}$		$\hat{X}$		$\tilde{X}$	
n	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$
100	0.65	2.52	0.83	4.05	0.82	3.18	0.90	4.57
200	0.35	1.78	0.74	4.56	0.69	3.60	0.80	5.42
500	0.21	1.54	0.64	5.56	0.42	3.12	0.60	5.95
1000	0.14	1.54	0.57	6.55	0.30	2.94	0.53	6.57
2000	0.11	1.45	0.51	7.72	0.23	3.15	0.46	7.59
$d=3$	$P=Q$				$P\neq Q$			
	$\hat{X}$		$\tilde{X}$		$\hat{X}$		$\tilde{X}$	
n	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$	$\|\cdot\|_{2\to\infty}$	$\|\cdot\|_F$
100	0.84	3.88	0.96	5.37	0.89	4.14	1.00	5.64
200	0.69	4.00	0.83	6.03	0.84	5.13	0.92	6.96
500	0.31	2.65	0.71	6.94	0.65	5.89	0.76	8.54
1000	0.21	2.43	0.62	8.01	0.42	4.99	0.57	8.67
2000	0.15	2.35	0.56	9.46	0.29	4.69	0.48	9.40

4.5.2 Hypothesis Testing

In this subsection, we conduct sample tests to test whether two graphs are from the same distribution, which is also to test whether all the latent positions are the same up to orthogonal transformation. We generate $\{X_i\}_{i=1}^n$ as an i.i.d. sample from a multivariate normal with mean 0 and identity covariance matrix. Also, we generate $Y_{ij} = X_{ij} + \epsilon\delta_{ij}$, where $\delta_{ij} \sim \mathcal{N}(0,1)$ independently, and $\epsilon \geq 0$ express the difference between two groups of latent positions or two graphs. We use the cosine similarity

$$P_{ij} = \frac{|X_i^\top||X_j|}{2\|X_i\|\|X_j\|} \qquad Q_{ij} = \frac{|Y_i^\top||Y_j|}{2\|Y_i\|\|Y_j\|} \tag{4.1}$$

We observe a pair of graphs $A \sim \text{Bernoulli}(P)$ and $B \sim \text{Bernoulli}(Q)$, and use our method to obtain latent position estimates $\hat{X}, \hat{Y}$ and ones $\tilde{X}, \tilde{Y}$ in a higher-dimensional space $\mathbb{R}^{d'}$ to test $\mathbb{H}_0 \colon P = Q$ against $\mathbb{H}_A \colon P \neq Q$ where the rejection region is calibrated via bootstrapping with significance level 0.05. (see Algorithm 3) We repeat the above steps for 100 Monte Carlo replicates to obtain an empirical estimate of the power for T. The results are presented

in Table 4.2 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $\epsilon \in \{0, 0.1, 0.3, 0.5\}$. We observe that the type I error of the test statistic is well-controlled and that the test statistic also exhibits power even for small values of ϵ and moderate values of n. Furthermore, using d'-dimensional estimated latent positions $\tilde{X}, \tilde{Y}$ can obtain greater powers than for d-dimensional $\hat{X}, \hat{Y}$.

Table 4.2: Outputs are the corresponding powers: Type I for $P = Q$ and (1-Type II) for $P \neq Q$.

$d = 2$	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.1$	
n	d	d'	d	d'	d	d'	d	d'
100	0.04	0.06	0.01	0.06	0.04	0.97	0.12	1
200	0.03	0.07	0.05	0.08	0.04	1	0.55	1
500	0.02	0.03	0.09	0.36	1	1	1	1
1000	0.01	0.02	1	1	1	1	1	1
$d = 3$	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.1$	
n	d	d'	d	d'	d	d'	d	d'
100	0.03	0.04	0.04	0.08	0	0.98	0.04	1
200	0.04	0.02	0.01	0.13	0.03	1	0.09	1
500	0.08	0.04	0.05	0.57	0.94	1	1	1
1000	0.06	0.04	0.45	1	1	1	1	1

We also consider a condition where we observe three graphs and conduct a similar sample test. Besides the above X and Y, we add another independent latent position matrix $Z \in \mathbb{R}^{n \times d}$ in the same way, then we observe another graph $C \sim \text{Bernoulli}(R)$. Using the same method on the Omnibus matrix, we also obtain latent position estimates $\hat{X}, \hat{Y}$ and ones $\tilde{X}, \tilde{Y}$ in a higher dimensional space $\mathbb{R}^{d'}$ to test $\mathbb{H}_0 \colon P = Q$ against $\mathbb{H}_A \colon P \neq Q$ where the rejection region is calibrated via bootstrapping with significance level 0.05. We repeat the above steps for 100 Monte Carlo replicates to obtain an empirical estimate of power for T. The results are presented in Table 4.3 for combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $\epsilon \in \{0, 0.1, 0.3, 0.5\}$. Also, we observe that the type I error of the test statistic is well-controlled and that the test statistic also exhibits power even for small values of ϵ and moderate values of n. Besides, compared with result from d-dimensional embedding (Levin et al. 2017), our methods lead to larger power.

Algorithm 3 Bootstrap procedure for testing procedure

procedure STATISTIC_SAMPLE(P,s)

(i) Generate adjacency matrices $(A^{(s)}, B^{(s)}) \sim$ Bernoulli(P) with marginal edge probabilities matrix P.

(ii) Compute the Omnibus matrix $M_A^{(s)}$ from $A^{(s)}$ and $B^{(s)}$.

(iii) Compute top-d'' eigenvector $\hat{V}^{(s)}$ from $M_A^{(s)}$, set $\begin{pmatrix} \hat{V}_1^{(s)} \\ \hat{V}_2^{(s)} \end{pmatrix} = \hat{V}^{(s)}$ with equal $\hat{V}_1^{(s)}, \hat{V}_2^{(s)}$, and compute top-$d'$ eigenvector $\hat{U}^{(s)}$ from $(\hat{V}_1^{(s)} + \hat{V}_2^{(s)})/2$.

(iv1) Let $\hat{R}_X^{(4)(s)} = \hat{U}^{(s)\top} A^2 \hat{U}^{(s)}$, and compute the fourth root of $\hat{R}_X^{(4)(s)}$ as $\hat{R}_X^{(s)}$, and $\hat{R}_Y^{(s)}$ with the same method from $B^{(s)}$ and $\hat{U}^{(s)}$.

(iv2) Let $\hat{R}_X^{(s)}, \hat{R}_Y^{(s)}$ be the d-dimensional adjacency spectral embedding of $\hat{U}^{(s)\top} A \hat{U}^{(s)}$ and $\hat{U}^{(s)\top} B \hat{U}^{(s)}$.

(v) Compute $\hat{X}^{(s)} = \hat{U}^{(s)} \hat{R}_X^{(s)}$ and $\hat{Y}^{(s)} = \hat{U}^{(s)} \hat{R}_Y^{(s)}$.

(vi1) **return** $\min_{W \in \mathcal{O}(d)} \|\hat{X}^{(s)} - \hat{Y}^{(s)} W\|_F$.

(vi2) **return** $\|\hat{X}^{(s)} - \hat{Y}^{(s)}\|_F$.

end procedure

Require: Adjacency matrices A and B, both of size $n \times n$, significance level $\alpha \in (0, 1)$, number of bootstrap samples m.

(A) Compute the Omnibus matrix M_A from A and B.

(B) Compute top-d'' eigenvector $\hat{V}$ from M_A, set $\begin{pmatrix} \hat{V}_1 \\ \hat{V}_2 \end{pmatrix} = \hat{V}$ with equal $\hat{V}_1, \hat{V}_2$, and compute top-d eigenvector $\hat{U}$ from $(\hat{V}_1 + \hat{V}_2)/2$.

(C1) Let $\hat{R}_X^{(4)} = \hat{U}^\top A^2 \hat{U}$, and compute the fourth root of $\hat{R}_X^{(4)}$ as $\hat{R}_X$, and $\hat{R}_Y$ with the same method from B and $\hat{U}$.

(C2) Let $\hat{R}_X, \hat{R}_Y$ be the d'-dimensional adjacency spectral embedding of $\hat{U}^\top A \hat{U}$ and $\hat{U}^\top B \hat{U}$.

(D) Compute $\hat{X} = \hat{U} \hat{R}_X$ and $\hat{Y} = \hat{U} \hat{R}_Y$.

(E1) Calculate the test statistics $T = \min_{W \in \mathcal{O}(d)} \|\hat{X} - \hat{Y} W\|_F$.

(E2) Calculate the test statistics $T = \|\hat{X} - \hat{Y}\|_F$.

for $s = 1$ to m **do**

(a) $T_P^{(s)} \leftarrow$ Statistic_Sample(P,s).

(b) $T_Q^{(s)} \leftarrow$ Statistic_Sample(Q,s).

end for

(F) Set $c_\alpha^{(P)}$ to be the $(1-\alpha) \times 100\%$ percentile of the $\{T_P^{(s)}\}_{s=1}^m$.

(F) Set $c_\alpha^{(Q)}$ to be the $(1-\alpha) \times 100\%$ percentile of the $\{T_Q^{(s)}\}_{s=1}^m$.

Output If $T > \max\{c_\alpha^{(P)}, c_\alpha^{(Q)}\}$ then reject $\mathbb{H}_0$; otherwise fail to reject $\mathbb{H}_0$.

Table 4.3: Outputs are the corresponding powers: Type I for $P = Q$ and (1-Type II) for $P \neq Q$.

$d = 2$	$\epsilon = 0$			$\epsilon = 0.01$			$\epsilon = 0.05$			$\epsilon = 0.1$		
n	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin
100	.06	.04	.04	.06	.04	.03	.89	.10	.06	1	.10	.34
200	.03	.03	.04	.10	.06	.04	1	.18	.07	1	.57	.81
500	.05	.06	.06	.36	.29	.05	1	1	.62	1	1	1
1000	.02	.02	.02	.99	1	.06	1	1	1	1	1	1
$d = 3$	$\epsilon = 0$			$\epsilon = 0.01$			$\epsilon = 0.05$			$\epsilon = 0.1$		
n	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin	$\tilde{X}$	$\hat{X}$	Levin
100	.03	.02	.02	.06	.03	.02	.94	0	.10	1	.01	.46
200	.03	.04	.03	.20	.05	.07	1	.04	.13	1	.05	.71
500	.04	.02	.08	.43	.06	.04	1	.98	.21	1	1	1
1000	.04	.03	.03	1	.79	.02	1	1	.96	1	1	1

4.5.3 Community Detection

In this subsection, we conduct clustering on two stochastic blockmodel graphs A, B with the same assignment for vertices. We assume that A is a K-block SBM where each vertex of A is assigned to some a block $k \in \{1, \ldots, K\}$ with probability $1/K$ and the marginal edge probabilities between vertices in block k and vertices in block ℓ are given by $0.45 - |k-\ell|/(2K)$ for any $k, \ell \in \{1, 2, \ldots, K\}$. Similarly, B is a K-block SBM with the same membership assignment as A and marginal edge probabilities $0.4 - |k - \ell|/(2K + 2)$. Recall that U is combination of two groups of latent positions, so it is also about combination of two assignments, which are the same in our setting. Hence, we can conduct clustering on $\hat{U}$ containing information from two graphs. For comparisions we also include the common method, that is, clustering on eigenvectors of $A+B$. We repeat 100 times to estimate the clusering rate. The results are presented in Table 4.4 for various combinations of $n \in \{100, 200, 500, 1000\}$ and $K \in \{2, 3, 4, 5, 6\}$. We observe that $\hat{U}$ is eligible for clustering, and it behaves better than the common way.

Table 4.4: Outputs are accuracy rates for clusters.

$n =$	100		200		500		1000	
	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$
k=2	1	1	1	1	1	1	1	1
k=3	0.66	0.70	0.74	0.90	1	1	1	1
k=4	0.50	0.52	0.51	0.60	0.74	0.86	0.99	0.99
k=5	0.42	0.44	0.43	0.47	0.57	0.60	0.76	0.80
k=6	0.38	0.38	0.38	0.41	0.47	0.49	0.57	0.59

For some edge probability setting, $A+B$ could blur the block structures. For example, the marginal edge probabilities of A is given by $0.45-|k-\ell|/(2K)$ and the one of B is given by $|k-\ell|/(2K)$ as the above. The results are presented in Table 4.5 for various combinations of $n \in \{100, 200, 500, 1000, 2000\}$ and $K \in \{2, 3, 4, 5, 6\}$. We observe that $\hat{U}$ behaves much better than the common way, for which the accuracy rates even do not converges to 1 as n increases. Furthermore, in this extreme condition, the common way almost leads to a complete random result. We can notice that there exists a probability matrix that is not positive semi-definite. However, the results for GRDPG support us to use $\hat{U}$ to conduct the community detection.

Table 4.5: The elements are accuracy rates for clusters

$n =$	100		200		500		1000		2000	
	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$	$A+B$	$\hat{U}$
k=2	0.55	0.99	0.53	1	0.52	1	0.51	1	0.51	1
k=3	0.40	0.62	0.38	0.69	0.36	0.97	0.35	1	0.35	1
k=4	0.33	0.46	0.31	0.53	0.29	0.71	0.27	0.76	0.27	0.80
k=5	0.30	0.39	0.27	0.43	0.24	0.57	0.23	0.62	0.22	0.63
k=6	0.28	0.35	0.24	0.37	0.21	0.48	0.20	0.51	0.19	0.52

4.6 Real Data Analysis

4.6.1 HNU1 Data

In this section, we consider connectomes constructed from the HNU1 study (Zuo et al. 2014). They are for diffusion magnetic resonance imaging (dMRI) records from 10 scans of each of 30 healthy adult subjects over one month, so we have $m = 300$ graphs. The vertex number $n = 200$ by registering the brain regions (Craddock et al. 2012). We binarize the graphs to replace all non-zero edge weights with 1, and leave unchanged edges of weight zero. The $mn \times mn$ Omnibus matrix has 400 non-zero singular values, and we select a dimension 11 for the Omnibus embedding by the profile-likelihood method of Zhu and Ghodsi (2006). Then we have a $(mn \times 11)$ Omnibus embedding $\hat{V}$, where we know

$$\hat{V} = \begin{pmatrix} \hat{V}^{(1)} \\ \hat{V}^{(2)} \\ \vdots \\ \hat{V}^{(m)} \end{pmatrix}$$

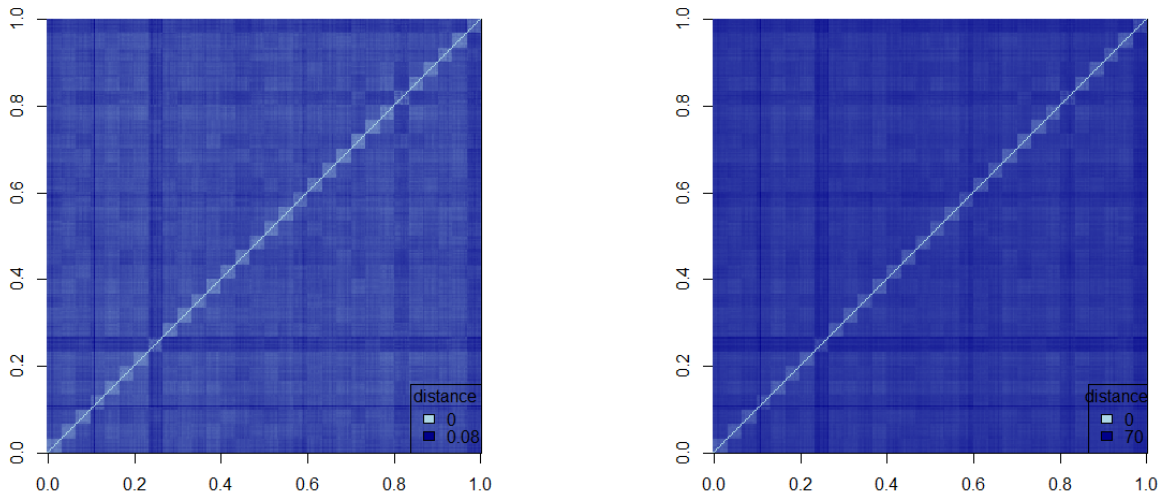
and $\hat{V}_i$ with the dimension $n \times 11$ reflects the latent positions of i-th graph. Every $11n$-dimensional vector $\text{vec}(\hat{V}_i)$ can express a position for each graph.

We consider the common L_2 distance and use the following naive method to classify scans into different subjects. This distance is the same as the Frobenius norm $\|\hat{V}^{(i)} - \hat{V}^{(f)}\|_F$. We select one scan from each subject as known labeled data, then for every remaining scan and its corresponding $11n$-dimensional vector $\text{vec}(\hat{V}_i)$, we classify it based on the label of the closest $\text{vec}(\hat{V}_j)$ over the labeled 30 scans. We repeat 100 times to select labeled data, and the average of the accuracy rate of classification is 98.2%. If we expand known labeled data, the rate can increase. Therefore, the Omnibus embedding can express the latent positions for graphs.

Also, we conduct clustering on all scans based on the Omnibus embedding, and consider all the $11n$-dimensional vectors, assuming there are 30 clusters. We remark that out of 300 scans from 30 subjects, only 1 subject scan is divided across clusters, and only two subjects are mistakenly clustered into the same cluster. If we directly conduct clustering on the original scans, considering the vectorization of the upper triangle, there are 5 subjects scans divided across clusters, and two pairs of subjects are mistakenly clustered into the same clusters, respectively. Therefore, Omnibus embedding could even better reflect relations among graphs.

Compared to the distance among graphs, the distance among latent positions of graphs

has similar block structures as in Fig 4.1. We observed that Omnibus embedding can also reflect the structure among graphs.



(a) Omnibus (b) whole scan

Figure 4.1: Comparison of Distances: Embedding vs. Graphs

4.6.2 COBRE Data

In this section, consider the COBRE data set (Aine et al. 2017), a collection of functional MRI scans of both schizophrenic and healthy patients. Each scan yields a graph on 264 vertices, corresponding to 264 brain regions of interest (Power et al. 2011), with edge weights given by correlations between BOLD (Blood Oxygenation Level Dependent) signals measured in those regions. The data set contains scans for 54 schizophrenic patients and 70 healthy controls, for a total of 124 brain graphs.

We keep 263 vertices that are involved in all graphs. Hence, we have $n = 263$ vertices and $m = 124$ brain graphs. We use Fisher transformation on correlation edges. There are five edges out of $(-1, 1)$, and they are a bit larger than 1 and in $(1.01, 1.04)$. Hence, we treat them as 1 for a strong correlation, and assign a large value for their Fisher transformation. We can conduct the normal error assumption based on the results in Section 4.4. We select the dimension $\hat{d}'' = 10$ of Omnibus embedding for the $(mn) \times (mn)$ Omnibus matrix by the

profile-likelihood method of Zhu and Ghodsi (2006). For the Omnibus embedding $\hat{V}$ as the above, we also consider the Frobenius norm $\|\hat{V}^{(i)} - \hat{V}^{(j)}\|_F$ as distance to conduct clustering. We have a clustering accuracy rate of 69%, compared with an accuracy rate of 63% for using $\hat{d} = 7$ for the Omnibus dimension from the average of the estimated dimension for each graph. Furthermore, if we directly conduct clustering on the original scans, considering the vectorization of the upper triangle, the clustering accuracy rate is only 53%.

We also use Hotelling's t^2 test to estimate whether each brain region has a significant difference between schizophrenic and healthy patients. Consider a multivariate analysis of variance (MANOVA), and MANOVA p-values are shown as in Fig 4.2. The significant brain region with the smallest p-value is the Fronto-parietal Task Control region, which previously linked to schizophrenia (Relión et al. 2019; Bunney and Bunney 2000; Fornito et al. 2012). For the other three with extremely small p-values, one of them is also the Fronto-parietal Task Control region and it is with the second smallest p-value. The other two regions are the Subcortical and Cerebellar regions. Patients with schizophrenia have decreased blood flow in the cerebellum, and emotion Schizophrenia is associated with subcortical abnormalities (Andreasen and Pierson 2008). Also, many subcortical areas also support cognition (Fan et al. 2019).

4.7 Discussion

In this chapter, we proposed a novel method for statistical inference on multiple random graphs using the Omnibus embedding technique. Our approach allows for the simultaneous analysis of multiple graphs, accommodating cases where the graphs may not share the same distribution, which is a significant advancement over existing methods that assume similar or identical distributions across graphs. We extended the traditional Random Dot Product Graph (RDPG) model to a more generalized framework, enabling more flexible and comprehensive analysis.

Our theoretical results established the convergence and asymptotic normality of the Omnibus embedding, demonstrating that the estimated latent positions closely approximate the true latent positions under both Bernoulli and Gaussian error settings. These guarantees ensure that our method is robust and reliable for various applications. We validated our method through extensive simulations, demonstrating its effectiveness in recovering latent positions, performing hypothesis testing, and conducting community detection. Our method showed superior performance compared to existing techniques, particularly in scenarios where graphs do not share the same distribution. We applied our method to real-world datasets, including the HNU1 connectome data and the COBRE dataset. Our method effectively

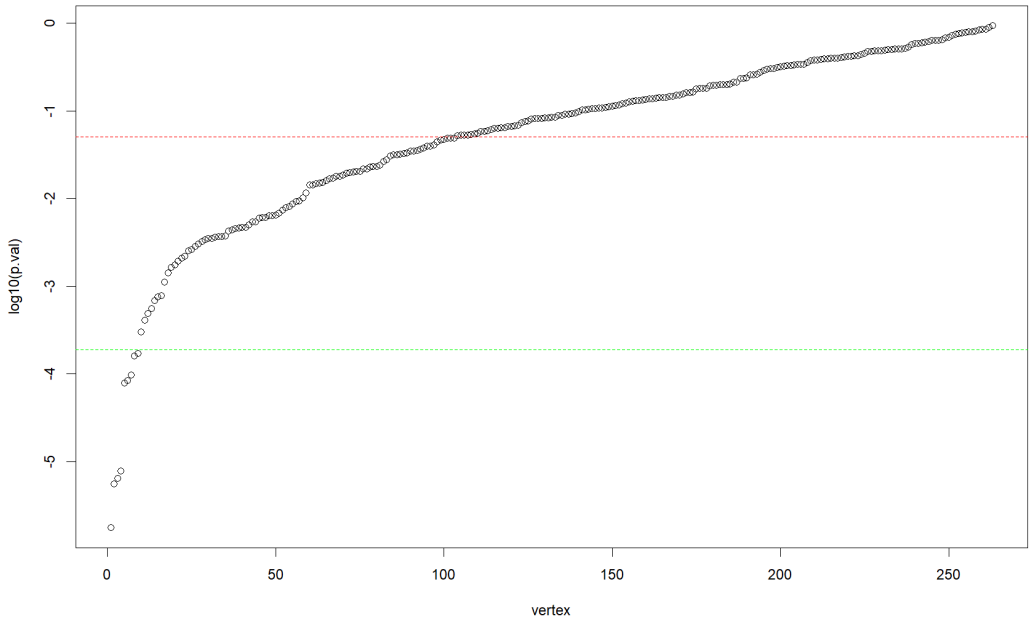


Figure 4.2: MANOVA p-values, with vertices sorted by significance and adjusted for multiple comparisons. The dotted lines indicate the p=0.05 threshold (red) and the threshold (green) after Bonferroni correction.

classified and clustered the graphs, showing high accuracy rates and identifying significant brain regions associated with schizophrenia. These applications highlight the practical utility of our approach in neuroscience and medical research.

While our method offers significant advancements, there are some directions for future research. One potential extension is to allow the number of graphs m, to grow unbounded, possibly exceeding the number of vertices. Additionally, as both the number of graphs and the size of each graph increase, ensuring computational efficiency becomes crucial. Developing more efficient algorithms and leveraging parallel computing techniques could further enhance the applicability of our method to large-scale datasets.

CHAPTER

5

CONCLUSIONS

In this dissertation, we aimed to investigate the applications and improvements of graph models in statistical inference. Our primary contributions are threefold. First, we enhanced a two-sample hypothesis testing method for Random Dot Product Graphs (RDPG), demonstrating its superior performance through theoretical analysis and extensive simulations. Second, we developed an independence testing procedure for inhomogeneous Erdos-Rényi random graphs, which effectively identifies dependencies between graph structures. Third, we introduced a novel joint inference method for multiple random graphs, which facilitates simultaneous analysis of more than two graphs and provides deeper insights into their underlying structures.

These findings have significant implications for various fields such as neuroscience, where understanding the structural differences between brain networks is crucial, and in machine learning, where improving the accuracy of network-based algorithms can enhance predictive modeling. Our enhanced two-sample hypothesis testing method, for instance, can be applied to detect changes in brain connectivity patterns associated with neurological diseases. Similarly, the independence testing procedure can be utilized to explore functional connectivity in brain networks, and the joint inference method can be used to analyze complex systems with multiple interacting networks.

Despite the promising results, our study has several limitations. One major limitation is the assumption of known vertex correspondence in the two-sample hypothesis testing

method, which may not always be practical in real-world applications. Additionally, our methods primarily focus on undirected and unweighted graphs, whereas many real-world networks are directed and weighted. Future research should address these limitations by developing techniques for hypothesis testing and independence testing without requiring known vertex correspondence and extending the methods to directed and weighted graphs. Another direction is the integration of our methods with advanced techniques to further enhance their predictive power and applicability to complex real-world problems.

REFERENCES

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

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452.

Aine, C., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., Hanlon, F. M., Houck, J. M., Jung, R. E., Lauriello, J., et al. (2017). Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics*, 15:343–364.

Alon, N., Krivelevich, M., and Sudakov, B. (1998). Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466.

Anderson, T. W. (2003). *An Introducion to Multivariate Statistical Analysis*. Wiley, 3rd edition.

Andreasen, N. C. and Pierson, R. (2008). The role of the cerebellum in schizophrenia. *Biological psychiatry*, 64(2):81–88.

Athreya, A., Cape, J., and Tang, M. (2022). Eigenvalues of stochastic blockmodel graphs and random graphs with low-rank edge probability matrices. *Sankhya A*, 84:36–63.

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., and Sussman, D. L. (2018). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92.

Bae, S.-H., Halperin, D., West, J. D., Rosvall, M., and Howe, B. (2017). Scalable and efficient flow-based community detection for large-scale graph analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3):1–30.

Banks, J., Moore, C., Neeman, J., and Netrapalli, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416.

Barak, B., Hopkins, S., Kelner, J., Kothari, P. K., Moitra, A., and Potechin, A. (2019). A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48:687–735.

Bedi, P. and Sharma, C. (2016). Community detection in social networks. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 6(3):115–135.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Bhattacharyya, S. and Chatterjee, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122.

Braverman, M., Ko, Y. K., Rubinstein, A., and Weinstein, O. (2017). Eth hardness for densest-k-subgraph with perfect completeness. In *Proceedings of the 2017 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1326–1341.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198.

Bunney, W. E. and Bunney, B. G. (2000). Evidence for a compromised dorsolateral prefrontal cortical parallel circuit in schizophrenia. *Brain Research Reviews*, 31(2-3):138–146.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

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

Cape, J., Tang, M., and Priebe, C. E. (2019b). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics.

Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Pozo, F. d., and Boccaletti, S. (2013). Emergence of network features from multiplexity. *Scientific reports*, 3(1):1344.

Carlsson, M. (2018). Perturbation theory for the matrix square root and matrix modulus. *arXiv preprint arXiv:1810.01464*.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214.

Chen, G., Arroyo, J., Athreya, A., Cape, J., Vogelstein, J. T., Park, Y., White, C., Larson, J., Yang, W., and Priebe, C. E. (2020). Multiple network embedding for anomaly detection in time series of graphs. *arXiv preprint arXiv:2008.10055*.

Chen, L., Vogelstein, J. T., Lyzinski, V., and Priebe, C. E. (2016). A joint graph inference case study: the *c. elegans* chemical and electrical connectomes. *Worm*, 5(2):e1142041.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

Craddock, R. C., James, G. A., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928.

Draves, B. and Sussman, D. L. (2020). Bias-variance tradeoffs in joint spectral embeddings. *arXiv preprint arXiv:2005.02511*.

Eichler, K. et al. (2017). The complete connectome of a learning and memory centre in an insect brain. *Nature*, 548(7666):175–182.

Fan, F., Xiang, H., Tan, S., Yang, F., Fan, H., Guo, H., Kochunov, P., Wang, Z., Hong, L. E., and Tan, Y. (2019). Subcortical structures and cognitive dysfunction in first episode schizophrenia. *Psychiatry Research: Neuroimaging*, 286:69–75.

Fan, J., Wang, W., and Zhong, Y. (2018). An [... formula...] eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research: JMLR*, 18.

Fishkind, D. E., Meng, L., Sun, A., Priebe, C. E., and Lyzinski, V. (2019). Alignment strength and correlation for graphs. *Pattern Recognition Letters*, 125:295–302.

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *Neuroimage*, 62(4):2296–2314.

Füredi, Z. and Komlós, J. (1981). The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241.

Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate-optimal graphon estimation. *Annals of Statistics*, 43(6):2624–2652.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(1):1980–2024.

Ghoshdastidar, D., Gutzeit, M., Carpentier, A., and Von Luxburg, U. (2020). Two-sample hypothesis testing for inhomogeneous random graphs. *Annals of Statistics*, 48(4):2208–2229.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592.

Guenther, W. C. (1964). Another derivation of the non-central chi-square distribution. *Journal of the American Statistical Association*, 59(307):957–960.

Haghighi, A. D., Ng., A., and Manning, C. D. (2005). Robust textual inference via graph matching. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 387–394.

Hasan, M. A. and Zaki, M. J. (2011). A survey of link prediction in social networks. *Social network data analytics*, pages 243–275.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

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

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Huang, S., Weng, H., and Feng, Y. (2023). Spectral clustering via adaptive layer aggregation for multi-layer networks. *Journal of Computational and Graphical Statistics*, 32(3):1170–1184.

Jones, A. and Rubin-Delanchy, P. (2020). The multilayer random dot product graph. *arXiv preprint arXiv:2007.10455*.

Kendall, M. (1990). *Rank correlation methods*. Oxford Univerity Press, 5th edition.

Kim, C. H., Jo, M., Lee, J., Bianconi, G., and Kahng, B. (2021). Link overlap influences opinion dynamics on multiplex networks of ashkin-teller spins. *Physical Review E*, 104(6):064304.

Kong, Z., Sun, L., Peng, H., Zhan, L., Chen, Y., and He, L. (2021). Multiplex graph networks for multimodal brain network analysis. *arXiv preprint arXiv:2108.00158*.

Korula, N. and Lattanzi, S. (2014). An efficient reconciliation algorithm for social networks. *Proceedings of the Very Large Data Bases Endowment*, 7:377–388.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Annals of Statistics*, 44:401–424.

Lei, J. and Lin, K. Z. (2023). Bias-adjusted spectral clustering in multi-layer stochastic block models. *Journal of the American Statistical Association*, 118(544):2433–2445.

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

Lesieur, T., Krzakala, F., and Deborová, L. (2015). Phase transitions in sparse PCA. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1635–1639.

Leung, D. and Drton, M. (2018). Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics*, 46:280–307.

Levin, K., Athreya, A., Tang, M., Lyzinski, V., Park, Y., and Priebe, C. E. (2017). A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv preprint arXiv:1705.09355*.

Li, T., Levina, E., and Zhu, J. (2020). Network cross-validation by edge resampling. *Biometrika*, 107:257–276.

Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Society.

Lu, L. and Peng, X. (2012). Spectra of edge-independent random graphs. *arXiv preprint arXiv:1204.6207*.

Lyzinski, V., Fishkind, D. E., Fiori, M., Vogelstein, J. T., Priebe, C. E., and Sapiro, G. (2016). Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:60–73.

Lyzinski, V., Fishkind, D. E., and Priebe, C. E. (2014a). Seeded graph matching for correlated erdös-rényi graphs. *Journal of Machine Learning Research*, 15:3513–3540.

Lyzinski, V. and Sussman, D. L. (2020). Matchability of heterogeneous networks pairs. *Information and Inference: A Journal of the IMA*, 9:749–783.

Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A., and Priebe, C. E. (2014b). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8:2905–2922.

Ma, Z., Marchette, D. J., and Priebe, C. E. (2012). Fusion and inference from multiple data sources in a commensurate space. *Statistical Analysis and Data Mining*, 5:187–193.

Madhawa, K. and Murata, T. (2020). Active learning for node classification: An evaluation. *Entropy*, 22(10):1164.

Miller, K. S. (1981). On the inverse of the sum of matrices. *Mathematics magazine*, 54(2):67–72.

Onaran, E., Garg, S., and Erkip, E. (2016). Optimal de-anonymization in random graphs with community structure. In *Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers*, pages 709–713.

Pantazis, K., Athreya, A., Arroyo, J., Frost, W. N., Hill, E. S., and Lyzinski, V. (2022). The importance of being correlated: Implications of dependence in joint spectral inference across multiple networks. *Journal of Machine Learning Research*, 23(141):1–77.

Pedarsani, P. and Grossglauser, M. (2011). On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1243.

Perry, A., Wein, A. S., and Bandeira, A. S. (2020). Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré, Probabilit'es et Statistiques*, 56:230–264.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Racz, M. and Sridhar, A. (2021). Correlated stochastic block models: Exact graph matching with applications to recovering communities. *Advances in Neural Information Processing Systems*, 34:22259–22273.

Relión, J. D. A., Kessler, D., Levina, E., and Taylor, S. F. (2019). Network classification with applications to brain connectomics. *The annals of applied statistics*, 13(3):1648.

Richiardi, J., Achard, S., Bunke, H., and Van De Ville, D. (2013). Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal processing magazine*, 30(3):58–70.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., and Van De Ville, D. (2011). Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626.

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2022). A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society, Series B.*, 84:1446–1473.

Sussman, D. L., Lyzinski, V., Park, Y., and Priebe, C. E. (2020). Matched filters for noisy induced subgraph detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2887–2900.

Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128.

Sussman, D. L., Tang, M., and Priebe, C. E. (2013). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., Park, Y., and Priebe, C. E. (2017a). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E. (2014). A nonparametric two-sample hypothesis testing problem for random dot product graphs. *arXiv preprint arXiv:1409.2344*.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E. (2017b). A nonparametric two-sample hypothesis testing problem for random graphs.

Tang, M., Sussman, D. L., and Priebe, C. E. (2013). Universally consistent vertex classification for latent positions graphs.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., and Chklovskii, D. B. (2011). Structural properties of the caenorhabditis elegans neuronal network. *PLoS computational biology*, 7:e1001066.

Vu, T., Chunikhina, E., and Raich, R. (2021). Perturbation expansions and error bounds for the truncated singular value decomposition. *Linear Algebra and its Applications*, 627:94–139.

Wang, Y. X. and Bickel, P. (2017). Likelihood based model selection for stochastic blockmodels. *Annals of Statistics*, 45:500–528.

Wihler, T. P. (2009). On the hölder continuity of matrix functions for normal matrices. *Journal of inequalities in pure and applied mathematics*, 10(10).

Winding, M. et al. (2023). The connectome of an insect brain. *Science*, 379.

Wu, Y. and Xu, J. (2021). Statistical problems with planted structures: Information-theoretical and computational limits. In *Information theoretic methods in data science*, pages 383–424. Cambridge University Press.

Wu, Y., Xu, J., and Yu, S. H. (2023). Testing correlation of unlabeled random graphs. *Annals of Applied Probability*.

Xiong, J., Shen, C., Arroyo, J., and Vogelstein, J. T. (2019). Graph independence testing. arXiv preprint #1906.03661.

Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5433–5442.

Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighborhood smoothing. *Biometrika*, 104:771–783.

Zhang, Y. and Tang, M. (2022). Perturbation analysis of randomized svd and its applications to high-dimensional statistics. *arXiv preprint arXiv:2203.10262*.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51:918–930.

Zhu, X., Lafferty, J., and Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3, pages 58–65.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1(1):1–13.

APPENDICES

APPENDIX

A

SUPPLEMENT TO CHAPTER 2

A.1 Proof of Theorem 1

Let
$W^* := \arg\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2$,
$W_x := \arg\min_{W \in \mathcal{O}(d)} \|XW - \hat{X}\|_F^2$,
$W_y := \arg\min_{W \in \mathcal{O}(d)} \|YW - \hat{Y}\|_F^2$.

Upper Bound:
Let $W_0 = \arg\min_{W \in \mathcal{O}(d)} \|XW_xW - YW_y\|_F^2$

$$
\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \\
\leq &\|\hat{X}W_0 - \hat{Y}\|_F^2 \\
= &\|\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y}\|_F^2 + \|XW_xW_0 - YW_y\|_F^2 \\
&+ \operatorname{tr}(\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y})^T(XW_xW_0 - YW_y) \\
= &\|\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y}\|_F^2 + \min_{W \in \mathcal{O}(d)} \|XW - Y\|_F^2 \\
&+ \operatorname{tr}(\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y})^T(XW_xW_0 - YW_y)
\end{aligned}
$$

And we have with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\mathrm{tr}(\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y})^T(XW_xW_0 - YW_y) \\
\leq&\sqrt{d}\|(\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y})^T(XW_xW_0 - YW_y)\|_F \\
\leq&\sqrt{d}\|(\hat{X}W_0 - XW_xW_0)^T(XW_xW_0 - YW_y)\|_F + \sqrt{d}\|(YW_y - \hat{Y})^T(XW_xW_0 - YW_y)\|_F \\
\sim&\sqrt{d}\|S_p^{-1/2}U_p^T(A-P)(XW_xW_0 - YW_y)\|_F + \sqrt{d}\|S_Q^{-1/2}U_Q^T(B-Q)(XW_xW_0 - YW_y)\|_F,
\end{aligned}$$

where we apply Lemma A.3 and Lemma A.4.

We consider only one item because of the similarity,
based on Lemma 7,

$$\begin{aligned}
&\|S_p^{-1/2}U_p^T(A-P)(XW_xW_0 - YW_y)\|_F \\
\leq&\|S_p^{-1/2}\|_2\|U_p^T(A-P)V\|_F\|M\|_2
\end{aligned}$$

where $\|M\|_2 = \|XW_xW_0 - YW_y\|_2 \leq \min_{W\in\mathcal{O}(d)}\|XW - Y\|_F = \mathbb{O}(1)$. We know $\|S_p^{-1/2}\|_2 = \mathbb{O}(\frac{1}{\log^{1+\epsilon/2} n})$, $\|U_p^T(A-P)V\|_F = \mathbb{O}(\log n)$ with the probability at least $1 - n^C$ for any $C > 0$ (based on Lemma 8), we have $\|S_p^{-1/2}U_p^T(A-P)(XW_xW_0 - YW_y)\|_F = \mathbb{O}(\log^{-\epsilon/2} n) \to 0$. So with the probability at least $1 - n^C$ for any $C > 0$

$$\mathrm{tr}(\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y})^T(XW_xW_0 - YW_y) \to 0$$

On the other hand, we have with the probability at least $1 - n^C$ for any $C > 0$

$$\begin{aligned}
&\|\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y}\|_F^2 \\
=&\|\hat{X}W_0 - XW_xW_0\|_F^2 + \|YW_y - \hat{Y}\|_F^2 + 2\mathrm{tr}(\hat{X}W_0 - XW_xW_0)^T(YW_y - \hat{Y}) \\
\sim&\|\hat{X} - XW_x\|_F^2 + \|YW_y - \hat{Y}\|_F^2 \qquad \text{(based on Lemma 9)} \\
\sim&C(X) + C(Y) \qquad \text{(based on Lemma 10)} \\
\sim&C(\hat{X}) + C(\hat{Y}) \qquad \text{(based on Lemma 6)}
\end{aligned}$$

(i)Under H_0, with the probability at least $1 - n^C$ for any $C > 0$,

$$\limsup \min_{W\in\mathcal{O}(d)} \frac{\|\hat{X}W - \hat{Y}\|_F^2}{2C(\hat{X})} \leq 1$$

(i)Under H_1, with the probability at least $1 - n^C$ for any $C > 0$,

$$\limsup \min_{W \in \mathcal{O}(d)} \frac{\|\hat{X}W - \hat{Y}\|_F^2}{C(\hat{X}) + C(\hat{Y}) + \min_{W \in \mathcal{O}(d)} \|XW - Y\|_F^2} \leq 1$$

Lower Bound:
Suppose $\|XW_xW^* - YW_y\|_F^2 = \omega(1)$, then

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \\
=&\|\hat{X}W^* - \hat{Y}\|_F^2 \\
\geq&\left(\|XW_xW^* - YW_y\|_F - \|\hat{X}W^* - XW_xW^* + YW_y - \hat{Y}\|_F\right)^2 \\
=&(\omega(1) - \theta(1))^2 \\
=&\omega(1)
\end{aligned}$$

which is contradictory to what we got about the upper bound.
So $\|XW_xW^* - YW_y\|_F^2 = \mathbb{O}(1)$.

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \\
=&\|\hat{X}W^* - \hat{Y}\|_F^2 \\
=&\|\hat{X}W^* - XW_xW^* + YW_y - \hat{Y}\|_F^2 + \|XW_xW^* - YW_y\|_F^2 \\
&+ \operatorname{tr}(\hat{X}W^* - XW_xW^* + YW_y - \hat{Y})^T(XW_xW^* - YW_y) \\
\geq&\|\hat{X}W^* - XW_xW^* + YW_y - \hat{Y}\|_F^2 + \min_{W \in \mathcal{O}(d)} \|XW - Y\|_F^2 \\
&+ \operatorname{tr}(\hat{X}W^* - XW_xW^* + YW_y - \hat{Y})^T(XW_xW^* - YW_y)
\end{aligned}$$

Based on the same way in the upper bound part, we have with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\operatorname{tr}(\hat{X}W^* - XW_xW^* + YW_y - \hat{Y})^T(XW_xW^* - YW_y) \to 0 \\
&\|\hat{X}W_0 - XW_xW_0 + YW_y - \hat{Y}\|_F^2 \sim C(X) + C(Y) \sim C(\hat{X}) + C(\hat{Y})
\end{aligned}$$

(i)Under H_0, with the probability at least $1 - n^C$ for any $C > 0$,

$$\liminf \min_{W \in \mathcal{O}(d)} \frac{\|\hat{X}W - \hat{Y}\|_F^2}{2C(\hat{X})} \geq 1$$

(i)Under H_1, with the probability at least $1 - n^C$ for any $C > 0$,

$$\liminf \min_{W \in \mathcal{O}(d)} \frac{\|\hat{X}W - \hat{Y}\|_F^2}{C(\hat{X}) + C(\hat{Y}) + \min_{W \in \mathcal{O}(d)} \|XW - Y\|_F^2} \geq 1$$

Conclusion:
with the probability at least $1 - n^C$ for any $C > 0$,
under H_0,
$\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \sim 2C(\hat{X})$;
under H_1:
$\min_{W \in \mathcal{O}(d)} \|\hat{X}W - \hat{Y}\|_F^2 \sim C(\hat{X}) + C(\hat{Y}) + \min_{W \in \mathcal{O}(d)} \|XW - Y\|_F^2$

A.2 Proof of Theorem 2

Let
$W^* := \arg\min_{W \in \mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2$,
$W_x := \arg\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW_x\|_F^2$,
$W_y := \arg\min_{W \in \mathcal{O}(d)} \|\hat{Y} - YW_y\|_F^2$.

Upper Bound:
Let $W_0 = \arg\min_{W \in \mathcal{O}(d)} \|\frac{XW_xW}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F^2$

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \\
\leq &\|\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \\
= &\|\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 + \|\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F^2 \\
&+ \mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) \\
= &\|\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 + \min_{W \in \mathcal{O}(d)} \|\frac{XW}{\|X\|_F} - \frac{Y}{\|Y\|_F}\|_F^2 \\
&+ \mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F})
\end{aligned}$$

Without loss of generality, we consider

$$
\begin{aligned}
&\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) \\
=&\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{\hat{X}W_0}{\|X\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) + \mathrm{tr}(\frac{\hat{X}W_0}{\|X\|_F} - \frac{XW_xW_0}{\|X\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) \\
\leq&\sqrt{d}\|(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{\hat{X}W_0}{\|X\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F})\|_F \\
&+ \sqrt{d}\|(\frac{\hat{X}W_0}{\|X\|_F} - \frac{XW_xW_0}{\|X\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F})\|_F \\
\leq&\sqrt{d}(1 - \frac{\|\hat{X}\|_F}{\|X\|_F})\|\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F + \sqrt{d}\|(\hat{X} - XW_x)^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F})\|_F \\
=&o((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2) \qquad \text{(The same way in Section A.1)},
\end{aligned}
$$

with the probability at least $1 - n^C$ for any $C > 0$.

So with the probability at least $1 - n^C$ for any $C > 0$,

$$
\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW_0}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) = o((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2)
$$

On the other hand, we have with the probability at least $1 - n^C$ for any $C > 0$,

$$
\begin{aligned}
&\|\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 \\
=&\|\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F}\|_F^2 + \|\frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 + \mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F} - \frac{XW_xW_0}{\|X\|_F})^T(\frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{XW_xW_0}{\|X\|_F})^T(\frac{YW_y}{\|\hat{Y}\|_F}-\frac{\hat{Y}}{\|Y\|_F})\\
=&\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{\hat{X}W_0}{\|X\|_F})^T(\frac{YW_y}{\|\hat{Y}\|_F}-\frac{YW_y}{\|Y\|_F})+\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{\hat{X}W_0}{\|X\|_F})^T(\frac{YW_y}{\|Y\|_F}-\frac{\hat{Y}}{\|Y\|_F})\\
&+\mathrm{tr}(\frac{\hat{X}W_0}{\|X\|_F}-\frac{XW_xW_0}{\|X\|_F})^T(\frac{YW_y}{\|\hat{Y}\|_F}-\frac{YW_y}{\|Y\|_F})+\mathrm{tr}(\frac{\hat{X}W_0}{\|X\|_F}-\frac{XW_xW_0}{\|X\|_F})^T(\frac{YW_y}{\|Y\|_F}-\frac{\hat{Y}}{\|Y\|_F})\\
\le&\sqrt{d}(1-\frac{\|\hat{X}\|_F}{\|X\|_F})(1-\frac{\|\hat{Y}\|_F}{\|Y\|_F})+\sqrt{d}(1-\frac{\|\hat{X}\|_F}{\|X\|_F})\frac{\|\hat{Y}-YW_y\|_F}{\|Y\|_F}\\
&+\sqrt{d}\frac{\|\hat{X}-XW_x\|_F}{\|X\|_F}(1-\frac{\|\hat{Y}\|_F}{\|Y\|_F})+\frac{1}{\|X\|_F\|Y\|_F}\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y)\\
=&o((\frac{1}{\|X\|_F}+\frac{1}{\|Y\|_F})^2)\qquad\text{(based on lemma 9).}
\end{aligned}
$$

So

$$
\begin{aligned}
&\|\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{XW_xW_0}{\|X\|_F}+\frac{YW_y}{\|\hat{Y}\|_F}-\frac{\hat{Y}}{\|Y\|_F}\|_F^2\\
=&\|\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{XW_xW_0}{\|X\|_F}\|_F^2+\|\frac{YW_y}{\|\hat{Y}\|_F}-\frac{\hat{Y}}{\|Y\|_F}\|_F^2+\mathrm{tr}(\frac{\hat{X}W_0}{\|\hat{X}\|_F}-\frac{XW_xW_0}{\|X\|_F})^T(\frac{YW_y}{\|\hat{Y}\|_F}-\frac{\hat{Y}}{\|Y\|_F})\\
\sim&\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW_x}{\|X\|_F}\|_F^2+\|\frac{YW_y}{\|\hat{Y}\|_F}-\frac{\hat{Y}}{\|Y\|_F}\|_F^2\\
\sim&\frac{C(X)}{\|X\|_F^2}+\frac{C(Y)}{\|Y\|_F^2}\qquad\text{(based on lemma 11)}\\
\sim&\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}.
\end{aligned}
$$

(i)Under H_0, there exists $\{a_n\}$ such that for sufficient large n, with the probability at least $1-n^C$ for any $C>0$,

$$\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2\le a_n\sim\frac{2C(\hat{X})}{\|\hat{X}\|_F^2};$$

(i)Under H_1, there exists $\{a_n\}$ such that for sufficient large n, with the probability at least $1-n^C$ for any $C>0$,

$$\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}W}{\|\hat{X}\|_F}-\frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2\le a_n\sim\frac{C(\hat{X})}{\|\hat{X}\|_F^2}+\frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}+\min_{W\in\mathcal{O}(d)}\|\frac{XW}{\|X\|_F}-\frac{Y}{\|Y\|_F}\|_F^2$$

Lower Bound:

Suppose $\|\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F^2 = \omega((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2)$, then

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \\
\geq&\big[\|\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F - \|\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F\big]^2 \\
=&(\omega(\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F}) - O(\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F}))^2 \\
=&\omega((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2)
\end{aligned}$$

which is contradictory to what we got about the upper bound.

So $\|\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F^2 = \mathbb{O}((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2)$.

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \\
=&\|\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \\
=&\|\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 + \|\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}\|_F^2 \\
&+ \operatorname{tr}(\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) \\
\geq&\|\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 + \min_{W\in\mathcal{O}(d)} \|\frac{XW}{\|X\|_F} - \frac{Y}{\|Y\|_F}\|_F^2 \\
&+ \operatorname{tr}(\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F})
\end{aligned}$$

Based on the same way in the upper bound part, we have with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\operatorname{tr}(\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F})^T(\frac{XW_xW^*}{\|X\|_F} - \frac{YW_y}{\|Y\|_F}) \sim o((\frac{1}{\|X\|_F} + \frac{1}{\|Y\|_F})^2) \\
&\|\frac{\hat{X}W^*}{\|\hat{X}\|_F} - \frac{XW_xW^*}{\|X\|_F} + \frac{YW_y}{\|\hat{Y}\|_F} - \frac{\hat{Y}}{\|Y\|_F}\|_F^2 \sim \frac{C(X)}{\|X\|_F^2} + \frac{C(Y)}{\|Y\|_F^2} \sim \frac{C(\hat{X})}{\|\hat{X}\|_F^2} + \frac{C(\hat{Y})}{\|\hat{Y}\|_F^2}
\end{aligned}$$

Then, with the probability at least $1 - n^C$ for any $C > 0$,

(i)Under H_0, there exists $\{a_n\}$ such that for sufficient large n,

$$\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \geq a_n \sim \frac{2C(\hat{X})}{\|\hat{X}\|_F^2};$$

(i)Under H_1, there exists $\{a_n\}$ such that for sufficient large n

$$\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \geq a_n \sim \frac{C(\hat{X})}{\|\hat{X}\|_F^2} + \frac{C(\hat{Y})}{\|\hat{Y}\|_F^2} + \min_{W\in\mathcal{O}(d)} \|\frac{XW}{\|X\|_F} - \frac{Y}{\|Y\|_F}\|_F^2$$

Conclusion:
with the probability at least $1 - n^C$ for any $C > 0$,
under H_0,
$\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \sim \frac{2C(\hat{X})}{\|\hat{X}\|_F^2}$;
under H_1:
$\min_{W\in\mathcal{O}(d)} \|\frac{\hat{X}W}{\|\hat{X}\|_F} - \frac{\hat{Y}}{\|\hat{Y}\|_F}\|_F^2 \sim \frac{C(\hat{X})}{\|\hat{X}\|_F^2} + \frac{C(\hat{Y})}{\|\hat{Y}\|_F^2} + \min_{W\in\mathcal{O}(d)} \|\frac{XW}{\|X\|_F} - \frac{Y}{\|Y\|_F}\|_F^2$

A.3 Proof of Theorem 3

Let
$W^* := \arg\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2$,
$W_x \in \mathcal{O}(d)$ such that $U_P S_P^{1/2} = XW_x$,
$W_y \in \mathcal{O}(d)$ such that $U_Q S_Q^{1/2} = YW_y$.

Upper Bound:
Let $W_0 = \arg\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(X)W_xW - \mathcal{P}(Y)W_y\|_F^2$

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2 \\
\leq&\|\mathcal{P}(\hat{X})W_0 - \mathcal{P}(\hat{Y})\|_F^2 \\
=&\|\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\|_F^2 + \|\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\|_F^2 \\
&+ \mathrm{tr}\big(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big) \\
=&\|\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\|_F^2 + \min_{W\in\mathcal{O}(d)} \|\mathcal{P}(X)W - \mathcal{P}(Y)\|_F^2 \\
&+ \mathrm{tr}\big(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)
\end{aligned}$$

Without loss of generality, we consider, with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\big(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\\
=&\mathrm{tr}(\mathcal{D}^{-1}\big(\hat{X}\big)\hat{X}W_0 - \mathcal{D}^{-1}(X)\hat{X}W_0\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\\
&+\mathrm{tr}\big(\mathcal{D}^{-1}(X)\hat{X}W_0 - \mathcal{D}^{-1}(X)XW_xW_0\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\\
\le&\sqrt{d}\|(\mathcal{D}^{-1}\big(\hat{X}\big)\hat{X}W_0 - \mathcal{D}^{-1}(X)\hat{X}W_0\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\|_F\\
&+\sqrt{d}\|\big(\mathcal{D}^{-1}(X)\hat{X}W_0 - \mathcal{D}^{-1}(X)XW_xW_0\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\|_F\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2) + \sqrt{d}\|\big(\mathcal{D}^{-1}(X)[\hat{X} - XW_x]\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\|_F\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2),
\end{aligned}$$

where we use the same way in Section A.1 combining Lemmas 14 and 13.

So, with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\mathrm{tr}\big(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW_0 - \mathcal{P}(Y)W_y\big)\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\|\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0 + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\|_F^2\\
=&\|\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0\|_F^2 + \|\mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\|_F^2\\
&+\mathrm{tr}(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0)^T(\mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y}))
\end{aligned}$$

And, with the probability at least $1 - n^C$ for any $C > 0$,

$$\begin{aligned}
&\mathrm{tr}(\mathcal{P}(\hat{X})W_0 - \mathcal{P}(X)W_xW_0)^T(\mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y}))\\
=&\mathrm{tr}([\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}^{-1}(X)]\hat{X}W_0)^T([\mathcal{D}^{-1}(\hat{Y}) - \mathcal{D}^{-1}(Y)]\hat{Y})\\
&+\mathrm{tr}([\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}^{-1}(X)]\hat{X}W_0)^T(\mathcal{D}^{-1}(Y)[Y - \hat{Y}W_y])\\
&+\mathrm{tr}(\mathcal{D}^{-1}(X)[X - \hat{X}W_x])^T([\mathcal{D}^{-1}(\hat{Y}) - \mathcal{D}^{-1}(Y)]\hat{Y})\\
&+\mathrm{tr}(\mathcal{D}^{-1}(X)[X - \hat{X}W_x])^T(\mathcal{D}^{-1}(Y)[Y - \hat{Y}W_y])\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2) + o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2) + o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2)\\
&\text{(based on Lemma 14)}\\
&+O((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2)\mathrm{tr}(\hat{X} - XW_x)^T(\hat{Y} - YW_y)\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2) \qquad \text{(based on lemma 9).}
\end{aligned}$$

So, with the probability at least $1-n^C$ for any $C>0$,

$$\begin{aligned}
&\|\mathcal{P}(\hat{X})W_0-\mathcal{P}(X)W_xW_0+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\|_F^2\\
\sim&\|\mathcal{P}(\hat{X})W_0-\mathcal{P}(X)W_xW_0\|_F^2+\|\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\|_F^2\\
\sim&C^*(X)+C^*(Y)\qquad\text{(based on Lemma 15)}\\
\sim&C^*(\hat{X})+C^*(\hat{Y})
\end{aligned}$$

Hence, with the probability at least $1-n^C$ for any $C>0$,
(i)Under H_0, there exists $\{a_n\}$ such that for sufficient large n,

$$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\le a_n\sim 2C^*(\hat{X});$$

(i)Under H_1, there exists $\{a_n\}$ such that for sufficient large n

$$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\le a_n\sim C^*(\hat{X})+C^*(\hat{Y})+\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F^2.$$

Lower Bound:
Suppose $\|\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\|_F^2=\omega((\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)^2)$, then

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\\
\ge&\big[\|\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\|_F-\|\mathcal{P}(\hat{X})W^*-\mathcal{P}(X)W_xW^*+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\|_F\big]^2\\
=&(\omega(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)-\theta(\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|))^2\\
=&\omega((\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)^2)
\end{aligned}$$

which is contradictory to what we got about the upper bound.
So $\|\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\|_F^2=\mathbb{O}((\|\mathcal{D}^{-1}(X)\|+\|\mathcal{D}^{-1}(Y)\|)^2)$.

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})W-\mathcal{P}(\hat{Y})\|_F^2\\
=&\|\mathcal{P}(\hat{X})W^*-\mathcal{P}(\hat{Y})\|_F^2\\
=&\|\mathcal{P}(\hat{X})W^*-\mathcal{P}(X)W_xW^*+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\|_F^2+\|\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\|_F^2\\
&+\operatorname{tr}\big(\mathcal{P}(\hat{X})W^*-\mathcal{P}(X)W_xW^*+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\big)\\
\ge&\|\mathcal{P}(\hat{X})W^*-\mathcal{P}(X)W_xW^*+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\|_F^2+\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(X)W-\mathcal{P}(Y)\|_F^2\\
&+\operatorname{tr}\big(\mathcal{P}(\hat{X})W^*-\mathcal{P}(X)W_xW^*+\mathcal{P}(Y)W_y-\mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW^*-\mathcal{P}(Y)W_y\big)
\end{aligned}$$

Based on the same way in the upper bound part, we have with the probability at least $1-n^C$ for any $C>0$,

$$\begin{aligned}
&\operatorname{tr}\big(\mathcal{P}(\hat{X})W^* - \mathcal{P}(X)W_xW^* + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\big)^T\big(\mathcal{P}(X)W_xW^* - \mathcal{P}(Y)W_y\big)\\
=&o((\|\mathcal{D}^{-1}(X)\| + \|\mathcal{D}^{-1}(Y)\|)^2)\\
&\|\mathcal{P}(\hat{X})W^* - \mathcal{P}(X)W_xW^* + \mathcal{P}(Y)W_y - \mathcal{P}(\hat{Y})\|_F^2 \sim C^*(\hat{X}) + C^*(\hat{Y})
\end{aligned}$$

Hence, with the probability at least $1-n^C$ for any $C>0$,
(i)Under H_0, there exists $\{a_n\}$ such that for sufficient large n,

$$\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2 \geq a_n \sim 2C^*(\hat{X});$$

(i)Under H_1, there exists $\{a_n\}$ such that for sufficient large n

$$\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2 \leq a_n \sim C^*(\hat{X}) + C^*(\hat{Y}) + \min_{W\in\mathcal{O}(d)} \|\mathcal{P}(X)W - \mathcal{P}(Y)\|_F^2.$$

Conclusion:
with the probability at least $1-n^C$ for any $C>0$,
under H_0,
$\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2 \sim 2C^*(\hat{X})$;
under H_1:
$\min_{W\in\mathcal{O}(d)} \|\mathcal{P}(\hat{X})W - \mathcal{P}(\hat{Y})\|_F^2$
$\sim C^*(\hat{X}) + C^*(\hat{Y}) + \min_{W\in\mathcal{O}(d)} \|\mathcal{P}(X)W - \mathcal{P}(Y)\|_F^2$.

A.4 Supplement results

Lemma 6. With the probability at least $1-n^C$ for any $C>0$,

$$C(\hat{M}) \sim C(M), \qquad C^*(\hat{M}) \sim C^*(M).$$

Proof. Follow Propositions A.1 and A.2 in Tang et al. (2017a) with Hoeffding's equality and the Borel-Cantelli Lemma. □

Lemma 7. For a matrix $D \in \mathbb{R}^{n\times d}$, there exist matrices V and M, where $V = [V_1|V_2\ldots|V_d]$ with $\|V_i\| = 1$ and $\|M\|_2 = \|D\|_2$, such that $D = VM$.

Proof. Do singular Value Decomposition on D. □

Lemma 8. $\|U_p^T(A-P)V\|_F = \mathbb{O}(\log^{1/2} n)$, with probability at least $1-n^{-C}$ for any $C>0$,where $V=[V_1|V_2\dots|V_d]$ with $\|V_i\|=1$.

Proof. Set u_i, v_i be i-th column of U_p and V, then

$$u_i(A-P)v_j = \sum_{k=1}^{n}\sum_{l=1}^{n}(A_{kl}-P_{kl})u_{ik}v_{jl}$$

where $|A_{kl}-P_{kl}|\le 1$ and $\mathbb{E}(A_{kl}-P_{kl})u_{ik}v_{jl}=0$, by Hoeffding's inequality, we get for some $C_2>0$,

$$\mathbb{P}(u_i(A-P)v_j \ge t) \le \exp(-\frac{2t^2}{C_2})$$

then for any $C>0$, let $t=(\frac{1}{2}cC_2\log n)^{1/2}$, where $c=C+2$, so

$$\mathbb{P}\Big(u_i(A-P)v_j \ge (\frac{1}{2}cC_2\log n)^{1/2}\Big) \le \exp(-c\log n) = n^{-c}$$

And then

$$\begin{aligned}
&\mathbb{P}\Big(u_i(A-P)v_j = \mathbb{O}(\log^{1/2} n)\Big)\\
\ge&\mathbb{P}\Big(u_i(A-P)v_j < (\frac{1}{2}cC_2\log n)^{1/2}\Big)\\
=&1-\mathbb{P}\Big(u_i(A-P)v_j \ge (\frac{1}{2}cC_2\log n)^{1/2}\Big)\\
\ge&1-n^{-c}
\end{aligned}$$

And because $U_p^T(A-P)V$ is $d\times d$ matrix with entries $u_i(A-P)v_j$,

$$\begin{aligned}
&\mathbb{P}\Big(\|U_p^T(A-P)V\|_F = \mathbb{O}(\log^{1/2} n)\Big)\\
\ge&1-d^2\mathbb{P}\Big(u_i(A-P)v_j \ge \frac{1}{2}cC_2\log n\Big)\\
=&1-d^2n^{-c} \ge 1-n^2n^{-c} = 1-n^{-(c-2)} = 1-n^{-C}
\end{aligned}$$

□

Lemma 9. $\mathrm{tr}(\hat{X}-XW_x)^T(\hat{Y}-YW_y)\to 0$ with probability at least $1-n^{-C}$ for any $C>0$.

Proof. We have

$$\hat{X} - XW_x = (A-P)U_P S_P^{-1/2} + M_1$$
$$\hat{Y} - YW_x = (B-Q)U_Q S_Q^{-1/2} + M_2,$$

where $\|M_1\|_F \to 0$ and $\|M_2\|_F \to 0$. Then,

$$\begin{aligned}
&\mathrm{tr}(\hat{X} - XW_x)^T(\hat{Y} - YW_y) \\
=&\mathrm{tr}[(A-P)U_P S_P^{-1/2} + M_1]^T[(B-Q)U_Q S_Q^{-1/2} + M_2] \\
=&\mathrm{tr}[(A-P)U_P S_P^{-1/2}]^T[(B-Q)U_Q S_Q^{-1/2}] + \mathrm{tr}M_1^T M_2 \\
&+ \mathrm{tr}[(A-P)U_P S_P^{-1/2}]^T M_2 + \mathrm{tr}M_1^T[(B-Q)U_Q S_Q^{-1/2}].
\end{aligned}$$

We know

$$\begin{aligned}
|\mathrm{tr}M_1^T M_2| &\le \|M_1\|_F \|M_2\|_F \to 0 \\
|\mathrm{tr}[(A-P)U_P S_P^{-1/2}]^T M_2| &\le \|(A-P)U_P S_P^{-1/2}\|_F \|M_2\|_F \to 0 \\
|\mathrm{tr}M_1^T[(B-Q)U_Q S_Q^{-1/2}] &\le \|M_1\|_F \|(B-Q)U_Q S_Q^{-1/2}\|_F \to 0.
\end{aligned}$$

So $\mathrm{tr}(\hat{X} - XW_x)^T(\hat{Y} - YW_y) \sim \mathrm{tr}[(A-P)U_P S_P^{-1/2}]^T[(B-Q)U_Q S_Q^{-1/2}]$. Next,

$$\mathrm{tr}[(A-P)U_P S_P^{-1/2}]^T[(B-Q)U_Q S_Q^{-1/2}] = \mathrm{tr}S_P^{-1/2}U_P^T(A-P)(B-Q)U_Q S_Q^{-1/2}.$$

We have $(S_P S_Q)^{-1/2} = \mathbb{O}(\sqrt{\frac{1}{\delta(P)\delta(Q)\gamma_2(P)\gamma_2(Q)}})$. Now we focus on $\mathrm{tr}U_P^T(A-P)(B-Q)U_Q$. The matrix is $d \times d$, so we consider $(U_P^T(A-P)(B-Q)U_Q)_{ii}$, which are diagonal entries for $1 \le i \le d$.

$$\begin{aligned}
&(U_P^T(A-P)(B-Q)U_Q)_{i,i} \\
=&(U_P)_i^T(A-P)(B-Q)(U_Q)_i \\
=&\mathrm{tr}(U_P)_i^T(A-P)(B-Q)(U_Q)_i \\
=&\mathrm{tr}(A-P)(B-Q)(U_Q)_i(U_P)_i^T \\
=:&\mathrm{tr}(A-P)C
\end{aligned}$$

So we have $C_{jk} = \sum_{l=1}^n (B-Q)_{jl}(U_Q)_{il}(U_P)_{ik}$. Based on Hoeffding's Inequality, we have with

probability at least $1 - n^{-C}$ for any $C > 0$,

$$C_{ij} = \mathbb{O}(\log^{1/2} n) \qquad \text{(By the similar way as in Lemma 8).}$$

Considering $\mathrm{tr}(A - P)C$, we have

$$\begin{aligned}
&\mathbb{P}[\mathrm{tr}(A-P)C \geq t] \\
=&\mathbb{P}\big[\mathrm{tr}(A-P)C \geq t | \max_{i,j} C_{ij} = \mathbb{O}(\log^{1/2} n)\big]\mathbb{P}\big[\max_{i,j} C_{ij} = \mathbb{O}(\log^{1/2} n)\big] \\
&+ \mathbb{P}\big[\mathrm{tr}(A-P)C \geq t | \max_{i,j} C_{ij} = \omega(\log^{1/2} n)\big]\mathbb{P}\big[\max_{i,j} C_{ij} = \omega(\log^{1/2} n)\big] \\
\leq&\mathbb{P}\big[\mathrm{tr}(A-P)C \geq t | \max_{i,j} C_{ij} = \mathbb{O}(\log^{1/2} n)\big] \times 1 + 1 \times n^{-C} \\
=&\mathbb{P}\big[\mathrm{tr}(A-P)C \geq t | \max_{i,j} C_{ij} = \mathbb{O}(\log^{1/2} n)\big] + n^{-C} \quad \text{for any } C > 0.
\end{aligned}$$

We know $\mathrm{tr}(A - P)C = \sum_{j,k}(A - P)_{kj}(C)_{jk}$, and $(A - P)$ is independent with C for that matrices A and B are independent, which we can obtain $\mathbb{E}[(A-P)_{kj}(C)_{jk}] = 0$. By Hoeffding's Inequality with the similar way, we have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\mathrm{tr}(A - P)C = \sum_{i,j} = \mathbb{O}(\log n).$$

Then $\mathrm{tr}S_P^{-1/2}U_P^T(A - P)(B - Q)U_Q S_Q^{-1/2} = \mathbb{O}(\frac{\log n}{\delta(P)\delta(Q)\gamma_2(P)\gamma_2(Q)})$ with probability at least $1 - n^{-C}$ for any $C > 0$. Hence we have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&\mathrm{tr}(\hat{X} - XW_x)^T(\hat{Y} - YW_y) \\
\sim&\mathrm{tr}S_P^{-1/2}U_P^T(A - P)(B - Q)U_Q S_Q^{-1/2} \\
=&\mathbb{O}(\frac{\log n}{\delta(P)\delta(Q)\gamma_2(P)\gamma_2(Q)}) \\
\to&0.
\end{aligned}$$

□

Lemma 10. $\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2 \sim C(X)$, with probability at least $1 - n^{-C}$ for any $C > 0$.

Proof. With probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2 \\
=&\min_{W \in \mathcal{O}(d)} \|\hat{X}W - X\|_F^2 \\
\geq&\min_{T \in \mathbb{R}^{d \times d}} \|\hat{X}T - X\|_F^2 \\
=&\|\hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}^T X - X\|_F^2 \\
=&\|(I - U_A U_A^T)U_P S_P^{1/2}\|_F^2 \\
=&\|(I - U_A U_A^T)PU_P S_P^{-1/2}\|_F^2 \\
=&\|(I - U_A U_A^T)(A - P)U_P S_P^{-1/2}\|_F^2 \\
\sim&\|(A - P)U_P S_P^{-1/2}\|_F^2 \\
\sim&C(X)
\end{aligned}$$

And on the other hand, there exists W_d such that with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2 \\
\leq&\|\hat{X} - XW_d\|_F^2 \\
\sim&C(X)
\end{aligned}$$

So $\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2 \sim C(X)$. □

Lemma 11. $\min_{W \in \mathcal{O}(d)} \|\frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW}{\|X\|_F}\|_F^2 \sim \frac{\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2}{\|X\|_F^2} \sim \frac{C(X)}{\|X\|_F^2}$, with probability at least $1 - n^{-C}$ for any $C > 0$.
Additional, $\|\frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW_2}{\|X\|_F}\|_F^2 \sim \frac{C(X)}{\|X\|_F^2}$, where $W_2 = \arg\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2$.

Proof. Let
$W_1 = \arg\min_{W \in \mathcal{O}(d)} \|\frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW}{\|X\|_F}\|_F^2$,
$W_2 = \arg\min_{W \in \mathcal{O}(d)} \|\hat{X} - XW\|_F^2$.

We have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$
\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW}{\|X\|_F} \|_F^2 \\
=& \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW_1}{\|X\|_F} \|_F^2 \\
\leq& \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW_2}{\|X\|_F} \|_F^2 \\
=& \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{\hat{X}}{\|X\|_F} \|_F^2 + \| \frac{\hat{X}}{\|X\|_F} - \frac{XW_2}{\|X\|_F} \|_F^2 + \mathbf{tr}(\frac{\hat{X}}{\|\hat{X}\|_F} - \frac{\hat{X}}{\|X\|_F})^T (\frac{\hat{X}}{\|X\|_F} - \frac{XW_2}{\|X\|_F}) \\
\leq& (1 - \frac{\|\hat{X}\|_F}{\|X\|_F})^2 + \frac{\|\hat{X} - XW_2\|_F^2}{\|X\|_F^2} + \sqrt{d} \| (\frac{\hat{X}}{\|\hat{X}\|_F} - \frac{\hat{X}}{\|X\|_F})^T (\frac{\hat{X}}{\|X\|_F} - \frac{XW_2}{\|X\|_F}) \|_F \\
\leq& o(\frac{1}{\|X\|_F^2}) + \frac{\|\hat{X} - XW_2\|_F^2}{\|X\|_F^2} + \sqrt{d} (1 - \frac{\|\hat{X}\|_F}{\|X\|_F}) \frac{\|\hat{X} - XW_2\|_F}{\|X\|_F} \\
=& o(\frac{1}{\|X\|_F^2}) + \frac{\|\hat{X} - XW_2\|_F^2}{\|X\|_F^2} + o(\frac{1}{\|X\|_F^2}) \\
\sim& \frac{\|\hat{X} - XW_2\|_F^2}{\|X\|_F^2} \\
\sim& \frac{C(X)}{\|X\|_F^2}
\end{aligned}
$$

Suppose we have $\|\hat{X} - XW_1\|_F^2 = \omega(1)$, then

$$
\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW}{\|X\|_F} \|_F^2 \\
=& \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{XW_1}{\|X\|_F} \|_F^2 \\
\geq& \big(\| \frac{\hat{X}}{\|X\|_F} - \frac{XW_1}{\|X\|_F} \|_F - \| \frac{\hat{X}}{\|\hat{X}\|_F} - \frac{\hat{X}}{\|X\|_F} \|_F \big)^2 \\
=& \big[\frac{\|\hat{X} - XW_2\|_F}{\|X\|_F} - (1 - \frac{\|\hat{X}\|_F}{\|X\|_F}) \big]^2 \\
=& (\omega(\frac{1}{\|X\|_F}) - o(\frac{1}{\|X\|_F^2}))^2 \\
=& \omega(\frac{1}{\|X\|_F^2})
\end{aligned}
$$

which is contradictory to what we got about the upper bound.
So $\|\hat{X} - XW_1\|_F^2 = \mathbb{O}(1)$.

On the other hand, with probability at least $1-n^{-C}$ for any $C>0$,

$$\begin{aligned}
&\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW}{\|X\|_F}\|_F^2\\
=&\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW_1}{\|X\|_F}\|_F^2\\
=&\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{\hat{X}}{\|X\|_F}\|_F^2+\|\frac{\hat{X}}{\|X\|_F}-\frac{XW_1}{\|X\|_F}\|_F^2+\mathbf{tr}(\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{\hat{X}}{\|X\|_F})^T(\frac{\hat{X}}{\|X\|_F}-\frac{XW_1}{\|X\|_F})\\
\geq&(1-\frac{\|\hat{X}\|_F}{\|X\|_F})^2+\|\frac{\hat{X}}{\|X\|_F}-\frac{XW_2}{\|X\|_F}\|_F^2-\sqrt{d}\|(\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{\hat{X}}{\|X\|_F})^T(\frac{\hat{X}}{\|X\|_F}-\frac{XW_1}{\|X\|_F})\|_F\\
\geq&o(\frac{1}{\|X\|_F^2})+\frac{\|\hat{X}-XW_2\|_F^2}{\|X\|_F^2}-\sqrt{d}(1-\frac{\|\hat{X}\|_F}{\|X\|_F})\frac{\|\hat{X}-XW_1\|_F}{\|X\|_F}\\
=&o(\frac{1}{\|X\|_F^2})+\frac{\|\hat{X}-XW_2\|_F^2}{\|X\|_F^2}-o(\frac{1}{\|X\|_F^2})\\
\sim&\frac{C(X)}{\|X\|_F^2}
\end{aligned}$$

So $\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW}{\|X\|_F}\|_F^2\sim\frac{C(X)}{\|X\|_F^2}$, with probability at least $1-n^{-C}$ for any $C>0$. Additional, we already have with probability at least $1-n^{-C}$ for any $C>0$,

$$\begin{aligned}
&\frac{C(X)}{\|X\|_F^2}\\
\sim&\min_{W\in\mathcal{O}(d)}\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW}{\|X\|_F}\|_F^2\\
\leq&\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW_2}{\|X\|_F}\|_F^2\\
\leq&o(\frac{1}{\|X\|_F^2})+\frac{\|\hat{X}-XW_2\|_F^2}{\|X\|_F^2}+o(\frac{1}{\|X\|_F^2})\\
\sim&\frac{C(X)}{\|X\|_F^2}
\end{aligned}$$

so $\|\frac{\hat{X}}{\|\hat{X}\|_F}-\frac{XW_2}{\|X\|_F}\|_F^2\sim\frac{C(X)}{\|X\|_F^2}$ □

Lemma 12. There exist $W_d\in\mathbb{R}^{d\times d}$ such that $\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W_d\|_F^2\sim C^*(X)$ with probability at least $1-n^{-C}$ for any $C>0$, where $C^*(X)=\mathbb{E}[\mathrm{tr}S_P^{-1/2}U_P^T(A-P)D^{-2}(A-P)U_PS_P^{-1/2}]$, $D=\mathcal{D}(X)$, with probability at least $1-n^{-C}$ for any $C>0$.

Proof. Follow the proof of Lemma A.5 in Tang et al. (2017a) with Logarithmic sobolev concentration inequality. □

Lemma 13. $\|U_P^T D^{-1}(A-P)U_P\|_F = \mathbb{O}(\frac{\log^{1/2} n}{\min \|X_i\|})$,
$\|U_P^T D^{-2}(A-P)U_P\|_F = \mathbb{O}(\frac{\log^{1/2} n}{\min \|X_i\|^2})$, where $D = \mathcal{D}(X)$.

Proof. The $ij-$th entry of $U_P^T D^{-1}(A-P)U_P$ is

$$\begin{aligned} & u_i^T(A-P)u_j \\ =& \sum_{k=1}^{n}\sum_{l=1}^{n}[(D^{-1}A)_{kl} - (D^{-1}P)_{kl}]u_{ik}u_{jl} \end{aligned}$$

With $[(D^{-1}A)_{kl} - (D^{-1}P)_{kl}] = \mathbb{O}(\frac{1}{\min \|X_i\|})$ and Hoeffding's inequality, for any $t > 0$, we have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$u_i^T D^{-1}(A-P)u_j = \mathbb{O}(\frac{\log^{1/2} n}{\min \|X_i\|}).$$

Because $U_P^T D^{-1}(A-P)U_P$ is a $d \times d$ matrix, with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\|U_P^T D^{-1}(A-P)U_P\|_F = \mathbb{O}(\frac{\log^{1/2} n}{\min \|X_i\|}).$$

Using the same way, we can get with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\|U_P^T D^{-2}(A-P)U_P\|_F = \mathbb{O}(\frac{\log^{1/2} n}{\min \|X_i\|^2}).$$

□

Lemma 14. $\|\big(\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}^{-1}(X)\big)\hat{X}\|_F^2 = o(\frac{1}{\min \|X_i\|^2})$,
$\|\big(\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}^{-1}(X)\big)X\|_F^2 = o(\frac{1}{\min \|X_i\|^2})$, with probability at least $1 - n^{-C}$ for any $C > 0$.

Proof. We have
$\|\big(\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}(X)\big)\hat{X}\|_F^2 = \sum_{i=1}^n \|\hat{X}_i\|^2\Big(\frac{1}{\|\hat{X}_i\|} - \frac{1}{\|X_i\|}\Big)^2 = \sum_{i=1}^n \frac{\big(\|\hat{X}_i\| - \|X_i\|\big)^2}{\|X_i\|^2}$.
And $\|\hat{X}_i\| - \|X_i\| = \frac{\|\hat{X}_i\|^2 - \|X_i\|^2}{\|\hat{X}_i\| + \|X_i\|} = \frac{A_i^T M_A A_i - P_i^T M P_i}{\|\hat{X}_i\| + \|X_i\|}$, where
$M_A = U_A S_A^{-1} U_A^T$, $M = U_P S_P^{-1} U_P^T$, A_i, P_i are ith row of A and P.

1:

We have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&|A_i^T M_A A_i - A_i^T M A_i| \\
=&\big|\|\hat{X}_i\|^2 - \|A_i^\top U_P S_P^{-1/2}\|^2\big| \\
=&\big|\|\hat{X}_i\| - \|AU_P S_P^{-1/2}\|\big| \cdot \big|\|\hat{X}_i\| + \|AU_P S_P^{-1/2}\|\big| \\
=&\mathbb{O}(a_i \frac{\log n}{\sqrt{n}}),
\end{aligned}$$

where $\sum_i a_i^2 = 1$, based on Lemmas A.3 and A.4 in Tang et al. (2017a).

2:

Let $E = A - P$, We have $2E_i^\top M P_i = \sum_{k=1}^n \sum_{l=1}^n E_{ik} P_{il} M_{kl}$,
we know
$\mathbb{E}(E_{ik} P_{il} M_{kl}) = 0$, $E_{ik} = \mathbb{O}(1)$, $\sum_l P_{il} = O(\delta(P))$, $\sum_k \sum_l M_{kl}^2 = \mathbb{O}(\frac{1}{(\delta(P)\gamma_2(P))^2})$.
By Hoeffding's inequality, we have $2E_i^\top M P_i = \mathbb{O}(\frac{\log^{1/2} n}{n})$, with probability at least $1 - n^{-C}$ for any $C > 0$.

3:

We have $\mathbb{E}(E_i^\top M E_i) = \sum_{k=1}^n \sum_{l=1}^n E_{ik} E_{il} M_{kl} = \sum_{k=1}^n \mathbb{E}(E_{ik}^2) M_{kk}$,
we know $\sum_k \mathbb{E}(E_{ik}^2) = \mathbb{O}(\delta(P))$, then by Hoeffding's inequality,
we have $\mathbb{E}(E_i^\top M E_i) = \mathbb{O}(\frac{\log^{1/2} n}{n})$, with probability at least $1 - n^{-C}$ for any $C > 0$.

4:

We have $\|\mathbb{E}_{ik}\|_{\varphi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}\|E_{ik}\|^p)^{1/p} \leq \mathbb{E}\|E_{ik}\| = \mathbb{O}(P_{ik})$.
By Hanson-Wright inequality, with probability at least $1 - n^{-C}$ for any $C > 0$, we have $|E_i M E_i - \mathbb{E}(E_i M E_i)| = \mathbb{O}(\frac{\log^{1/2} n}{n})$.

So with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&|A_i^T M_A A_i - P_i^T M P_i| \\
\leq &|A_i^T M_A A_i - A_i^T M A_i| + |A_i^T M A_i - P_i^T M P_i| \\
= &|A_i^T M_A A_i - A_i^T M A_i| + |2E_i M P_i + E_i M E_i| \\
\leq &|A_i^T M_A A_i - A_i^T M A_i| + |2E_i M P_i| + |\mathbb{E}(A_i^T M A_i)| + |A_i^T M A_i - \mathbb{E}(A_i^T M A_i)| \\
= &\mathbb{O}(\frac{\log^{1/2} n}{n} + a_i \frac{\log n}{\sqrt{n}}).
\end{aligned}$$

Then with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\|\hat{X}_i\| - \|X_i\| = \frac{A_i^T M_A A_i - P_i^T M P_i}{\|\hat{X}_i\| + \|X_i\|} = \mathbb{O}(\frac{\log^{1/2} n}{n|X_i|} + a_i \frac{\log n}{\sqrt{n}}). \tag{A.1}$$

Therefore, with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\|\big(\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}(X)\big)\hat{X}\|_F^2 = \sum_{i=1}^{n} \frac{\big(\|\hat{X}_i\| - \|X_i\|\big)^2}{\|X_i\|^2} = o(\frac{1}{\min \|X_i\|^2})$$

And by the same way, with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\|\big(\mathcal{D}^{-1}(\hat{X}) - \mathcal{D}(X)\big)X\|_F^2 = o(\frac{1}{\min \|X_i\|^2}).$$

□

Lemma 15. $\min_{W \in \mathcal{O}(d)} \|\mathcal{P}(\hat{X}) - \mathcal{P}(X)W\|_F^2 \sim C^*(X)$ with probability at least $1 - n^{-C}$ for any $C > 0$.

Proof. We have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&\min_{W \in \mathcal{O}(d)} \|\mathcal{P}(\hat{X}) - \mathcal{P}(X)W\|_F \\
=&\|\mathcal{P}(\hat{X}) - \mathcal{P}(X)W^*\|_F \\
=&\|\mathcal{D}^{-1}(\hat{X})\hat{X} - \mathcal{D}^{-1}(X)XW^*\|_F^2 \\
\geq&\|\mathcal{D}^{-1}(X)(\hat{X} - XW^*)\|_F - \|\mathcal{D}^{-1}(\hat{X})\hat{X} - \mathcal{D}^{-1}(X)\hat{X}\|_F \\
=&\|\mathcal{D}^{-1}(X)(\hat{X} - XW^*)\|_F - o(\frac{1}{\min \|X_i\|}) \qquad \text{(based on Lemma 14)}
\end{aligned}$$

And

$$\begin{aligned}
&\|\mathcal{D}^{-1}(X)(\hat{X} - XW_x)\|_F \\
\geq&\min_{T \in \mathbb{R}^{d \times d}} \|\mathcal{D}^{-1}(X)\hat{X} - \mathcal{D}^{-1}(X)XT\|_F \\
=&\|\mathcal{D}^{-1}(X)\hat{X} - \mathcal{D}^{-1}(X)XT\|_F \\
=&\|D^{-1}\hat{X} - D^{-1}X(X^TD^{-2}X)^{-1}X^TD^{-2}\hat{X}\|_F \qquad \big(D := \mathcal{D}(X)\big) \\
=&\|D^{-1}U_AS_A^{1/2} - D^{-1}U_PS_P^{1/2}(S_P^{1/2}U_P^TD^{-2}U_PS_P^{1/2})^{-1}S_P^{1/2}U_P^TD^{-2}U_AS_A^{1/2}\|_F \\
=&\|D^{-1}U_AS_A^{1/2} - D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}U_AS_A^{1/2}\|_F \\
=&\|D^{-1}U_AS_A^{1/2} - D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}[U_PU_P^T + (I - U_PU_P^T)]U_AS_A^{1/2}\|_F \\
=&\|D^{-1}U_AS_A^{1/2} - D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}U_PU_P^TU_AS_A^{1/2} \\
&- D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(I - U_PU_P^T)U_AS_A^{1/2}\|_F \\
=&\|D^{-1}U_AS_A^{1/2} - D^{-1}U_PU_P^TU_AS_A^{1/2} - D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(I - U_PU_P^T)U_AS_A^{1/2}\|_F \\
\geq&\|D^{-1}(I - U_PU_P^T)U_AS_A^{1/2}\|_F - \|D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(I - U_PU_P^T)U_AS_A^{1/2}\|_F
\end{aligned}$$

We have with probability at least $1 - n^{-C}$ for any $C > 0$,

$$\begin{aligned}
&\|D^{-1}(I - U_PU_P^T)U_AS_A^{1/2}\|_F \\
\sim&\|D^{-1}(A - P)U_PS_P^{-1/2}\|_F \\
\sim&\sqrt{C^*(X)},
\end{aligned}$$

and

$$
\begin{aligned}
&\|D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(I-U_PU_P^T)U_AS_A^{1/2}\|_F\\
\sim&\|D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(A-P)U_PS_P^{-1/2}\|_F\\
\leq&\|D^{-1}\|_2\|U_P\|_2\|(U_P^TD^{-2}U_P)^{-1}\|_2\|U_P^TD^{-2}(A-P)U_P\|_F\|S_P^{-1/2}\|_2\\
=&\mathbb{O}(\frac{1}{\min\|X_i\|})\mathbb{O}(\max\|X_i\|^2)\mathbb{O}(\frac{\log^{1/2}n}{\min\|X_i\|^2})\mathbb{O}(\frac{1}{\sqrt{\gamma_2(P)}}) \qquad \text{(based on Lemma 13)}\\
=&o(\frac{1}{\min\|X_i\|}).
\end{aligned}
$$

So with probability at least $1-n^{-C}$ for any $C>0$,

$$
\begin{aligned}
&\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W\|_F\\
\geq&\|\mathcal{D}^{-1}(X)(\hat{X}-XW^*)\|_F-o(\frac{1}{\min\|X_i\|})\\
\geq&\|D^{-1}(I-U_PU_P^T)U_AS_A^{1/2}\|_F-\|D^{-1}U_P(U_P^TD^{-2}U_P)^{-1}U_P^TD^{-2}(I-U_PU_P^T)U_AS_A^{1/2}\|_F\\
&-o(\frac{1}{\min\|X_i\|})\\
=&\|D^{-1}(I-U_PU_P^T)U_AS_A^{1/2}\|_F-o(\frac{1}{\min\|X_i\|})-o(\frac{1}{\min\|X_i\|})\\
\sim&\sqrt{C^*(X)}-o(\frac{1}{\min\|X_i\|})-o(\frac{1}{\min\|X_i\|})\\
\sim&\sqrt{C^*(X)}
\end{aligned}
$$

On the other hand, by Lemma 12, there exist W_d, such that with probability at least $1-n^{-C}$ for any $C>0$,
$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W\|_F\leq\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W\|_F\sim\sqrt{C^*(X)}$,
therefore, $\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W\|_F\sim\sqrt{C^*(X)}$, that is
$\min_{W\in\mathcal{O}(d)}\|\mathcal{P}(\hat{X})-\mathcal{P}(X)W\|_F^2\sim C^*(X)$, with probability at least $1-n^{-C}$ for any $C>0$. $\square$

APPENDIX

B

SUPPLEMENT TO CHAPTER 3

B.1 Proof of Theorem 4

Let $\mathcal{S} = \{(0,0),(0,1),(1,0),(1,1)\}$. Let $t = n(n-1)$. Now denote by

$$\mathcal{S}^t = \{(s_1, s_2, \ldots, s_t) \colon s_i \in \mathcal{S}\}$$

the set of tuples of length t whose elements are from $\mathcal{S}$. We can then view any realization (A_n, B_n) from the R-ER(P) model as corresponding to some element of $\mathcal{S}^t$. Let A_{ij} and B_{ij} denote the ijth element of A_n and B_n, respectively; note that, for ease of exposition, we dropped the index n from these notations. The second moment for the likelihood ratio

between $\mathcal{P}_n$ and $\mathcal{Q}_n$ is then given by

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}_n}\Big[\Big(\frac{\mathcal{P}_n(A_n,B_n)}{\mathcal{Q}_n(A_n,B_n)}\Big)^2\Big] &= \sum_{(A_n,B_n)\in\mathcal{S}^t}\frac{\mathbb{P}(A_n,B_n)^2}{\mathbb{Q}(A_n,B_n)} \\
&= \sum_{(A_n,B_n)\in\mathcal{S}^t}\prod_{i<j}\frac{\mathbb{P}(A_{ij},B_{ij})^2}{\mathbb{Q}(A_{ij},B_{ij})} \\
&= \prod_{i<j}\Big[\Big(\frac{\mathbb{P}_{ij}(1,1)^2}{\mathbb{Q}_{ij}(1,1)}+\frac{2\mathbb{P}_{ij}(1,0)^2}{\mathbb{Q}_{ij}(1,0)}+\frac{\mathbb{P}_{ij}(0,0)^2}{\mathbb{Q}_{ij}(0,0)}\Big) \\
&= \prod_{i<j}(1+R_{ij}^2)
\end{aligned}$$

The last equality in the above display is derived as follows (see also Eq. (3.3))

$$\begin{aligned}
\frac{\mathbb{P}_{ij}(1,1)^2}{\mathbb{Q}_{ij}(1,1)}+\frac{2\mathbb{P}_{ij}(1,0)^2}{\mathbb{Q}_{ij}(1,0)}+\frac{\mathbb{P}_{ij}(0,0)^2}{\mathbb{Q}_{ij}(0,0)} &= P_{ij}^2+2P_{ij}(1-P_{ij})R_{ij}+(1-P_{ij})^2R_{ij}^2 \\
&\quad +2P_{ij}(1-P_{ij})-4P_{ij}(1-P_{ij})R_{ij} \\
&\quad +2P_{ij}(1-P_{ij})R_{ij}^2 \\
&\quad +(1-P_{ij})^2+2P_{ij}(1-P_{ij})R_{ij}+P_{ij}^2R_{ij}^2 \\
&= 1+R_{ij}^2
\end{aligned}$$

We thus obtain

$$\mathbb{E}_{\mathcal{Q}_n}\Big[\Big(\frac{\mathcal{P}_n(A_n,B_n)}{\mathcal{Q}_n(A_n,B_n)}\Big)^2\Big] = \prod_{i<j}(1+R_{ij}^2). \tag{B.1}$$

Theorem 4 follows directly from Eq. (B.1) and the following technical lemma.

Lemma 16. $\limsup\prod_{i<j}(1+R_{ij}^2)<\infty$ iff $\limsup\|R\|_F^2<\infty$.

Proof. First suppose that $\limsup_{n\to\infty}\|R\|_F^2\le C$ for some finite constant $C>0$. Denote $N=n(n-1)/2$. Then by Jensen's inequality we have, for all but a finite number of n, that

$$\begin{aligned}
\log\Big(\prod_{i<j}(1+R_{ij}^2)\Big) &= \sum_{i<j}\log(1+R_{ij}^2) \\
&\le N\log\Big(1+\frac{1}{N}\sum_{i<j}R_{ij}^2\Big) \\
&\le \log\Big[\Big(1+\frac{1}{2N}\|R\|_F^2\Big)^N\Big]\le\log\Big[\Big(1+\frac{C}{2N}\Big)^N\Big]\le C/2
\end{aligned}$$

Conversely, suppose $\limsup_{n\to\infty}\sum_{i<j}\log(1+R_{ij}^2)\le C$ for some finite constant $C>0$. Then,

as $R_{ij}^2 \leq 1$ for all $\{i, j\}$ and $R_{ii} = 0$ for all i, we have

$$\begin{aligned}
\frac{1}{2}\|R\|_F^2 &\leq \sum_{i<j}\Big[e^{\log(1+R_{ij}^2)} - 1\Big] \\
&= \sum_{k=1}^{\infty}\sum_{i<j}\frac{\log^k(1+R_{ij}^2)}{k!} \\
&\leq \sum_{k=1}^{\infty}\sum_{i<j}\frac{\log(1+R_{ij}^2)\times\log^{k-1}2}{k!} \\
&= \Big(\sum_{i<j}\log(1+R_{ij}^2)\Big)\sum_{k=1}^{\infty}\frac{\log^{k-1}2}{k!} \leq \frac{C}{\log 2}.
\end{aligned}$$

as desired. □

B.2 Proof of Theorem 5

Suppose we are given an adjacency matrix S sampled from a planted clique model with edges probability $p = \frac{1}{2}$ and *unknown* clique size s_0. Let us generate a pair of *undirected* random graphs (A, B) as follows. The collection $\{(A_{ij}, B_{ij})\}$ for $i < j$ are *independent* bivariate random variables and furthermore, for any pair $i < j$,

$$\begin{aligned}
\mathbb{P}(A_{ij} = B_{ij} = 1 \mid S_{ij} = 0) = \mathbb{P}(A_{ij} = B_{ij} = 0 \mid S_{ij} = 0) = 0.5, \\
\mathbb{P}(A_{ij} = 1, B_{ij} = 0 \mid S_{ij} = 1) = \mathbb{P}(A_{ij} = 0, B_{ij} = 1 \mid S_{ij} = 1) = 0.5.
\end{aligned}$$

Let $\xi_{ij} = 1$ if vertices i and j is part of the planted clique in S and $\xi_{ij} = 0$ otherwise. Note that we can view the ξ_{ij} as deterministic quantities by assuming that the vertices forming the planted clique are chosen prior to adding the random edges in S. In particular, $S_{ij} = 1$ whenever $\xi_{ij} = 1$ and $S_{ij} \sim$ Bernoulli(0.5) otherwise. We therefore have

$$\begin{aligned}
\mathbb{P}(A_{ij} = 1, B_{ij} = 1) &= \mathbb{P}(A_{ij} = B_{ij} = 1 \mid S_{ij} = 0) \times \mathbb{P}(S_{ij} = 0) \\
&= 0.5 \times \big(\mathbb{P}(S_{ij} = 0, \xi_{ij} = 0) + \mathbb{P}(S_{ij} = 0, \xi_{ij} = 1)\big) \\
&= \tfrac{1}{4} \times 1\{\xi_{ij} = 0\}.
\end{aligned}$$

Similar reasonings yield

$$\mathbb{P}(A_{ij}=0,B_{ij}=0)=\mathbb{P}(A_{ij}=B_{ij}=0\mid S_{ij}=0)\times\mathbb{P}(S_{ij}=0)=\tfrac{1}{4}\times 1\{\xi_{ij}=0\},$$
$$\mathbb{P}(A_{ij}=0,B_{ij}=1)=0.5\times\mathbb{P}(S_{ij}=1)=\tfrac{1}{4}\times 1\{\xi_{ij}=0\}+\tfrac{1}{2}\times 1\{\xi_{ij}=1\},$$
$$\mathbb{P}(A_{ij}=1,B_{ij}=0)=\tfrac{1}{4}\times 1\{\xi_{ij}=0\}+\tfrac{1}{2}\times 1\{\xi_{ij}=1\},$$

and hence (A,B) is a realization of a R-correlated Erdős-Rényi graph with edges probability $p=\frac{1}{2}$ and $R_{ij}=-\xi_{ij}$ for all $\{i,j\}$ (see Eq. (3.3)).

Now suppose that either $s_0=0$ or $s_0=n^{1/4}$. Then, given S and the pair (A,B) randomly generated from S, we have

$$\mathbb{H}_0^{(1)}\colon \|R\|_F=0 \iff \mathbb{H}_0^{(2)}\colon S \text{ has no planted clique}$$
$$\mathbb{H}_A^{(1)}\colon \|R\|_F=n^{1/4} \iff \mathbb{H}_A^{(2)}\colon S \text{ has a planted clique of size at least } n^{1/4}.$$

Therefore, for any given instance $S\sim\text{PlantedClique}(n,1/2,s_0)$, there exists an instance (A,B) from R-ER$(1/2)$ where R is such that (1) the pair (A,B) is generated in polynomial time and (2) the planted clique problem on S is equivalent to deciding between the null and alternative hypothesis for $\|R\|_F$. Thus, assuming the Planted Clique conjecture holds, i.e., the PlantedClique problem requires quasi-polynomial time, there is no efficient algorithm for deciding between $\|R\|_F=0$ versus $\|R\|_F>0$ in this setting.

B.3 Proof of Theorem 6

Let $h_n=\rho_n h$ and $g_n=\gamma_n g$. Now recall Eq. (3.12). Then C corresponds to the adjacency matrix of a latent position graph with latent positions $\{(X_i,Y_i)\}_{i=1}^n$ and link function $f_n\colon \mathbb{R}^{d+d'}\times\mathbb{R}^{d+d'}\mapsto[0,1]$ given by

$$f_n((x,y),(x',y'))=h_n(x,x')+(1-g_n(y,y'))h_n(x,x')(1-h_n(x,x')). \tag{B.2}$$

Next suppose that $\|R\|_F=0$. Then by Theorem 3 in Xu (2018), for all $c>0$ there exists a constant C such that with probability at least $1-n^{-c}$

$$\|\hat{P}-P\|_F\le C(n\rho_n)^{\frac{s+d}{2s+d}},\quad\text{and}\quad \|\hat{H}-2P+P\circ P\|_F\le C(n\rho_n)^{\frac{s+d}{2s+d}}. \tag{B.3}$$

simultaneously. Similarly, suppose $\|R\| > 0$ holds. Once again by Theorem 3 in Xu (2018), there exists a constant C' such that with probability at least $1 - n^{-c}$,

$$\|\hat{P} - P\|_F \leq C(n\rho_n)^{\frac{s+d}{2s+d}}, \quad \text{and} \quad \|\hat{H} - H\|_F \leq C(n\rho_n)^{\frac{s+d+d'}{2s+d+d'}} \tag{B.4}$$

simultaneously. We note that the upper bound for $\|\hat{H} - H\|_F$ in Eq. (B.4) is larger than that in Eq. (B.3) and this is due mainly to the fact that if $\|R\|_F > 0$ then we will be using the latent positions $Z_i = (X_i, Y_i) \in \mathbb{R}^{d+d'}$ together with a link function f_n that is also at least s times continuously differentiable. We therefore have, with probability at least $1 - n^{-c}$, that

$$\|\hat{P} \circ \hat{P} - P \circ P\|_F^2 = \sum_{i,j} (\hat{P}_{ij} + P_{ij})^2 (\hat{P}_{ij} - P_{ij})^2 \leq 4\|\hat{P} - P\|_F^2 \leq 4C^2 (n\rho_n)^{\frac{2s+2d}{2s+d}}.$$

and hence, for $\|R\| = 0$ we have

$$\begin{aligned} \|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F &\leq \|\hat{H} - 2P + P \circ P\|_F + 2\|\hat{P} - P\|_F + \|\hat{P} \circ \hat{P} - P \circ P\|_F \\ &\leq 4C(n\rho_n)^{\frac{s+d}{2s+d}} \end{aligned}$$

with probability at least $1 - n^{-c}$. Similarly, for $\|R\| > 0$ we have

$$\begin{aligned} \|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F &\geq \|H - 2P - P \circ P\|_F - \big(\|\hat{H} - H\|_F + 2\|\hat{P} - P\|_F + \|\hat{P} \circ \hat{P} - P \circ P\|_F\big) \\ &\geq \|H - 2P - P \circ P\|_F - 4C(n\rho_n)^{\frac{s+d+d'}{2s+d+d'}} \\ &= \|R \circ (P - P \circ P)\|_F - 4C(n\rho_n)^{\frac{s+d+d'}{2s+d+d'}} \end{aligned}$$

Now let $\alpha = \frac{s+d+d'}{2s+d+d'}$ and note that $\frac{s+d}{2s+d} \leq \alpha$ for any choice of $s \geq 0, d \geq 0$ and $d' \geq 0$. Define $T(A, B)$ as the test statistic

$$T(A, B) = \frac{\|\hat{H} - 2\hat{P} + \hat{P} \circ \hat{P}\|_F}{(n\rho_n)^{\alpha} \log^{1/2} n}.$$

If $\|R\|_F = 0$ then $T(A, B) \to 0$ as $n \to \infty$. Furthermore if $\|R \circ (P - P \circ P)\|_F = \Omega((n\rho_n)^{\alpha'})$ for any $\alpha' > \alpha$ then $T(A, B) \to \infty$ as $n \to \infty$. Thus rejecting $\mathbb{H}_0$ for large values of $T(A, B)$ leads to an asymptotically valid and consistent test.

B.4 Proof of Corollary 1

If f and g are infinitely differentiable then, in place of Eq. (B.3) and Eq. (B.4), we have

$$\|\hat{P} - P\|_F = O((n\rho_n)^{1/2} \log^{d/2}(n\rho_n)), \quad \|\hat{H} - 2P + P \circ P\|_F$$

under $\mathbb{H}_0$ and

$$\|\hat{P} - P\|_F = O((n\rho_n)^{1/2} \log^{d/2}(n\rho_n)), \quad \|\hat{H} - H\|_F$$

under $\mathbb{H}_A$. See Theorem 4 in Xu (2018) for a statement of these bounds. The remaining steps follow the same argument as that presented in the proof of Theorem 6. We omit the details.

B.5 Proof of Theorem 7

Recall that the vertices of C are clustered using a community detection algorithm which guarantees exact recovery (see e.g., Abbe (2017); Gao et al. (2017); Lyzinski et al. (2014b)). We therefore have $\hat{\tau} = \tau$ asymptotically almost surely. Let us now condition on the event that $\hat{\tau} = \tau$. Then for any $k, \ell \in \{1, 2, \ldots, K\}$, the collection $\{(A_{ij}, B_{ij}) \colon \tau_i = k, \tau_j = \ell\}$ are iid bivariate random vectors with Pearson correlation $\rho_{k\ell}$. The central limit theorem then implies

$$\sqrt{n_{k\ell}}(\hat{\rho}_{k\ell} - \rho_{k\ell}) \xrightarrow{d} \mathcal{N}(0, (1 - \rho_{k\ell}^2)^2).$$

Furthermore, as $\hat{\rho}_{k\ell}$ depends only on the edges from vertices in the kth block to vertices in the ℓth block, the $\{\hat{\rho}_{k\ell}\}$ are *mutually* independent. Now suppose that the null hypothesis is true. Then $\rho_{k\ell} \equiv 0$ and the $\sqrt{n_{k\ell}}\hat{\rho}_{k\ell}$ are iid standard normals. In other words we have

$$\sum_{k \le \ell} n_{k\ell} \hat{\rho}_{k\ell}^2 \xrightarrow{d} \chi^2_{K(K+1)/2}$$

under $\mathbb{H}_0$. Next suppose that the alternative hypothesis is true and that there exists a constant $\mu > 0$ such that

$$\sum_{k \le \ell} n_{k\ell} \rho_{k\ell}^2 \to \mu.$$

Then for any k, ℓ, the term $\sqrt{n_{k\ell}}\rho_{k\ell}$ is bounded and thus, by Slutsky's theorem, we have

$$\sqrt{n_{k\ell}}(\hat{\rho}_{k\ell} - \rho_{k\ell}) \xrightarrow{d} \mathcal{N}(0, 1)$$

for all k, ℓ. We can then follow the same arguments as that in Guenther (1964) and show that the limiting distribution of $\sum_{k \le \ell} n_{k\ell} \hat{\rho}_{k\ell}^2$ depends on the $\{\rho_{k\ell}\}$ only through the quantity

$\sum_{k\leq \ell} n_{k\ell}\rho_{k\ell}^2$. Hence, as $\sum_{k\ell} n_{k\ell}\rho_{k\ell}^2 \to \mu$ we have by Slutksy's theorem that

$$\sum_{k\leq l} n_{kl}\hat{\rho}_{kl}^2 \xrightarrow{d} \chi^2_{K(K+1)/2}(\mu)$$

as desired.

APPENDIX

C

SUPPLEMENT TO CHAPTER 4

C.1 Proof of Theorems 10 and 12

For simplicity, set $X = X^{(1)}$, $P = P^{(1)}$ in the proof.

Let $R_X^{(4)} = U^\top P^2 U$, $R_X^{(2)}$ be the non-negative square root of $R_X^{(4)}$, and $R_X^{(1)}$ be non-negative the square root of $R_X^{(2)}$. Based on Theorem 7.2.6 in Horn and Johnson (2012), we know $R_X^{(2)} = U^\top |P| U$ with uniqueness.

Set W be the orthogonal square matrix W in Lemma 3 such that $\sqrt{2n}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(2n) \to \mathcal{N}(0, I_{d'})$. Then we have

$$\begin{aligned}
&\hat{U}^\top A^2 \hat{U} - W^\top R_X^{(4)} W \\
=&\hat{U}^\top A^2 \hat{U} - W^\top U^\top P^2 U W \\
=&\hat{U}^\top A^2 (\hat{U} - UW) + \hat{U}^\top (A^2 - P^2) UW + (\hat{U} - UW)^\top P^2 UW.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&(\hat{U} - UW)^\top P^2 UW \\
=&(\tilde{E}U\Sigma^{-1}W + R)^\top P^2 UW \\
=&W^\top \Sigma^{-1} U^\top \tilde{E} P^2 UW + R^\top P^2 UW \\
=&W^\top \Sigma^{-1} U^\top \tilde{E} U_P S_P U_P^\top PUW + R^\top P^2 UW,
\end{aligned}$$

where we know by Lemma 19

$$\begin{aligned}
&\|W^\top \Sigma^{-1} U^\top \tilde{E} U_P S_P U_P^\top PUW\| \\
\leq&\|\Sigma^{-1}\|\|U^\top \tilde{E} U_P\|\|S_P\|\|U_P^\top\|\|P\|\|U\| \\
=&O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}\big)\theta(n\rho)\theta(n\rho) \\
=&O_{\mathbb{P}}(\sqrt{\log n}n\rho),
\end{aligned}$$

and

$$\begin{aligned}
&\|R^\top P^2 UW\| \\
\leq&\|R^\top\|\|P\|\|P\|\|U\| \\
\leq&\sqrt{n}O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}\sqrt{\log n})\theta(n\rho)\theta(n\rho) \\
=&O_{\mathbb{P}}(\sqrt{\log n}n\rho).
\end{aligned}$$

Hence, we have

$$\|(\hat{U} - UW)^\top P^2 UW\| = O_{\mathbb{P}}(\sqrt{\log n}n\rho).$$

Similarly,

$$\|\hat{U}^\top A^2(\hat{U} - UW)\| = O_{\mathbb{P}}(\sqrt{\log n}n\rho).$$

Also, by Lemma 17, we have

$$\begin{aligned}
&\|\hat{U}^\top (A^2 - P^2)UW\| \\
\leq&\|\hat{U}^\top U_A S_A U_A^\top (A - P)UW\| + \|\hat{U}^\top (A - P)U_P S_P U_P^\top UW\| \\
=&O_{\mathbb{P}}(\sqrt{\log n}n\rho).
\end{aligned}$$

Thus,

$$\|\hat{U}^\top A^2 \hat{U} - W^\top R_X^{(4)} W\|_2 = O_{\mathbb{P}}(\sqrt{\log n}n\rho).$$

Consider Theorem 1 in Vu et al. (2021), and set Vu et al. (2021)'s $X := U^\top P^2 U/(n^2\rho^2)$,

Vu et al. (2021)'s $\Delta = (\hat{U}^\top A^2 \hat{U} - U^\top P^2 U)/(n^2\rho^2)$.

Then we know in Vu et al. (2021), $\delta_j = 0$ for $j > d$ and $\|\Delta\|_2 = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)) = o_{\mathbb{P}}(\sigma_d/2) = o_{\mathbb{P}}(1)$. So by Equation (13) in Theorem 1 in Vu et al. (2021), we have

$$\begin{aligned}\hat{R}_X^{(4)}/(n^2\rho^2) =& W^\top R_X^{(4)} W/(n^2\rho^2) + (\hat{U}^\top A^2 \hat{U} - U^\top P^2 U)/(n^2\rho^2) \\ & - U_2 U_2^\top (\hat{U}^\top A^2 \hat{U} - U^\top P^2 U) U_2 U_2^\top /(n^2\rho^2) \\ & + O(\|(\hat{U}^\top A^2 \hat{U} - U^\top P^2 U)/(n^2\rho^2)\|_F^2).\end{aligned}$$

So we have

$$\begin{aligned}& \|\hat{R}_X^{(4)} - W^\top R_X^{(4)} W\|/(n^2\rho^2) \\ \leq & O_{\mathbb{P}}(\sqrt{\log n} n\rho/(n^2\rho^2)) + O_{\mathbb{P}}(\sqrt{\log n} n\rho/(n^2\rho^2)) + O_{\mathbb{P}}((\log n) n^2\rho^2//(n^4\rho^4)) \\ = & O_{\mathbb{P}}(\sqrt{\log n} n\rho),\end{aligned}$$

then $\|\hat{R}_X^{(4)} - W^\top R_X^{(4)} W\| = O_{\mathbb{P}}(\sqrt{\log n} n\rho)$.

Consider Theorem 3.1 and Equation (3.1) in Carlsson (2018), let Carlsson (2018)'s $A := W^\top R_X^{(4)} W/(n^2\rho^2)$, Carlsson (2018)'s $E := (\hat{R}_X^{(4)} - W^\top R_X^{(4)} W)/(n^2\rho^2)$.

Then we know in Carlsson (2018),

$$\|\Lambda_\alpha\| = \theta(1), \|\hat{E}\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)),$$

then

$$\|B\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)), \|C\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)).$$

On the other hand, $(A + E) = \hat{R}_X^{(4)}$ has rank d, so $(\Lambda_\alpha + \hat{E}) = U^*(A + E)U$ has rank at most d (actually d). Also, we know $\Lambda_{\alpha^+} + B$ is invertible with probability at least $1 - n^{-C}$ for any $C > 0$ because $\lambda_{\alpha^+} = \theta(1)$ and $\|B\| = o_{\mathbb{P}}(1)$. By Guttman rank additivity formula, we have

$$\begin{aligned}& \operatorname{rank}(\Lambda_\alpha + \hat{E}) \\ = & \operatorname{rank}(\Lambda_{\alpha^+} + B) + \operatorname{rank}(C^\top (\Lambda_{\alpha^+} + B)^{-1} C + D - C^\top (\Lambda_{\alpha^+} + B)^{-1} C) \\ = & \operatorname{rank}(\Lambda_{\alpha^+} + B) + \operatorname{rank}(D),\end{aligned}$$

then we have $\operatorname{rank}(D) = \operatorname{rank}(\Lambda_\alpha + \hat{E}) - \operatorname{rank}(\Lambda_{\alpha^+} + B) = d - d = 0$, so $D = 0$.

Next, we have in Carlsson (2018),

$$\|[(\cdot)^{1/2}, \alpha]_1\| = \theta(1), \| \begin{pmatrix} B & C \\ C^\top & D^{1/2} \end{pmatrix} \| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)),$$

and

$$O(\|E\|^{3/2}) = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}\sqrt{\frac{\sqrt{\log n}}{n\rho}}).$$

Now we can obtain

$$\|(\hat{R}_X^{(2)} - W^\top R_X^{(2)} W)/(n\rho)\| = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}) + O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}\sqrt{\frac{\sqrt{\log n}}{n\rho}}) = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}),$$

so we have

$$\|\hat{R}_X^{(2)} - W^\top R_X^{(2)} W\| = O_{\mathbb{P}}(\sqrt{\log n}).$$

As the above, we consider Theorem 3.1 and Equation (3.1) in Carlsson (2018), let Carlsson (2018)'s $A := W^\top R_X^{(2)} W/(n\rho)$, Carlsson (2018)'s $E := (\hat{R}_X^{(2)} - W^\top R_X^{(2)} W)/(n\rho)$.

Then we know in Carlsson (2018),

$$\|\Lambda_\alpha\| = \theta(1), \|\hat{E}\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)),$$

then

$$\|B\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)), \|C\| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)).$$

On the other hand, $(A+E) = \hat{R}_X^{(2)}$ has rank d, so $(\Lambda_\alpha + \hat{E}) = U^*(A+E)U$ has rank at most d (actually d). Also, we know $\Lambda_{\alpha^+} + B$ is invertible with probability at least $1 - n^{-C}$ for any $C > 0$ with the same reason as the above. By Guttman rank additivity formula, we have

$$\begin{aligned} &\text{rank}(\Lambda_\alpha + \hat{E}) \\ =&\text{rank}(\Lambda_{\alpha^+} + B) + \text{rank}(C^\top(\Lambda_{\alpha^+} + B)^{-1}C + D - C^\top(\Lambda_{\alpha^+} + B)^{-1}C) \\ =&\text{rank}(\Lambda_{\alpha^+} + B) + \text{rank}(D), \end{aligned}$$

then we have $\text{rank}(D) = \text{rank}(\Lambda_\alpha + \hat{E}) - \text{rank}(\Lambda_{\alpha^+} + B) = d - d = 0$, so $D = 0$.

Next, we have in Carlsson (2018),

$$\|[(\cdot)^{1/2}, \alpha]_1\| = \theta(1), \| \begin{pmatrix} B & C \\ C^\top & D^{1/2} \end{pmatrix} \| = O_{\mathbb{P}}(\sqrt{\log n}/(n\rho)),$$

and

$$O(\|E\|^{3/2}) = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}\sqrt{\frac{\sqrt{\log n}}{n\rho}}).$$

Now we can obtain

$$\|(\hat{R}^{(1)} - W^\top R_X^{(1)} W)/(\sqrt{n\rho})\| = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}) + O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}\sqrt{\frac{\sqrt{\log n}}{n\rho}}) = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{n\rho}),$$

so we have

$$\|\hat{R}^{(1)} - W^\top R_X^{(1)} W\| = O_{\mathbb{P}}(\sqrt{\frac{\log n}{n\rho}}).$$

Let $R_X^{(2)} = U^\top |P| U =: U_X S_X U_X^\top$ be the eigen-decomposition, then $R_X^{(1)} = U_X S_X^{1/2} U_X^\top$. On the other hand,

$$U R_X^{(1)} U_X U_X^\top R_X^{(1)} U^\top = U U^\top U_P |S_P| U_P^\top U U^\top = U_P |S_P| U_P^\top$$

for $\mathcal{C}(U_P) \subset \mathcal{C}(U)$. So, there exists $W_X \in \mathcal{O}(d)$ such that $U R_X^{(1)} U_X = U_P |S_P|^{1/2} W_X$. Also, based on $U_P |S_P|^{1/2} = XQ$ for some $Q \in \mathcal{O}(p_X, q_X)$, we have $U R_X^{(1)} U_X = XQW_X$.

Now we have $U R_X^{(1)} = U U_X S_X^{1/2} U_X^\top = U R_X^{(1)} U_X U_X^\top = XQW_X U_X^\top$. Hence,

$$\begin{aligned}
&\tilde{X} - XQW_X U_X^\top W \\
=&\hat{U}\hat{R}_X^{(1)} - U R_X^{(1)} W \\
=&(\hat{U} - UW) W^\top R_X^{(1)} W + \hat{U}(\hat{R}_X^{(1)} - W^\top R_X^{(1)} W).
\end{aligned}$$

We have

$$\begin{aligned}
&\|\hat{U}(\hat{R}_X^{(1)} - W^\top R_X^{(1)} W)\|_{2\to\infty} \\
\le&\|\hat{U}\|_{2\to\infty} \|\hat{R}_X^{(1)} - W^\top R_X^{(1)} W\| \\
=&O_{\mathbb{P}}(\frac{1}{\sqrt{n}}\sqrt{\frac{\log n}{n\rho}}).
\end{aligned}$$

So we have

$$\|\tilde{X} - XQW_X U_X^\top W\| = O_{\mathbb{P}}(1), \|\tilde{X} - XQW_X U_X^\top W\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{\sqrt{n}}).$$

It satisfy Theorem 10, with setting $(U^*)^\top := W_X U_X^\top W$.

On the other hand,

$$(\tilde{X} - XQW_X U_X^\top W)_i = ((\hat{U} - UW)W^\top R_X^{(1)} W)_i + o_{\mathbb{P}}(\frac{1}{\sqrt{n}}),$$

that is,

$$W\tilde{X}_i - U_X W_X^\top Q^\top X_i = R_X^{(1)}(W\hat{U}_i - U_i) + o_{\mathbb{P}}(\frac{1}{\sqrt{n}}).$$

Let $U^* := U_X W_X^\top$, we have $(U^*)^\top U^* = I$. Let new $Q := Q^\top$, we also have $Q \in \mathcal{O}(p_X, q_X)$. Then, there exist matrices $U^* \in \mathbb{R}^{d' \times d}$ satisfying $(U^*)^\top U^* = I_d$ and $W \in \mathcal{O}(d')$ such that $W\tilde{X}_i - U^* QX = R_X^{(1)}(W\hat{U}_i - U_i) + o_{\mathbb{P}}(\frac{1}{\sqrt{n}})$.

It satisfy Theorem 12. Here we require the normality of $W\hat{U}_i - U_i$, which also requires Assumption 2.

C.2 Proof of Theorems 9 and 11

For simplicity, set $X = X^{(1)}$, $P = P^{(1)}$ in the proof.

Let R_X be such that $X = UR_X$, and we know $R_X = U^\top X$, so it is unique. Set W be the orthogonal square matrix W in Lemma 3 such that $\sqrt{2n}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(2n) \to \mathcal{N}(0, I_{d'})$. In the proof of Theorem 12, we have $\|\hat{R}_X^{(2)} - W^\top R_X^{(2)} W\| = O_{\mathbb{P}}(\sqrt{\log n})$. Also, we have $\hat{R}_X \hat{R}_X^\top = \hat{R}_X^{(2)}$ and $W^\top R_X Q Q^\top R_X^\top W = W^\top U^\top X Q Q^\top X^\top U W = W^\top U^\top U_P |S_P| U_P^\top U W = W^\top R_X^{(2)} W$. Based on Lemma A.1 in Tang et al. (2013) we know that there exists $W_X \in \mathcal{O}(d)$ such that

$$\|\hat{R}_X - W^\top R_X Q W_X\| = O_{\mathbb{P}}(\sqrt{\frac{\log n}{n\rho}}).$$

Hence,

$$\begin{aligned}
&\hat{X} - XQW_X \\
=&\hat{U}\hat{R}_X - UR_X QW_X \\
=&(\hat{U} - UW)W^\top R_X QW_X + \hat{U}(\hat{R}_X - W^\top R_X QW_X),
\end{aligned}$$

and we have

$$\begin{aligned}&\|\hat{U}(\hat{R}_X - W^\top R_X Q W_X)\|_{2\to\infty}\\ \leq&\|\hat{U}\|_{2\to\infty}\|\hat{R}_X - W_X^\top R_X Q W\|\\ =&O_{\mathbb{P}}(\frac{1}{\sqrt{n}}\sqrt{\frac{\log n}{n\rho}}).\end{aligned}$$

So we have

$$\|\tilde{X} - XQW_X U_X^\top W\| = O_{\mathbb{P}}(1), \|\tilde{X} - XQW_X U_X^\top W\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\log n}}{\sqrt{n}}).$$

It satisfy Theorem 9, with setting $(U^*)^\top := W_X U_X^\top W$.

On the other hand, follow the similar way in the proof of Lemma 3, we have

$$\sqrt{mn}(\bar{V}^*)^{-1/2}((Q^\top)^{-1}W_X\hat{X}_i - X_i)/(mn) \to I_{d'}$$

for $1 \leq i \leq n$, where $\bar{V}^*$ is defined in Lemma 3.

Based on $(Q^\top)^{-1} = I_{p_X,q_X} Q I_{p_X,q_X} = (Q^{-1})^\top \in \mathcal{O}(p_X, q_X)$, if we let new $Q = (Q^\top)^{-1}$, we have $Q \in \mathcal{O}(p_X, q_X)$. Then there exist matrices $W_X \in \mathcal{O}(d)$ and $Q \in \mathcal{O}(p_X, q_X)$ such that

$$\sqrt{2n}(\bar{V}^*)^{-1/2}(QW_X\hat{X}_i - X_i)/(2n) \to I_d,$$

where $\bar{V}^*$ is defined in Lemma 3, $\hat{X}_i$ and X_i for $1 \leq i \leq n$ are the rows of $\hat{X}$ and X. It satisfy Theorem 11.

C.3 Proof or Lemmas 1, 2, and 3

Proof of Lemma 1

For simplicity, we consider the condition where $m = 2$. For $m > 2$, it is similar.

We have

$$\begin{aligned}&XI_{p_1,q_1}X^\top = P = V_1 S V_1^\top,\\ &YI_{p_2,q_2}Y^\top = Q = V_2 S V_2^\top,\\ &\frac{1}{2}(XI_{p_1,q_1}X^\top + YI_{p_2,q_2}Y^\top) = \frac{1}{2}(P+Q) = V_1 S V_2^\top = V_2 S V_1^\top.\end{aligned}$$

So

$$\begin{aligned}
&\mathcal{C}(X) = \mathcal{C}(XI_{p_1,q_1}X^\top) = \mathcal{C}(V_1SV_1^\top) \subset \mathcal{C}(V_1),\\
&\mathcal{C}(Y) = \mathcal{C}(YI_{p_2,q_2}Y^\top) = \mathcal{C}(V_2SV_2^\top) \subset \mathcal{C}(V_2),\\
&\mathcal{C}(X) = \mathcal{C}(XI_{p_1,q_1}X^\top) = \mathcal{C}(2V_2SV_1 - YI_{p_2,q_2}Y^\top) \subset \mathcal{C}(V_2),\\
&\mathcal{C}(Y) = \mathcal{C}(YI_{p_2,q_2}Y^\top) = \mathcal{C}(2V_1SV_2 - XI_{p_1,q_1}X^\top) \subset \mathcal{C}(V_1).
\end{aligned}$$

Also,

$$M_P = \begin{pmatrix} X & 0 & \frac{1}{2}Y & \frac{1}{2}Y \\ \frac{1}{2}X & \frac{1}{2}X & Y & 0 \end{pmatrix} \begin{pmatrix} I_{p_1,q_1} & & & \\ & -I_{p_1,q_1} & & \\ & & I_{p_2,q_2} & \\ & & & -I_{p_2,q_2} \end{pmatrix} \begin{pmatrix} X & 0 & \frac{1}{2}Y & \frac{1}{2}Y \\ \frac{1}{2}X & \frac{1}{2}X & Y & 0 \end{pmatrix}^\top.$$

Hence,

$$\mathcal{C}(\begin{pmatrix} V_1 \\ V_2 \end{pmatrix}) = \mathcal{C}(M_P) \subset \mathcal{C}(\begin{pmatrix} X & 0 & \frac{1}{2}Y & \frac{1}{2}Y \\ \frac{1}{2}X & \frac{1}{2}X & Y & 0 \end{pmatrix}),$$

and then

$$\begin{aligned}
\mathcal{C}(V_1) \subset \mathcal{C}(\begin{pmatrix} X & 0 & \frac{1}{2}Y & \frac{1}{2}Y \end{pmatrix}) = \mathcal{C}(\begin{pmatrix} X|Y \end{pmatrix}),\\
\mathcal{C}(V_2) \subset \mathcal{C}(\begin{pmatrix} \frac{1}{2}X & \frac{1}{2}X & Y & 0 \end{pmatrix}) = \mathcal{C}(\begin{pmatrix} X|Y \end{pmatrix}).
\end{aligned}$$

Now, we have

$$\mathcal{C}(V_1) = \mathcal{C}(V_2) = \mathcal{C}(\begin{pmatrix} X|Y \end{pmatrix}),$$

besides,

$$\mathcal{C}(U) = \mathcal{C}(V_iV_i^\top) = \mathcal{C}(V_i),$$

thus,

$$\mathcal{C}(U) = \mathcal{C}(V_1) = \mathcal{C}(V_2) = \mathcal{C}(\begin{pmatrix} X|Y \end{pmatrix}.$$

Proof of Lemmas 2 and 3

For simplicity, we consider the condition where it is for RDPG and $m = 2$. For GRDPG and $m > 2$, it is similar.

Let $E = \begin{pmatrix} E_{\text{up}} \\ E_{\text{down}} \end{pmatrix} := M_A - M_P$, where E_{up} and E_{down} have the same size.

By Cape et al. (2019a) and Proposition 2 we have

$$\hat{V} - V = \begin{pmatrix} \hat{V}_1 \\ \hat{V}_2 \end{pmatrix} - \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} W = EVS^{-1}W + R,$$

where $\|R\|_{2\to\infty} = O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}(\log n)^{2\delta})$ $(2 < 2\delta)$, so

$$\hat{V}_1 - V_1 W = E_{\text{up}} V S^{-1} W + R_{\text{up}},$$

$$\hat{V}_2 - V_2 W = E_{\text{down}} V S^{-1} W + R_{\text{down}},$$

where R_{up} and R_{down} are the half upper and lower parts of R, respectively.

Hence, with high probability,

$$\begin{aligned}
\|\tilde{E}\| =& \|\frac{1}{2}(\hat{V}_1\hat{V}_1^\top + \hat{V}_2\hat{V}_2^\top) - V_1V_1^\top\| \\
\leq& \|E_{\text{up}}VS^{-1}V_1^\top\| + \|V_1S^{-1}V^\top E_{\text{up}}^\top\| + \|E_{\text{down}}VS^{-1}V_2^\top\| + \|V_2S^{-1}V^\top E_{\text{down}}^\top\| + o_{\mathbb{P}}(\frac{1}{\sqrt{n\rho}}) \\
=& O_{\mathbb{P}}(\frac{1}{\sqrt{n\rho}}) + o_{\mathbb{P}}(\frac{1}{\sqrt{n\rho}}) \\
=& O_{\mathbb{P}}(\frac{1}{\sqrt{n\rho}}),
\end{aligned}$$

for $\|E\| = O_{\mathbb{P}}(\sqrt{n\rho})$ (from Theorem 6 in Lu and Peng (2012)) and $\|S^{-1}\| = O(\frac{1}{n\rho})$. Based on

Bernstein inequalities, we have $\|EV\|_{2\to\infty} = O_{\mathbb{P}}(\sqrt{\rho \log n})$, so

$$\begin{aligned}
\|\tilde{E}\|_{2\to\infty} =& \|\frac{1}{2}(\hat{V}_1\hat{V}_1^\top + \hat{V}_2\hat{V}_2^\top) - V_1V_1^\top\|_{2\to\infty} \\
\leq& \|E_{\text{up}}VS^{-1}V_1^\top\|_{2\to\infty} + \|V_1S^{-1}V^\top E_{\text{up}}^\top\|_{2\to\infty} \\
&+ \|E_{\text{down}}VS^{-1}V_2^\top\|_{2\to\infty} + \|V_2S^{-1}V^\top E_{\text{down}}^\top\|_{2\to\infty} + o_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}) \\
\leq& \|E_{\text{up}}V\|_{2\to\infty}\|S^{-1}V_1^\top\| + \|V_1S^{-1}\|_\infty\|V^\top E_{\text{up}}^\top\|_{2\to\infty} \\
&+ \|E_{\text{down}}V\|_{2\to\infty}\|S^{-1}V_2^\top\| + \|V_2S^{-1}\|_\infty\|V^\top E_{\text{down}}^\top\|_{2\to\infty} + o_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}) \\
=& O_{\mathbb{P}}(\sqrt{\rho \log n})O(\frac{1}{n\rho}) + \mathbb{O}(\frac{1}{n\rho\sqrt{n}})\mathbb{O}_{\mathbb{P}}(\sqrt{n\rho \log n}) \\
&+ O_{\mathbb{P}}(\sqrt{\rho \log n})O(\frac{1}{n\rho}) + \mathbb{O}(\frac{1}{n\rho\sqrt{n}})\mathbb{O}_{\mathbb{P}}(\sqrt{n\rho \log n}) + o_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}) \\
=& O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}).
\end{aligned}$$

Furthermore, $\|\tilde{E}U\|_{2\to\infty} \leq \|\tilde{E}\|_{2\to\infty} = \mathbb{O}_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho})$, and

$$\begin{aligned}
&\|\tilde{E}^k U\|_{2\to\infty} \leq \|\tilde{E}^k\|_{2\to\infty} \\
\leq& \|\tilde{E}\|_{2\to\infty}\|\tilde{E}\|_2^{k-1} \\
=& \mathbb{O}_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}(n\rho)^{-(k-1)/2}).
\end{aligned}$$

By Davis-Kahan Theorem, we have $\min_{W\in\mathcal{O}(d')}\|\hat{U}-UW\|_F = O_{\mathbb{P}}((n\rho)^{-1/2})$, then

$$\begin{aligned}
\min_{W\in\mathcal{O}(d')}\|U^T\hat{U}-W\|_F &\le \|U^T\hat{U}-W^*\|_F \\
&= \|\Sigma^*-I\|_F \\
&= \sqrt{\sum_i(1-\sigma_i)^2} \\
&\le \sum_i(1-\sigma_i) \\
&\le \sum_i(1-\sigma_i^2) \\
&= \|\sin\Theta(U,\hat{U})\|_F^2 \\
&\le \min_{W\in\mathcal{O}(d')}\|\hat{U}-UW\|_F^2 \\
&= O_{\mathbb{P}}((n\rho)^{-1}),
\end{aligned}$$

where $W^* = W_1W_2$, $W_1\Sigma^*W_2$ is SVD of $U^T\hat{U}$, and σ_i's are singular values of $U^T\hat{U}$.

Additionally, based on Lemma 6.7 in Cape et al. (2019b) and Lemma 19, we have

$$\begin{aligned}
&\|\Sigma U^T\hat{U}-U^T\hat{U}\hat{\Sigma}\|_F \\
\le&\|U^\top(\frac{1}{2}(V_1V_1^\top+V_2V_2^\top)-\frac{1}{2}(\hat{V}_1\hat{V}_1^\top+\hat{V}_2\hat{V}_2^\top))\hat{U}\|_F \\
\le&\|U^\top\tilde{E}UW^*\|_F+\|U^\top\tilde{E}(I-UU^\top)\hat{U}\|_F+\|U^\top\tilde{E}U(W^*-U^\top\hat{U})\|_F \\
\le&\sqrt{d'}(\|U^\top\tilde{E}UW^*\|+\|U^\top\tilde{E}(I-UU^\top)\hat{U}\|+\|U^\top\tilde{E}U(W^*-U^\top\hat{U})\|) \\
\le&O(\|U^\top\tilde{E}U\|)+O(\|U^\top\|\|\tilde{E}\|\|(I-UU^\top)\hat{U}\|)+O(\|U^\top\tilde{E}U\|\|(W^*-U^\top\hat{U})\|) \\
=&O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}\big)+O_{\mathbb{P}}\big((n\rho)^{-1/2}(n\rho)^{-1/2}\big)+O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}(n\rho)^{-1}\big) \\
=&O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}\big)
\end{aligned}$$

and

$$\begin{aligned}
&\|\Sigma^{-k}U^T\hat{U}-U^T\hat{U}\hat{\Sigma}^{-k}\|_F \\
\le&\|\Sigma U^T\hat{U}-U^T\hat{U}\hat{\Sigma}\|_F\|H\|_F \\
=&O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}\big)O_{\mathbb{P}}(kC_H^{-(k+1)}) \\
=&O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}kC_H^{-(k+1)}\big),
\end{aligned}$$

where entries $h_{ij}=\frac{\sum_{l=0}^{k-1}\hat{\lambda}_j^{k-1-l}\lambda_i^l}{\hat{\lambda}_j^k\lambda_i^k}$, $\lambda_i,\hat{\lambda}_i$ are diagonal entries of $\Sigma,\hat{\Sigma}$, and the constant $C_H>0$.

By Equation (11) in Cape et al. (2019a), we have

$$\begin{aligned}\hat{U} &= \sum_{k=0}^{\infty} \tilde{E}^k M \hat{U} \hat{\Sigma}^{-(k+1)} \\ &= M\hat{U}\hat{\Sigma}^{-1} + \tilde{E}M\hat{U}\hat{\Sigma}^{-2} + \sum_{k=2}^{\infty} \tilde{E}^k M \hat{U} \hat{\Sigma}^{-(k+1)} \\ &=: S_1 + S_2 + S_3.\end{aligned}$$

We have

$$\begin{aligned}S_1 &= U\Sigma U^T \hat{U}\hat{\Sigma}^{-1} \\ &= UW + (UU^T\hat{U} - UW) + (U\Sigma U^T\hat{U}\hat{\Sigma}^{-1} - UU^T\hat{U}) \\ &= UW + R_1,\end{aligned}$$

where $\|R_1\|_{2\to\infty} = O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}\sqrt{\log n})$ for

$$\begin{aligned}&\|UU^T\hat{U} - UW\|_{2\to\infty} \le \|U\|_{2\to\infty}\|U^T\hat{U} - W\| = O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}) \\ &\|U\Sigma U^T\hat{U}\hat{\Sigma}^{-1} - UU^T\hat{U}\|_{2\to\infty} = O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}\sqrt{\log n}).\end{aligned}$$

Also,

$$\begin{aligned}S_2 &= \tilde{E}U\Sigma U^T\hat{U}\hat{\Sigma}^{-2} \\ &= \tilde{E}U\Sigma^{-1}W + (\tilde{E}U\Sigma^{-1}U^T\hat{U} - \tilde{E}U\Sigma^{-1}W) + (\tilde{E}U\Sigma U^T\hat{U}\hat{\Sigma}^{-2} - \tilde{E}U\Sigma^{-1}U^T\hat{U}) \\ &= \tilde{E}U\Sigma^{-1} + R_2,\end{aligned}$$

where $\|R_2\|_{2\to\infty} = O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{(n\rho)^2}\sqrt{\log n})$ for

$$\begin{aligned}
\|\tilde{E}U\Sigma^{-1}U^T\hat{U} - \tilde{E}U\Sigma W\|_{2\to\infty} &= \|\tilde{E}U\|_{2\to\infty}\|\Sigma^{-1}\|\|U^T\hat{U} - W\| \\
&= \|\tilde{E}U\|_{2\to\infty} O_{\mathbb{P}}\big((n\rho)^{-1}\big) \\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}(n\rho)^{-1}) \\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{(n\rho)^2}), \\
\|\tilde{E}U\Sigma U^T\hat{U}\hat{\Sigma}^{-2} - \tilde{E}U\Sigma^{-1}U^T\hat{U}\|_{2\to\infty} &= \|\tilde{E}U\|_{2\to\infty}\|\Sigma\|\|\Sigma^{-2}U^T\hat{U} - U^T\hat{U}\hat{\Sigma}^{-2}\|_F \\
&= \|\tilde{E}U\|_{2\to\infty} O_{\mathbb{P}}((n\rho)^{-1}\sqrt{\log n}) \\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{n\rho}(n\rho)^{-1}\sqrt{\log n}) \\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho \log n}}{(n\rho)^2}\sqrt{\log n}).
\end{aligned}$$

Also,

$$\begin{aligned}
S_3 &= \sum_{k=2}^{\infty} \tilde{E}^k U\Sigma U^T\hat{U}\hat{\Sigma}^{-(k+1)} \\
&= \sum_{k=2}^{\infty} \tilde{E}^k U\Sigma^{-k}W + (\tilde{E}^k U\Sigma^{-k}U^T\hat{U} - \tilde{E}^k U\Sigma^{-k}W) + (\tilde{E}^k U\Sigma U^T\hat{U}\hat{\Sigma}^{-(k+1)} - \tilde{E}^k U\Sigma^{-k}U^T\hat{U}),
\end{aligned}$$

so

$$
\begin{aligned}
&\|S_3\|_{2\to\infty}\\
\le&\sum_{k=2}^{\infty} O_{\mathbb{P}}(\sqrt{\rho\log n}(n\rho)^{-(k+1)/2}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}(n\rho)^{-(k+1)/2}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}k(C_H^2 n\rho)^{-(k+1)/2}\sqrt{\log n})\\
=&\sum_{k=2}^{\infty} O_{\mathbb{P}}(\sqrt{\rho\log n}(n\rho)^{-(k+1)/2}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}k(C_H^2 n\rho)^{-(k+1)/2}\sqrt{\log n})\\
=&O_{\mathbb{P}}(\sqrt{\rho\log n})O(\sum_{k=2}^{\infty}(n\rho)^{-(k+1)/2}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\sqrt{\log n})O(\sum_{k=2}^{\infty}k(C_H^2 n\rho)^{-(k+1)/2})\\
=&O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\frac{1}{\sqrt{n\rho}})O(\sum_{k=0}^{\infty}(\sqrt{n\rho})^{-k}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\sqrt{\log n}(C_H\sqrt{n\rho})^{-3})O(\sum_{k=0}^{\infty}(k+2)(C_H\sqrt{n\rho})^{-k})\\
=&O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\frac{1}{\sqrt{n\rho}})O(\frac{1}{1-\frac{1}{\sqrt{n\rho}}}) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\sqrt{\log n}(C_H\sqrt{n\rho})^{-3})O(\frac{3}{1-\frac{1}{C_H\sqrt{n\rho}}}+\frac{1}{(C_H\sqrt{n\rho}-1)^2}-1)\\
=&O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\frac{1}{\sqrt{n\rho}})O(1) + O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}\sqrt{\log n}(C_H\sqrt{n\rho})^{-3})O(1)\\
=&O_{\mathbb{P}}\big(n^{-\frac{1}{2}}(n\rho)^{-1}\sqrt{\log n}\big),
\end{aligned}
$$

for

$$
\begin{aligned}
\|\tilde{E}^k U\Sigma^{-k}W\|_{2\to\infty} &\le \|\tilde{E}^k U\|_{2\to\infty}\|\Sigma^{-k}\|\\
&= O_{\mathbb{P}}(\sqrt{\rho\log n}(n\rho)^{-(k+1)/2})\\
\|\tilde{E}^k U\Sigma^{-k}U^T\hat{U} - \tilde{E}^k U\Sigma^{-k}W\|_{2\to\infty} &\le \|\tilde{E}^k U\|_{2\to\infty}\|\Sigma^{-k}\|\|U^T\hat{U}-W\|\\
&= \|\tilde{E}^k U\|_{2\to\infty}O_{\mathbb{P}}\big((n\rho)^{-1}\big)\\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}(n\rho)^{-(k+1)/2}),\\
\|\tilde{E}^k U\Sigma U^T\hat{U}\hat{\Sigma}^{-(k+1)} - E^k U\Sigma^{-k}U^T\hat{U}\|_{2\to\infty} &= \|\tilde{E}^k U\|_{2\to\infty}\|\Sigma\|\|U^T\hat{U}\hat{\Sigma}^{-(k+1)} - \Sigma^{-(k+1)}U^T\hat{U}\|\\
&= \|\tilde{E}^k U\|_{2\to\infty}O_{\mathbb{P}}\big((n\rho)^{-1}\sqrt{\log n}kC_H^{-(k+1)}\big)\\
&= O_{\mathbb{P}}(\frac{\sqrt{\rho\log n}}{n\rho}k(C_H^2 n\rho)^{-(k+1)/2}\sqrt{\log n}).
\end{aligned}
$$

Now we have

$$\hat{U} - UW = \tilde{E}U\Sigma^{-1}W + R^*,$$

where $\|R^*\|_{2\to\infty} = O_{\mathbb{P}}(n^{-1/2}(n\rho)^{-1}\sqrt{\log n})$.

We have

$$\tilde{E}U = E_{\text{up}} V S^{-1} V_1^\top U + E_{\text{down}} V S^{-1} V_2^\top U + \tilde{R},$$

where $\|\tilde{R}\|_{2\to\infty} = O_{\mathbb{P}}(\frac{(\log n)^{2\delta}}{n\rho\sqrt{n}})$ based on Lemma 18.

Following the same process as in Proposition 2, we have

$$\begin{aligned}
&E_{\text{up}} V S^{-1} V_1^\top U + E_{\text{down}} V S^{-1} V_2^\top U \\
=&\begin{pmatrix}(A-P) & (B-Q)\end{pmatrix} \begin{pmatrix} I_n & I_n/2 & I_n/2 & 0 \\ 0 & I_n/2 & I_n/2 & I_n \end{pmatrix} \begin{pmatrix} V S^{-1} V_1^\top U \\ V S^{-1} V_2^\top U \end{pmatrix} \\
=&\begin{pmatrix}(A-P) & (B-Q)\end{pmatrix} \begin{pmatrix} V_1 S^{-1} W_1^\top + \frac{1}{2} V_1 S^{-1} W_2^\top + \frac{1}{2} V_2 S^{-1} W_1^\top \\ V_2 S^{-1} W_2^\top + \frac{1}{2} V_1 S^{-1} W_2^\top + \frac{1}{2} V_2 S^{-1} W_1^\top \end{pmatrix},
\end{aligned}$$

where W_1, W_2 satisfy $V_1 = U W_1$ and $V_2 = U W_2$. Then we also have $W_1^\top U^\top = V_1^\top$, $W_2^\top U^\top = V_2^\top$, $U^\top V_1 = W_1$, and $U^\top V_2 = W_2$.

On the other hand, we have

$$V S V^\top = \begin{pmatrix} P & (P+Q)/2 \\ (P+Q)/2 & Q \end{pmatrix},$$

and

$$V S^{-1} V^\top = \begin{pmatrix} V_1 S^{-1} V_1^\top & V_1 S^{-1} V_2^\top \\ V_2 S^{-1} V_1^\top & V_2 S^{-1} V_2^\top \end{pmatrix}.$$

Therefore,

$$\begin{aligned}
&\begin{pmatrix} P & (P+Q)/2 \\ (P+Q)/2 & Q \end{pmatrix} \begin{pmatrix} V_1 S^{-1} V_1^\top & V_1 S^{-1} V_2^\top \\ V_2 S^{-1} V_1^\top & V_2 S^{-1} V_2^\top \end{pmatrix} \\
=&V S V^\top V S^{-1} V^\top = V V^\top \\
=&\begin{pmatrix} V_1 V_1^\top & V_1 V_2^\top \\ V_2 V_1^\top & V_2 V_2^\top \end{pmatrix}.
\end{aligned}$$

Considering the element in the first row and the first column, we have

$$P(V_1 S^{-1} V_1^\top) + \frac{P+Q}{2}(V_2 S^{-1} V_1^\top) = V_1 V_1^\top,$$

and considering the one in the second row and the second column, we have

$$Q(V_2S^{-1}V_2^\top)+\frac{P+Q}{2}(V_1S^{-1}V_2^\top)=V_2V_2^\top,$$

Then, we have

$$\begin{aligned}
&\begin{pmatrix}P & Q\end{pmatrix}\begin{pmatrix}V_1S^{-1}V_1^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\\ V_2S^{-1}V_2^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\end{pmatrix}\\
=&P(V_1S^{-1}V_1^\top)+\frac{P}{2}(V_1S^{-1}V_2^\top)+\frac{P}{2}(V_2S^{-1}V_1^\top)+Q(V_2S^{-1}V_2^\top)+\frac{Q}{2}(V_1S^{-1}V_2^\top)+\frac{Q}{2}(V_2S^{-1}V_1^\top)\\
=&P(V_1S^{-1}V_1^\top)+\frac{P+Q}{2}(V_2S^{-1}V_1^\top)+Q(V_2S^{-1}V_2^\top)+\frac{P+Q}{2}(V_1S^{-1}V_2^\top)\\
=&V_1V_1^\top+V_2V_2^\top=2U\Sigma U^\top,
\end{aligned}$$

which has rank d', so $\begin{pmatrix}V_1S^{-1}V_1^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\\ V_2S^{-1}V_2^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\end{pmatrix}$ has rank at least d'. Furthermore, based on

$$\begin{pmatrix}V_1S^{-1}W_1^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\\ V_2S^{-1}W_2^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\end{pmatrix}\begin{pmatrix}U^\top & 0\\ 0 & U^\top\end{pmatrix}=\begin{pmatrix}V_1S^{-1}V_1^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\\ V_2S^{-1}V_2^\top+\frac{1}{2}V_1S^{-1}V_2^\top+\frac{1}{2}V_2S^{-1}V_1^\top\end{pmatrix},$$

we know $\begin{pmatrix}V_1S^{-1}W_1^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\\ V_2S^{-1}W_2^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\end{pmatrix}\in\mathbb{R}^{2n\times d'}$ has rank at least d', so it is full column rank.

Follow the same way in Proposition 2, we have

$$\sqrt{2n}(\bar{V}^*)^{-1/2}(W\hat{U}_i-U_i)/(2n)\to I_{d'}$$

for $1\le i\le n$, where $\bar{V}^*=(2n)^{-1}(\Xi_iV^*)^\top(\Xi_iV^*)$ and $V^*=\begin{pmatrix}I_n & I_n/2 & I_n/2 & 0\\ 0 & I_n/2 & I_n/2 & I_n\end{pmatrix}\begin{pmatrix}VS^{-1}V_1^\top U\Sigma^{-1}\\ VS^{-1}V_2^\top U\Sigma^{-1}\end{pmatrix}=\begin{pmatrix}V_1S^{-1}W_1^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\\ V_2S^{-1}W_2^\top+\frac{1}{2}V_1S^{-1}W_2^\top+\frac{1}{2}V_2S^{-1}W_1^\top\end{pmatrix}$.

Consider $\Xi_i^{**}:=[\mathrm{diag}(\begin{pmatrix}XX_i\\ YY_i\end{pmatrix})-\mathrm{diag}^2(\begin{pmatrix}XX_i\\ YY_i\end{pmatrix})][I-\mathrm{diag}(e_i)]$, $\bar{V}^*_{\mathrm{new}}=(2n)^{-1}(\Xi_i^*V^*)^\top(\Xi_i^*V^*)$, and $\bar{V}^*_{\mathrm{int}}=(2n)^{-1}(\Xi_i^{**}V^*)^\top(\Xi_i^{**}V^*)$. Based on Miller (1981), we have

$$(\bar{V}^*_{\mathrm{new}})^{-1}=(\bar{V}^*_{\mathrm{int}})^{-1}-\frac{1}{1+g_1}(\bar{V}^*_{\mathrm{int}})^{-1}(\bar{V}^*_{\mathrm{new}}-\bar{V}^*_{\mathrm{int}})(\bar{V}^*_{\mathrm{int}})^{-1},$$

where $g_1 = \text{tr}(\bar{V}^*_{\text{new}} - \bar{V}^*_{\text{int}})(\bar{V}^*_{\text{int}})^{-1}$, and

$$(\bar{V}^*_{\text{int}})^{-1} = (\bar{V}^*)^{-1} - \frac{1}{1+g_2}(\bar{V}^*)^{-1}(\bar{V}^*_{\text{int}} - \bar{V}^*)(\bar{V}^*)^{-1},$$

where $g_2 = \text{tr}(\bar{V}^*_{\text{int}} - \bar{V}^*)(\bar{V}^*)^{-1}$. We know $g_1, g_2 \geq 0$ for the psd of $(\bar{V}^*_{\text{new}} - \bar{V}^*_{\text{int}}), (\bar{V}^*_{\text{int}})^{-1}, (\bar{V}^*_{\text{int}} - \bar{V}^*), (\bar{V}^*)^{-1}$. Then,

$$\begin{aligned}
\|(\bar{V}^*_{\text{new}})^{-1} - (\bar{V}^*_{\text{int}})^{-1}\| &\leq |\frac{1}{1+g_1}|\|(\bar{V}^*_{\text{int}})^{-1}\|^2\|\bar{V}^*_{\text{new}} - \bar{V}^*_{\text{int}}\| = O(\frac{1}{\rho}) \\
\|(\bar{V}^*_{\text{int}})^{-1} - (\bar{V}^*)^{-1}\| &\leq |\frac{1}{1+g_2}|\|(\bar{V}^*)^{-1}\|^2\|\bar{V}^*_{\text{int}} - \bar{V}^*\| = O(\frac{1}{\rho}),
\end{aligned}$$

so $\|(\bar{V}^*_{\text{new}})^{-1} - (\bar{V}^*)^{-1}\| = O(\frac{1}{\rho})$. Based on Wihler (2009), we have

$$\|(\bar{V}^*_{\text{new}})^{-1/2} - (\bar{V}^*)^{-1/2}\| \leq d'\|(\bar{V}^*_{\text{new}})^{-1} - (\bar{V}^*)^{-1}\|^{1/2} = O(\frac{1}{\sqrt{\rho}}).$$

Now we have

$$\begin{aligned}
&\|\sqrt{2n}(\bar{V}^*_{\text{new}})^{-1/2}(W\hat{U}_i - U_i)/(2n) - \sqrt{2n}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(2n)\| \\
\leq &\frac{1}{\sqrt{2n}}O(\frac{1}{\sqrt{\rho}})\|W\hat{U}_i - U_i\| \\
\to &0.
\end{aligned}$$

Hence, we have $\sqrt{2n}(\bar{V}^*_{\text{new}})^{-1/2}(W\hat{U}_i - U_i)/(2n) \to I_{d'}$. That is,

$$\sqrt{2n}(\bar{V}^*)^{-1/2}(W\hat{U}_i - U_i)/(2n) \to I_{d'}$$

for $1 \leq i \leq n$, where $\bar{V}^* = (2n)^{-1}(\Xi_i^* V^*)^\top(\Xi_i^* V^*)$ and $V^* = \begin{pmatrix} I_n & I_n/2 & I_n/2 & 0 \\ 0 & I_n/2 & I_n/2 & I_n \end{pmatrix}\begin{pmatrix} VS^{-1}V_1^\top U\Sigma^{-1} \\ VS^{-1}V_2^\top U\Sigma^{-1} \end{pmatrix} = \begin{pmatrix} V_1S^{-1}W_1^\top + \frac{1}{2}V_1S^{-1}W_2^\top + \frac{1}{2}V_2S^{-1}W_1^\top \\ V_2S^{-1}W_2^\top + \frac{1}{2}V_1S^{-1}W_2^\top + \frac{1}{2}V_2S^{-1}W_1^\top \end{pmatrix}$.

C.4 Supplement Results

Proposition 2. If Assumptions 1 holds, M_A and M_P satisfy the Assumptions 1-5 in Cape et al. (2019a).

Proof. For simplicity, we consider the condition where it is for RDPG and $m = 2$. For GRDPG and $m > 2$, it is similar.

We have $\rho = \omega(\frac{\log^2 n}{n})$, so it satisfies Assumption 1.

Based on $\frac{\sigma_1(M_P)}{\sigma_{d''}(M_P)} \leq C_{10}$ in Assumption 2, it satisfies Assumption 2 in Cape et al. (2019a).

Set $E = M_A - M_P =: \begin{pmatrix} E_1 & E_3 \\ E_3 & E_2 \end{pmatrix}$, where E_1, E_2, E_3 have the same size. Then consider

$$E = \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix} + \begin{pmatrix} 0 & E_3^* \\ E_3^{*\top} & 0 \end{pmatrix} + \begin{pmatrix} 0 & E_3^{*\top} \\ E_3^* & 0 \end{pmatrix} + \begin{pmatrix} 0 & E_3 - E_3^* - E_3^{*\top} \\ E_3 - E_3^{*\top} - E_3^* & 0 \end{pmatrix}$$

where E_i^* is the upper-triangle (off diagonal) of E_i. The first three matrices satisfy the conditions for Theorem 6 in Lu and Peng (2012), so their spectral norms are $O_{\mathbb{P}}(\sqrt{n\rho})$. The fourth matrix is fixed with $2n$ non-zero entries of $\theta(\rho)$, so its spectral norm is $O(\sqrt{n\rho^2})$. Overall, $\|E\| = O_{\mathbb{P}}(\sqrt{n\rho})$. Therefore, it satisfies Assumption 3.

Consider

$$\begin{aligned} E =& \begin{pmatrix} E_1^* + E_1^{*\top} & 0 \\ 0 & E_2^* + E_2^{*\top} \end{pmatrix} + \begin{pmatrix} 0 & E_3^* + E_3^{*\top} \\ E_3^* + E_3^{*\top} & 0 \end{pmatrix} \\ &+ \begin{pmatrix} E_1 - E_1^* - E_1^{*\top} & 0 \\ 0 & E_2 - E_2^* - E_2^{*\top} \end{pmatrix} + \begin{pmatrix} 0 & E_3 - E_3^* - E_3^{*\top} \\ E_3 - E_3^{*\top} - E_3^* & 0 \end{pmatrix} \\ =:& E^{(1)} + E^{(2)} + E^{(3)} + E^{(4)} \end{aligned}$$

In $E^{(1)}$ and $E^{(2)}$, the entries satisfy $\mathbb{E}(e_{ij}/\sqrt{n\rho})^{g'} = O(\rho/(\sqrt{n\rho})^{g'})$. By Proposition 3.2 in Zhang and Tang (2022), we have there exist $C_{E^*} > 0$, such that for any $\delta > 1$, any integer $g \leq \log 2n$, and any vector $u \in \mathbb{R}^{2n}$ not dependent on $E^{(l)}$, where $l = 1, 2$, there exists $\nu > 0$ such that,

$$\begin{aligned} |\langle E^{(l)g} u, e_i \rangle| &\leq 2(\frac{\sqrt{C_{E^*}}}{4})^g \sqrt{2n\rho}^g (\log 2n)^{\delta g} \|u\|_\infty \\ &\leq (\sqrt{C_{E^*} 2n\rho})^g (\log 2n)^{\delta g} \|u\|_\infty \end{aligned}$$

with probability at least $1 - \exp[-\nu(\log n)^\delta]$.

$E^{(3)}$ is a diagonal matrix, and $|(E^{(3)})_{ii}| = \theta(\rho)$, so there exists $C_3 > 0$ such that $|(E^{(3)})_{ii}| \leq$

$\rho\sqrt{C_3}$. Then

$$\begin{aligned}
&|\langle E^{(3)g}u, e_i\rangle| \\
=&|(E^{(3)})^g_{ii} e_i^\top u| \\
\leq&|(E^{(3)})_{ii}|^g \|u\|_\infty \\
\leq&(\sqrt{C_3})^g \rho^g \|u\|_\infty \\
\leq&(\sqrt{C_3 2n\rho})^g (\log 2n)^{\delta g} \|u\|_\infty.
\end{aligned}$$

For $E^{(4)g}$, if g is even, then $E^{(4)}g$ is a diagonal matrix, and $|(E^{(4)})_{ii}|^g = \theta(\rho^g)$. By the same way as above, there exists $C_4 > 0$ such that,

$$|\langle E^{(4)g}u, e_i\rangle| \leq (\sqrt{C_4 2n\rho})^g (\log 2n)^{\delta g} \|u\|_\infty.$$

When g is odd, $|\langle E^{(4)g}u, e_i\rangle| = |(E_3)_{kk}(E^{(3)})^{g-1}_{ii} e_i^\top u| = \theta(\rho^g)|e_i^\top u|$, where $k \in [1, n]$ and $k \equiv i (\mathrm{mod}\ n)$. So there exists $C_4^* > 0$ such that $|(E_3)_{kk}(E^{(3)})^{g-1}_{ii}| \leq \rho\sqrt{C_4^*}$. Then,

$$\begin{aligned}
&|\langle E^{(4)g}u, e_i\rangle| \\
=&|(E_3)_{kk}(E^{(3)})^{g-1}_{ii} e_i^\top u| \\
\leq&(\sqrt{C_4^*})^g \rho^g \|u\|_\infty \\
\leq&(\sqrt{C_4^* 2n\rho})^g (\log 2n)^{\delta g} \|u\|_\infty.
\end{aligned}$$

Let C_E satisfy $\sqrt{C_E} \geq \sqrt{C_{E^*}} + \sqrt{C_3} + \sqrt{C_4} + \sqrt{C_4^*}$. Then we have there exist $C_{E^*} > 0$, such that for any $\delta > 1$ and any integer $g \leq \log 2n$, for each standard basis vector e_i and for each column vector v of V, we have

$$|\langle E^g v, e_i\rangle| \leq (C_{E^*} 2n\rho)^{g/2} (\log 2n)^{\delta g} \|u\|_\infty$$

with probability at least $1 - \exp[-\nu(\log n)^\delta]$. Hence, it satisfies Assumption 4.

Set $E_1^\circ = E_1^* + E_1^{*\top}$ and $E_2^\circ = E_2^* + E_2^{*\top}$. Consider

$$\begin{aligned}
E =& \begin{pmatrix} E_1^\circ & \frac{1}{2}(E_1^\circ + E_2^\circ) \\ \frac{1}{2}(E_1^\circ + E_2^\circ) & E_2^\circ \end{pmatrix} \\
&+ \begin{pmatrix} E_1 - E_1^* - E_1^{*\top} & 0 \\ 0 & E_2 - E_2^* - E_2^{*\top} \end{pmatrix} + \begin{pmatrix} 0 & E_3 - E_3^* - E_3^{*\top} \\ E_3 - E_3^{*\top} - E_3^* & 0 \end{pmatrix} \\
=:& E^{(5)} + E^{(3)} + E^{(4)}
\end{aligned}$$

Consider the upper n lines

$$(E^{(5)}V)_i = (\begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix} \begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_i = \sum_{k=1}^{2n} \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k$$

with independent terms, where $1 \le i \le n$. We have $\mathbb{E}(\begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k) = 0$ and $\mathrm{Var}(\begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k) = (\Xi_i \begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k (\Xi_i \begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k^\top$, where the square of $\Xi_i \in \mathbb{R}^{2n \times 2n}$ is the covariance matrix of $\begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_i$, and we know Ξ_i is diagonal with all positive values except that $(\Xi_i)_{ii} = (\Xi_i)_{n+i,n+i} = 0$. Set

$$\bar{V} = (2n)^{-1} \sum_{k=1}^{2n} \mathrm{Var}(\begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)$$

and $\nu^2 = \lambda_{\min}(\bar{V})$, so

$$\begin{aligned} \bar{V} =& (2n)^{-1} (\Xi_i \begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)^\top (\Xi_i \begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V) \\ =:& (2n)^{-1} (\Xi_i V^*)^\top (\Xi_i V^*) \\ =& (2n)^{-1} \sum_{k=1}^{2n} (\Xi_i)_{kk}^2 V_k^* V_k^{*\top} \end{aligned}$$

has eigenvalues of $\theta(n^{-1}\rho)$ with full rank. It follows that, there exists constant $C_\xi > 0$ such that

$$\begin{aligned} \sum_{k=1}^{2n} (\Xi_i)_{kk}^2 V_k^* V_k^{*\top} + \xi V_i^* V_i^{*\top} + \xi V_{n+i}^* V_{n+i}^{*\top} - \sum_{k=1}^{2n} \frac{\xi}{2} V_k^* V_k^{*\top} \succ 0 \\ \sum_{k=1}^{2n} \frac{\xi}{3} V_k^* V_k^{*\top} - (\xi V_i^* V_i^{*\top} + \xi V_{n+i}^* V_{n+i}^{*\top}) \succ 0, \end{aligned}$$

where ξ is the smallest non-zero values in Ξ_i^2. Similarly,

$$\sum_{k=1}^{2n} (\Xi_i)_{kk}^2 V_k^* V_k^{*\top} \sum_{k=1}^{2n} (\Xi_i)_{kk}^2 \mathbf{1}_{\{(\Xi_i)_{kk}^2 = \theta(\rho)\}} V_k^* V_k^{*\top} \succ 0$$

By Chebyshev's inequality, we have $\mathbb{P}(\| \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)\|^2 > \epsilon n\nu^2) = O(\frac{(\Xi_i)^4_{kk}}{n^2\rho^2})$ for any $\epsilon > 0$. So the condition

$$\lim_{n\to\infty} \frac{1}{2n\nu^2} \sum_{k=1}^{2n} \mathbb{E}(\| \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)\|^2 \mathbf{1}_{\{\| \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)\|^2 > \epsilon n\nu^2\}})$$

$$\leq \lim_{n\to\infty} \frac{1}{2n\nu^2} \sum_{k=1}^{2n} \mathbb{E}[C_5 n^{-1} \mathbf{1}_{\{\| \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)\|^2 > \epsilon n\nu^2\}}]$$

$$= \lim_{n\to\infty} \frac{1}{2n\nu^2} \sum_{k=1}^{2n} C_5 n^{-1} \mathbb{P}(\| \begin{pmatrix} E_1^\circ & E_2^\circ \end{pmatrix}_{ik} (\begin{pmatrix} I_n & I_n/2 \\ 0 & I_n/2 \end{pmatrix} V)_k)\|^2 > \epsilon n\nu^2)$$

$$\leq \lim_{n\to\infty} \theta(\frac{1}{n^2\rho}) = 0,$$

for some constant $C_5 > 0$.

By Multivariate Linderberg-Feller CLT, we have $\sqrt{2n}\bar{V}^{-1/2}(E^{(5)}V)_i/(2n) \to I_{d''}$ for $1 \leq i \leq n$. Then we know that $\rho^{-1/2}(E^{(5)}V)_i$ approximates to a centered multivariate normal distribution with Covariance matrix whose elements are $\theta(1)$ or 0, for $1 \leq i \leq n$. It is the same for $n+1 \leq i \leq 2n$.

For $E^{(3)}$, $(E^{(3)}V)_i = E^{(3)}_{ij} V_j = O(n^{-1/2})$ for some j, so $\rho^{-1/2}(E^{(3)}V)_i \to 0$ and

$$\sqrt{n}\bar{V}^{-1/2}(E^{(3)}V)_i/(2n) \to 0.$$

It is the same for $E^{(4)}$.

Hence, $\sqrt{2n}\bar{V}^{-1/2}(EV)_i/(2n) \to I_{d''}$ for $1 \leq i \leq n$. Then we know that $\rho^{-1/2}(EV)_i$ approximates to a centered multivariate normal distribution with Covariance matrix whose elements are $\theta(1)$ or 0, for all $1 \leq i \leq 2n$.

So it satisfies Assumption 5. □

Lemma 17. If matrices $U \in \mathbb{R}^{n\times d_1}$ and $V \in \mathbb{R}^{n\times d_2}$ satisfy $\sum_i U_{ik}^2 = \sum_i V_{il}^2 = 1$ for any $1 \leq k \leq d_1$ and $1 \leq l \leq d_2$, then $U^\top(A-P)V = O\big(\sqrt{\log n}\big)$, with probability at least $1 - n^{-C}$ for any $C > 0$.

Proof. We have

$$\begin{aligned}(U^\top(A-P)V)_{kl} &= \sum_{i=1}^n\sum_{j=1}^n (A-P)_{ij}U_{ik}V_{jl} \\ &= \sum_{i=1}^n\sum_{j=1}^n (A-P)_{ij}U_{ik}V_{jl}I\big(i\neq j\big) + \sum_{i=1}^n (A-P)_{ii}U_{ik}V_{il} \\ &=: \text{term}_1 + \text{term}_2,\end{aligned}$$

where $\mathbb{E}(A-P)_{ij}U_{ik}V_{jl} = 0$ and they are independent. By Hoeffding's inequality, we get

$$\begin{aligned}\mathbb{P}(|\text{term}_1| \geq t) &\leq 2\exp(-\frac{2t^2}{\sum_{i=1}^n\sum_{j=1}^n (A-P)_{ij}^2U_{ik}^2V_{jl}^2I\big(i\neq j\big)}) \\ &\leq 2\exp(-\frac{2t^2}{\sum_{i=1}^n\sum_{j=1}^n U_{ik}^2V_{jl}^2}) \\ &= 2\exp(-2t^2).\end{aligned}$$

Also,

$$\begin{aligned}|\text{term}_2| &\leq |\sum_{i=1}^n U_{ik}V_{il}| \\ &\leq \sqrt{(\sum_{i=1}^n U_{ik}^2)(\sum_{i=1}^n V_{il}^2)} \\ &= 1.\end{aligned}$$

Then for any $C > 0$, let $t = (\frac{C}{2}\log n + \log 2)^{1/2}$, we have

$$\mathbb{P}\Big(|U^\top(A-P)V)_{kl} - \text{term}_2| \geq (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \leq 2\exp(-C\log n - \log 2) = n^{-C}.$$

So

$$\begin{aligned}&\mathbb{P}\Big(|U^\top(A-P)V)_{kl} - \text{term}_2| = \mathbb{O}(\log^{1/2} n)\Big) \\ \geq&\mathbb{P}\Big(|U^\top(A-P)V)_{kl} - \text{term}_2| < (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \\ =&1 - \mathbb{P}\Big(|U^\top(A-P)V)_{kl} - \text{term}_2| \geq (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \\ \geq&1 - n^{-C}.\end{aligned}$$

Based on $|\text{term}_2| \le 1$, we know $(U^\top(A-P)V)_{kl}$ is still $O\big(\sqrt{\log n}\big)$, with probability at least $1-n^{-C}$ for any $C>0$. □

Lemma 18. If matrices $U \in \mathbb{R}^{n\times d_1}$ and $V \in \mathbb{R}^{2n\times d_2}$ satisfying $\sum_i U_{ik}^2 = \sum_i V_{il}^2 = 1$ for any $1 \le k \le d_1$ and $1 \le l \le d_2$, then $U^\top E^* V = O\big(\sqrt{\log n}\big)$, with probability at least $1-n^{-C}$ for any $C>0$, where E^* is either of E_{up} or E_{down} defined in the proof of Lemma 3.

Proof. We have

$$\begin{aligned}(U^\top E^* V)_{kl} &= \sum_{i=1}^{n}\sum_{j=1}^{2n} E^*_{ij}U_{ik}V_{jl} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2n} E^*_{ij}U_{ik}V_{jl}I\big((j-i)(j-i-n)\neq 0\big) + \sum_{i=1}^{n} E^*_{ii}U_{ik}V_{il} + E^*_{i(i+n)}U_{ik}V_{(i+n)l} \\ &= \text{term}_1 + \text{term}_2 :,\end{aligned}$$

where $\mathbb{E}E^*_{ij}U_{ik}V_{jl} = 0$ and they are independent. by Hoeffding's inequality, we get

$$\begin{aligned}\mathbb{P}(|\text{term}_1| \ge t) &\le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}\sum_{j=1}^{2n} E^{*\,2}_{ij}U^2_{ik}V^2_{jl}I\big((j-i)(j-i-n)\neq 0\big)}\Big) \\ &\le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}\sum_{j=1}^{2n} U^2_{ik}V^2_{jl}}\Big) \\ &= 2\exp(-2t^2).\end{aligned}$$

Also,

$$\begin{aligned}|\text{term}_2| &\le |\sum_{i=1}^{n} U_{ik}V_{il} + U_{ik}V_{(i+n)l}| \\ &\le \sqrt{(2\sum_{i=1}^{n}U^2_{ik})(\sum_{i=1}^{2n}V^2_{il})} \\ &= \sqrt{2}.\end{aligned}$$

Then for any $C>0$, let $t = (\frac{C}{2}\log n + \log 2)^{1/2}$, we have

$$\mathbb{P}\Big(|U^\top E^* V)_{kl} - \text{term}_2| \ge (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \le 2\exp(-C\log n - \log 2) = n^{-C}.$$

So

$$\begin{aligned}
&\mathbb{P}\Big(|U^\top E^* V)_{kl} - \text{term}_2| = \mathbb{O}(\log^{1/2} n)\Big) \\
\geq &\mathbb{P}\Big(|U^\top E^* V)_{kl} - \text{term}_2| < (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \\
= &1 - \mathbb{P}\Big(|U^\top E^* V)_{kl} - \text{term}_2| \geq (\frac{C}{2}\log n + \log 2)^{1/2}\Big) \\
\geq &1 - n^{-C}.
\end{aligned}$$

Based on $|\text{term}_2| \leq \sqrt{2}$, we know $(U^\top E^* V)_{kl}$ is still $O\big(\sqrt{\log n}\big)$, with probability at least $1 - n^{-C}$ for any $C > 0$. □

Lemma 19. If matrices $U \in \mathbb{R}^{n\times d_1}$ and $N \in \mathbb{R}^{n\times d_2}$ satisfying $\sum_i U_{ik}^2 = \sum_i N_{il}^2 = 1$ for any $1 \leq k \leq d_1$ and $1 \leq l \leq d_2$, then $U^\top \tilde{E} N = O\big((n\rho)^{-1}\sqrt{\log n}\big)$, where $\tilde{E}$ is defined in the proof of Lemma 3.

Proof. In the proof of Lemma 3, we have

$$\begin{aligned}
\tilde{E} =&E_{\text{up}} V S^{-2} V^\top E_{\text{up}}^\top + E_{\text{up}} V S^{-1} W R_{\text{up}} + E_{\text{up}} V S^{-1} V_1^\top + E_{\text{up}} W^\top S^{-1} V^\top E_{\text{up}}^\top \\
&+ R_{\text{up}} R_{\text{up}}^\top + R_{\text{up}} W^\top V_1^\top + V_1 S^{-1} V E_{\text{up}}^\top + V_1 R_{\text{up}}^\top \\
&+ E_{\text{down}} V S^{-2} V^\top E_{\text{down}}^\top + E_{\text{down}} V S^{-1} W R_{\text{down}} + E_{\text{down}} V S^{-1} V_1^\top \\
&+ E_{\text{down}} W^\top S^{-1} V^\top E_{\text{down}}^\top + R_{\text{down}} R_{\text{down}}^\top + R_{\text{down}} W^\top V_1^\top + V_1 S^{-1} V E_{\text{down}}^\top + V_1 R_{\text{down}}^\top.
\end{aligned}$$

Based on Lemma 18,

$$\begin{aligned}
(U^\top \tilde{E} N)_{kl} =&O\big((n\rho)^{-2}\log n\big) + O\big((n\rho)^{-1}\sqrt{\log n}\big) + O\big((n\rho)^{-1}\sqrt{\log n}\big) + o\big(n^{-\frac{1}{2}}(n\rho)^{-\frac{3}{2}}\sqrt{\log n}\big) \\
&+ o\big(n^{-\frac{1}{2}}(n\rho)^{-1}\big) + o\big(n^{-\frac{1}{2}}(n\rho)^{-\frac{1}{2}}\big) + O\big((n\rho)^{-1}\sqrt{\log n}\big) + o\big(n^{-\frac{1}{2}}(n\rho)^{-\frac{1}{2}}\big) \\
=&O\big((n\rho)^{-1}\sqrt{\log n}\big).
\end{aligned}$$

□

C.5 Addition for Gaussian Error

The proof for Theorems 13, 15, 10, 16 and Lemmas 4, 5 is similar for GRDPG, except that some bounds would be changed. For instance, we might use $\min(\sqrt{\log n}, \sqrt{n\rho})$ to replace $\sqrt{\log n}$ because ρ can be $o(\log n)$ for Gaussian Error.

The below lemma is similar as Lemma 17 but for Gaussian Error. The proof is sightly different.

Lemma 20. If matrices $U \in \mathbb{R}^{n\times d_1}$ and $V \in \mathbb{R}^{n\times d_2}$ satisfying $\sum_i U_{ik}^2 = \sum_i V_{il}^2 = 1$ for any $1 \le k \le d_1$ and $1 \le l \le d_2$, then $U^\top (A-P)V = O\big(1+\sqrt{\rho \log n}\big)$, with probability at least $1 - n^{-C}$ for any $C > 0$.

Proof. We have

$$\begin{aligned}(U^\top (A-P)V)_{kl} &= \sum_{i=1}^n \sum_{j=1}^n (A-P)_{ij} U_{ik} V_{jl} \\ &= \sum_{i=1}^n \sum_{j=1}^n (A-P)_{ij} U_{ik} V_{jl} I\big(i \neq j\big) + \sum_{i=1}^n (A-P)_{ii} U_{ik} V_{il} \\ &=: \mathrm{term}_1 + \mathrm{term}_2.\end{aligned}$$

We have

$$\begin{aligned}|\mathrm{term}_2| &\le |\sum_{i=1}^n U_{ik} V_{il}| \\ &\le \sqrt{(\sum_{i=1}^n U_{ik}^2)(\sum_{i=1}^n V_{il}^2)} \\ &= 1,\end{aligned}$$

and Gaussian distribution term_1 has

$$\begin{aligned}\mathrm{Var}(\mathrm{term}_1) &\le 2\sum_{i=1}^n \sum_{j=1}^n \mathrm{Var}\big((A-P)_{ij}\big) U_{ik}^2 V_{jl}^2 I\big(i \neq j\big) \\ &= \theta(\rho) \sum_{i=1}^n \sum_{j=1}^n U_{ik}^2 V_{jl}^2 I\big(i \neq j\big) \\ &= \Omega(\rho).\end{aligned}$$

We know there exists a positive $\rho_2 = \Omega(\rho)$, such that $\frac{\mathrm{term}_1}{\sqrt{\rho_2}} \sim \mathcal{N}(0,1)$. As $\mathcal{N}(0,1)$ is a sub-Gaussian distribution, there exist a $c > 0$, such that for every $t \ge 0$, we have

$$\mathbb{P}(|\frac{\mathrm{term}_1}{\sqrt{\rho_2}}| \ge t) \le 2\exp(-t^2/c).$$

Set $t = \sqrt{cC\log n + c\log 1/2}$, we have

$$\mathbb{P}\big(|\mathrm{term}_1| \ge \sqrt{\rho_2}\sqrt{cC\log n + c\log 1/2}\big) \le n^{-C}.$$

So

$$\mathbb{P}\big(|\text{term}_1| = O(\sqrt{\rho \log n})\big) \geq 1 - n^{-C}.$$

Based on $|\text{term}_2| \leq 1$, we have with probability at least $1 - n^{-C}$, $(U^\top (A - P)V)_{kl} = O(1 + \sqrt{\rho \log n})$). □