To be determined

by

Runze Tang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

July, 2017

Abstract

Abstract goes here.

Primary Reader: Carey Priebe

Secondary Reader: Minh Tang

Acknowledgments

Thanks!

Dedication

This thesis is dedicated to ...

Contents

Abstract **ii**

Acknowledgments **iii**

List of Tables **ix**

List of Figures **x**

1 Introduction **1**

1.1 Section 1

1.1.1 Subsection 1

1.2 Optional table of contents heading 2

1.2.1 Another subsection 2

1.2.1.1 Subsubsection 2

2 Random Graph Models **4**

2.1 Graphs and Random Graphs 5

2.1.1 Basic Concepts of Graphs 5

2.1.2 Random Graphs 6

2.2 Unweighted Random Graph Models 7

2.2.1 Independent Edge Model 7

2.2.2 Random Dot Product Graph 8

2.2.3 Stochastic Blockmodel 12

2.3 Weighted Random Graph Models 16

2.3.1 Weighted Independent Edge Model 16

2.3.2 Weighted Random Dot Product Graph 17

2.3.3 Weighted Stochastic Blockmodel 18

3 A Law of Large Graphs **20**

3.1 Model 27

3.2 Methods 28

3.2.1 Adjacency Spectral Embedding 28

3.2.2 Choosing Dimension 29

3.2.3 Graph Diagonal Augmentation 30

3.3 Estimators 31

3.3.1 Element-wise sample mean 31

3.3.2 Low-Rank Estimator 32

3.4 Theoretical Results 35

3.5 Finite Sample Toy Model Simulations 45

3.6 CoRR Brain Graphs Experiment 49
3.6.1 Dataset Description 49
3.6.2 Experiment Results 50
3.6.3 Exploration of Dimension Selection Procedures 56
3.6.4 Interpretability of Low-rank Methods 57
3.6.5 Challenges of the CoRR Dataset 60
3.6.6 Lobe Structure behind the Low-rank Methods 64
3.7 Synthetic Data Analysis for Full Rank IEM 67
3.8 Appendix: Proofs for Theory Results 71

4 Robust Generalizations of the Law of Large Graphs 76

4.1 Contamination Model 79
4.2 Estimators 81
4.2.1 Entry-wise Maximum Likelihood Estimator 81
4.2.2 Estimator $\widetilde{P}^{(1)}$ Based on Adjacency Spectral Embedding of $\widehat{P}^{(1)}$ 83
4.2.2.1 Rank-d Approximation 83
4.2.3 Entry-wise Maximum Lq-likelihood Estimator $\widehat{P}^{(q)}$ 86
4.2.4 Estimator $\widetilde{P}^{(q)}$ Based on Adjacency Spectral Embedding $\widehat{P}^{(q)}$. 88
4.3 Theoretical Results 89
4.3.1 MLE vs. MLqE 90
4.3.2 MLE vs. ASE of MLE 92
4.3.3 MLqE vs. ASE of MLqE 94

4.3.4 MLE vs. ASE of MLqE 95
4.3.5 Summary 96
4.4 Extensions 97
4.5 Simulations 99
4.5.1 Simulation Setting 99
4.5.2 Simulation Results 100
4.6 CoRR Brain Graphs Experiment 103
4.7 Appendix: Proofs for Theory Results 108
4.7.1 Outline of the Proofs 108
4.7.2 $\widehat{P}^{(q)}$ vs. $\widehat{P}^{(1)}$ 110
4.7.3 ASE Procedure of $\widehat{P}^{(1)}$ 117
4.7.4 $\widetilde{P}^{(1)}$ vs. $\widehat{P}^{(1)}$ 134
4.7.5 $\widetilde{P}^{(q)}$ vs. $\widehat{P}^{(q)}$ 145
4.7.6 $\widetilde{P}^{(q)}$ vs. $\widetilde{P}^{(1)}$ 156
4.7.7 Other Proofs 156

5 Discussion **159**

Vita **166**

List of Tables

1.1 This is what I want to have in the LOT 1

List of Figures

2.1 Example illustrating the stochastic blockmodel 14

3.1 Heat maps of the population mean, the sample mean, and the low-rank estimator 23
3.2 Example illustrating different estimates under the stochastic blockmodel 36
3.3 Asymptotic scaled relative efficiency in a 2-block SBM 43
3.4 Finite sample relative efficiency based on simulations 48
3.5 Relative efficiencies of two estimators for the CoRR data set 52
3.6 Heat plots of absolute estimation error for both estimators 54
3.7 Top 5 regions of the brain and top 50 connections between regions with the largest differences between two estimators 55
3.8 Comparison of MSE of two estimators for three atlases at three sample sizes for the CoRR data 58
3.9 Brain plots colored by the first 4 dimensions of embedding for the Desikan atlas 59
3.10 Relative error of the low-rank approximation of the population mean 61
3.11 Histogram of the population mean 61
3.12 Histogram of mean graph for Desikan atlas 63
3.13 Violin plot of the permutation test 68
3.14 Comparison of two estimators for synthetic data analysis 70

4.1 Roadmap among the data and four estimators 82
4.2 Relationship among four estimators 97
4.3 Simulation results based on different contamination ratios 102
4.4 Simulation results based on different parameters for robust estimator 103
4.5 Screeplot and the histogram of the eigenvalues of the mean of 114 graphs based on m2g pipeline 105
4.6 Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes 107

Chapter 1

Introduction

Introduction.

A citation [?]. A citation without brackets ?. Multiple citations [???].

1.1 Section

This is a section. Here's a reference to a different section: 1.1.1.

1.1.1 Subsection

This is a subsection.

Table 1.1: This is a caption.

A	B
a1	b1
a2	b2
a3	b3

Table 1.1 ... continued

A	B
a4	b4

1.2 Section with linebreaks in the name

This is another section.

1.2.1 Another subsection

1.2.1.1 Subsubsection

1.2.1.1.1 Heading level below subsubsection

And I quote:

> La la la.

No ident after end of quote.

Another paragraph with a list:

- Item 1
- Item 2

Again, we don't indent here.

Chapter 2

Random Graph Models

Our main focus of this work is a collection of random graphs. As a first step, we will introduce some models for a single random graph in this chapter. These important components will not only lead to our model for a collection of graphs but also motivate our estimators later in Chapter 3 and Chapter 4.

We start the chapter with some basic concepts of graphs and random graphs in Section 2.1. Then in Section 2.2, random graph models for unweighted graphs are introduced. In particular, it is concerned with independent edge model, random dot product graph, and stochastic blockmodel. Their relationship will also be discussed. Section 2.3 generalizes all three models introduced in Section 2.2 so that they can be adapted to weighted graphs.

2.1 Graphs and Random Graphs

In this section we will introduce some basic concepts of graphs and random graphs.

2.1.1 Basic Concepts of Graphs

A *graph* is an ordered pair $G = (V, E)$ comprising a set V of vertices together with a set E of edges. Denote the number of vertices $|V|$ to be n. Without loss of generality, we assume $V = [n] = \{1, 2, \ldots, n\}$. Each edge is associated with two vertices. We say a graph has no self-loops if each edge is associated with two distinct vertices. And a graph is undirected if all its edges have no orientation. Thus for a undirected graph without self-loops, we have the edge set $E \subset \{\{u, v\} : u, v \in V, u \neq v\}$. This will be our basic setting throughout the entire paper.

Every graph can be represented in the form of adjacency matrix $a \in \mathcal{A}$. For unweighted graphs, $\mathcal{A} = \{0, 1\}^{n \times n}$ so that the adjacency matrices are binary. For $i, j \in [n]$, $a_{ij} = 1$ indicates that there is an edge from vertex i to vertex j and $a_{ij} = 0$ otherwise. Note that a is always symmetric since we assume the graphs to be undirected in this work. Thus $a_{ij} = 1$ means there is an edge between vertex i and vertex j and $a_{ji} = 1$ as well. For weighted graphs, each edge is assigned with a positive real-valued weight, i.e. $\mathcal{A} = \mathbb{R}_{\geq 0}^{n \times n}$. Similarly, $a_{ij} > 0$ denotes the weight assigned to the edge between vertex i and vertex j, and $a_{ij} = 0$ means there is no edge between these two vertices.

As mentioned above, we assume the graph has no self-loops and are undirected. So every adjacency matrix a is hollow and symmetric, that is $a_{ii} = 0$ for $i \in [n]$ and $a_{ij} = a_{ji}$ for $i, j \in [n]$. This is always assumed without specific clarifications in later sections.

2.1.2 Random Graphs

A random graph is a graph with fixed vertex set whose edges are randomly distributed with respect to some distributions. Mathematically, a random graph $A : \Omega \mapsto \mathcal{A}$ is a map from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the space of all adjacency matrices on n vertices.

Example 2.1.1 (Erdős-Rényi Graphs) *The first random graph model is the* Erdős-Rényi graphs (ER) *introduced by [Gilbert, 1959]. They are unweighted graphs where each edge is present with the same probability $p \in [0, 1]$ independently. In our setting (undirected graphs without self-loops), that is $A_{ij} \overset{iid}{\sim}$ Bernoulli(p) for $i, j \in [n]$ with $i < j$. Thus for $a \in \mathcal{A}$ we have*

$$\mathbb{P}(A = a) = \prod_{i<j} p^{a_{ij}}(1-p)^{1-a_{ij}}.$$

In later sections, we will introduce other random graph models which capture different properties of the graphs in practice respectively.

2.2 Unweighted Random Graph Models

In this section, we will focus on unweighted graphs and introduce three important models, i.e. independent edge model, random dot product graph, and stochastic blockmodel.

2.2.1 Independent Edge Model

As introduced in Example 2.1.1, we see ER model is quite restrictive since all edges follow the same Bernoulli distribution with parameter p. Here we consider a much more general model.

Definition 2.2.1 (Independent Edge Model) *Under an* independent edge model (IEM) *proposed by Bollobás et al. [2007], for $i, j \in [n]$ and $i < j$, the edge between vertex i and vertex j is present with probability $p_{ij} \in [0, 1]$ independently, i.e. $A_{ij} \overset{ind}{\sim}$ Bernoulli(p_{ij}). Let $P = (p_{ij})_{i,j=1}^{n} \in [0, 1]^{n \times n}$ be the parameter matrix consists of all probabilities for Bernoulli distributions, then the model is denoted by IEM(P). Thus for $a \in \mathcal{A}$ we have*

$$\mathbb{P}(A = a) = \prod_{i<j} p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}}.$$

Note that the graphs considered in this paper are always assumed to be undirected without self-loops, thus the parameter matrix P needs to be symmetric and hollow. However, for convenience, we still define the parameters to be an n-by-n matrix while

only $\binom{n}{2}$ of them are effective.

Compared to the ER model, IEM allows the probabilities of the existence of the edges to vary while keeping the independence. Different probabilities give IEM the freedom to include a wide range of random graph distributions, but definitely not all of them because of independence assumption. The most general model is a multinomial distribution with $2^{\binom{n}{2}}$ choices from all possible unweighted and undirected graphs with no self-loops on n vertices, which imposes no assumptions on the graph structure. While such model is instructive sometimes, IEM will be the most general model we consider in this paper.

2.2.2 Random Dot Product Graph

For a graph, the adjacencies between vertices generally depend on unobserved properties of the corresponding vertices. For instance, in a social network setting, people are more likely to be friends if they have shared interests; in a connectomics setting, the two brain regions with similar properties will have similar connectivity patterns compared to other regions of the brain. The latent positions model (LPM) proposed by [**?**] captures such structure, where each vertex is associated with a latent position that influences the adjacencies for that vertex. In particular, we are interested in the case that the latent positions are random in this work.

Definition 2.2.2 (Latent Position Model) *Let the latent position for vertex i be*

a random variable $X_i : \Omega \mapsto \mathcal{X}$*, where* $\mathcal{X}$ *is the latent space and* $i \in [n]$*. Let the link function be a symmetric map* $\kappa : \mathcal{X}^2 \mapsto [0, 1]$*. Then under a* latent position model (LPM)*, for* $i, j \in [n]$ *and* $i < j$*, conditioned on their respective latent positions* X_i *and* X_j*,*

$$A_{ij}|(X_i, X_j) \overset{ind}{\sim} \text{Bernoulli}(\kappa(X_i, X_j)).$$

Thus for $a \in \mathcal{A}$ *we have*

$$\mathbb{P}(A = a|X_1, \ldots, X_n) = \prod_{i<j} \kappa(X_i, X_j)^{a_{ij}} (1 - \kappa(X_i, X_j))^{1-a_{ij}}.$$

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ *for some distribution* F *on* $\mathcal{X}$*, then the model is denoted by* $\text{LPM}(\mathcal{X}, F, \kappa)$*.*

Among all the LPMs, we are particularly interested in the random dot product graph (RDPG) [**??**].

Definition 2.2.3 (Random Dot Product Graph) *Let the latent position for vertex* i *be a random variable* $X_i : \Omega \mapsto \mathcal{X} \subset \mathbb{R}^d$*, where* $i \in [n]$ *and* $\mathcal{X}$ *is the latent space such that* $x^\top y = \sum_{i=1}^d x_i y_i \in [0, 1]$ *for* $x, y \in \mathcal{X}$*. Then under a* random dot product graph (RDPG)*, for* $i, j \in [n]$ *and* $i < j$*, conditioned on their respective latent positions* X_i *and* X_j*,*

$$A_{ij}|(X_i, X_j) \overset{ind}{\sim} \text{Bernoulli}(X_i^\top X_j).$$

Thus for $a \in \mathcal{A}$ we have

$$\mathbb{P}(A = a | X_1, \ldots, X_n) = \prod_{i<j} (X_i^\top X_j)^{a_{ij}} (1 - X_i^\top X_j)^{1-a_{ij}}.$$

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ for some distribution F on $\mathcal{X}$, then the model is denoted by $\mathrm{RDPG}(\mathcal{X}, F)$. *Typically, we write all latent positions together as $X = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d}$, where the i-th row $X_i^\top$ corresponds to the latent position for vertex i.*

From the definition above, we see RDPG is actually a special case of LPM with latent space $\mathcal{X} \subset \mathbb{R}^d$ such that $x^\top y \in [0, 1]$ for $x, y \in \mathcal{X}$ and link function $\kappa(x, y) = x^\top y$. If d is much smaller than the number of vertices n, which is likely to be the case in practice, RDPG is then a more parsimonious model compared to IEM, requiring only $n \cdot d$ parameters rather than $\binom{n}{2}$.

The direction and magnitude of the latent position, which are determined by properties of the corresponding vertex, are the most important factors in RDPG due to the nature of dot product. Vertices with latent positions pointing in similar directions are more likely to have an edge between them compared to those with different directions. Similarly, the magnitude of the latent positions encodes the vertices' overall tendency to form edges. A larger magnitude potentially leads to more edges incident with the vertex.

Conditioned on the latent positions X, the RDPG now can be considered to be an $\mathrm{IEM}(P)$ with $P = XX^\top$, i.e. an edge between vertex i and vertex j is present with

probability $P_{ij} = X_i^\top X_j$. Note that the probability matrix p is the outer product of the latent position matrix X with itself. This imposes two important properties on P under RDPG, namely that P is positive semidefinite (PSD) and $\text{rank}(P) = \text{rank}(X) \leq d$. On the other hand, this also suggests that for certain circumstances when the probability matrix P might not be positive semi-definite, one may want to use some other LPM which preserves the low-rank property instead.

Example 2.2.4 *Consider a LPM with link function $\kappa(x, y) = x^\top K y$ where $K \in \{-1, 0, 1\}^{d \times d}$ is a diagonal matrix with $K_{ii} = 1$ for $i \leq d'$ and $K_{ii} = -1$ otherwise for $1 < d' < d$. Unlike RDPG, this LPM can be applied to the situation where the low-rank probability matrix P is indefinite. For example $P \in [0, 1]^{n \times n}$ with d non-zero eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d'} > 0 > \lambda_{d'+1} \geq \cdots \geq \lambda_d$.*

The LPM in Example 2.2.4 above is a natural extension of RDPG to the non-PSD case. We will see later that it motivates another estimator analogous to the one we are going to propose. Details are discussed in Remark **??**. RT: TODO

As one may notice, RDPG is non-identifiable. For any orthonormal matrix $W \in \mathbb{R}^{d \times d}$, the rotated latent positions XW is equivalent to the original X since $P = XX^\top = (XW)(XW)^\top$. In later sections, we will sometimes consider the equivalent class of latent positions up to rotations instead.

Importantly, since P is the outer product of the latent positions X in RDPG, it also motivates the low-rank estimator based on spectral decomposition which will be discussed in details later.

2.2.3 Stochastic Blockmodel

One of the most important structures for graphs is the community structure in which vertices are clustered into different communities such that vertices of the same community behave similarly. This structural property is captured by the stochastic blockmodel (SBM) [Holland et al., 1983], where each vertex is assigned to a block and the probability that an edge exists between two vertices depends only on their respective block memberships.

Definition 2.2.5 (Stochastic Blockmodel) *Consider a K-block* stochastic blockmodel (SBM) *with block probability matrix $B \in [0,1]^{K\times K}$, and the vector of block memberships $\tau \in [K]^n$, where for each $i \in [n]$, $\tau_i = k$ means vertex i is a member of block k. We have for $i, j \in [n]$ and $i < j$,*

$$A_{ij} \overset{ind}{\sim} \text{Bernoulli}(B_{\tau_i,\tau_j}).$$

Thus for $a \in \mathcal{A}$ we have

$$\mathbb{P}(A = a) = \prod_{i<j} B_{\tau_i,\tau_j}^{a_{ij}}(1 - B_{\tau_i,\tau_j})^{1-a_{ij}}.$$

Such model is denoted by $\text{SBM}(\tau, B)$.

In some scenarios, the block probability matrix B and block membership vector τ in SBM are assumed to be random. For example, each vertex can be assigned to

a block independently according to a probability vector $\rho \in (0,1)^K$ with $\sum_{k=1}^K \rho_k = 1$ such that $\mathbb{P}(\tau_i = k) = \rho_k$. Nevertheless, the definition above still holds when conditioning on B and τ. And then the SBM now can be considered to be an IEM(P) with $P = B_{\tau,\tau}$, i.e. an edge between vertex i and vertex j is present with probability $P_{ij} = B_{\tau_i,\tau_j}$.

Example 2.2.6 *Consider a 5-block SBM with*

$$B = \begin{bmatrix} 0.90 & 0.27 & 0.05 & 0.10 & 0.30 \\ 0.27 & 0.67 & 0.02 & 0.26 & 0.14 \\ 0.05 & 0.02 & 0.44 & 0.25 & 0.33 \\ 0.10 & 0.26 & 0.25 & 0.70 & 0.18 \\ 0.30 & 0.14 & 0.33 & 0.18 & 0.58 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.22 & 0.39 & 0.05 & 0.16 & 0.18 \end{bmatrix}.$$

We sample a graph with $n = 200$ vertices under this SBM and plot the corresponding probability matrix $P = B_{\tau,\tau}$ and the adjacency matrix A in Figure 2.1. While A is a noisy version of P, the structure of 25 blocks can be seen clearly in both figures as a result of 5 different blocks among vertices.

As discussed in Section 2.2.2, the probability matrix P in an RDPG is positive semidefinite. And now we argue that an SBM with a positive semidefinite B can always be parameterized as an RDPG. Firstly due to the positive semidefiniteness of B, we can decompose $B = \nu\nu^\top$ where $\nu \in \mathbb{R}^{K\times d}$. Define $\nu_1, \ldots, \nu_K$ such that the rows

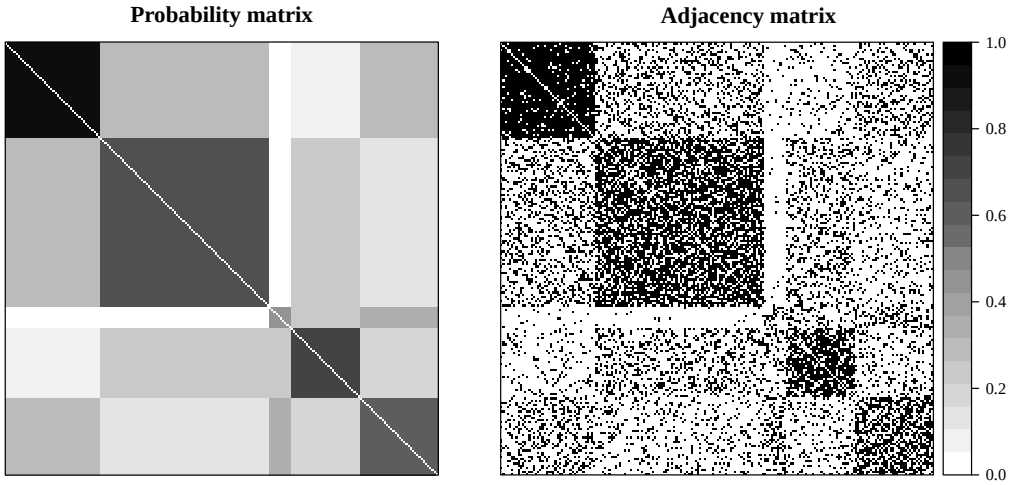


Figure 2.1: Example illustrating the stochastic blockmodel. The parameters are given in Example 2.2.6. The left figure shows the probability matrix P with $K = 5$ blocks and $n = 200$ vertices and the right figure shows an adjacency matrix A sampled under SBM(P). While A is a noisy version of P, much of the structure of P is preserved in A, a property we will exploit in our estimation procedure.

of ν are given by $\nu_1^\top, \ldots, \nu_K^\top$. Then ν_k can be regarded as the shared latent position for all vertices assigned to block k. Define $X = [X_1, \ldots, X_n]^\top = [\nu_{\tau_1}, \ldots, \nu_{\tau_n}]^\top \in \mathbb{R}^{n \times d}$. Then the SBM now can be parameterized as an RDPG since

$$\mathbb{P}(A_{ij} = 1) = B_{\tau_i, \tau_j} = \nu_{\tau_i}^\top \nu_{\tau_j} = X_i^\top X_j \in [0, 1].$$

Definition 2.2.7 (SBM as RDPG) *Consider an* RDPG($\mathcal{X}, F$) *where F is a distribution on K point masses $\nu_1, \ldots, \nu_K \in \mathcal{X}$ such that $\mathbb{P}(X_i = \nu_k) = \rho_k$ for $i \in [n]$ and $k \in [K]$, where the block proportion vector $\rho \in (0, 1)^K$ with $\sum_{k=1}^K \rho_k = 1$. Then we denote this model by* SBM(ρ, B) *where $B = \nu^\top \nu$ with $\nu = [\nu_1, \ldots, \nu_K]^\top$. Similarly, we can define $\tau \in [K]^n$ such that $X_i = \nu_{\tau_i}$.*

For notational convenience we will refer to the sub-model of SBM with positive semidefinite B as *SBM*. As shown above, the SBM can be regard as an RDPG where all vertices in the same block have identical latent positions. In this work, we will always analyze SBM in an RDPG setting.

Remark 2.2.8 *As we mentioned, under the SBM, all vertices in the same block have identical latent positions. Rather than allowing vertices differ from each other as RDPG, SBM presumes all nodes within the same block have the same expected degree. To better describe complex networks in some situations, a bunch of generalizations of the SBM have been explored in order to incorporate the local variation of vertices to the block structure. Airoldi et al. [2008] proposed mixed membership stochastic blockmodels, which associates each vertex with multiple blocks with a probability vector rather than a single block as SBM requires. Also, in order to model variation of the expected degrees of different vertices within the same block, Karrer and Newman [2011] proposed degree-corrected SBM, which assigns additional parameters to each vertex to adjust the expected degree relatively. All these generalizations are trying to drag SBM towards RDPG a little bit so that the model can capture variations among vertices while keeping the community structure.*

2.3 Weighted Random Graph Models

In this section, we shift our focus to the case where graphs are weighted and generalize the three models introduced in Section 2.2 respectively, i.e. the weighted independent edge model (WIEM) in Section 2.3.1, the weighted random dot product graph model (WRDPG) in Section 2.3.2, and the weighted stochastic blockmodel (WSBM) as a WRDPG in Section 2.3.3.

As a reminder, for weighted graphs, each edge is assigned with a positive real-valued weight, i.e. $\mathcal{A} = \mathbb{R}_{\geq 0}^{n \times n}$. So the adjacency matrices are not binary any more.

2.3.1 Weighted Independent Edge Model

As introduced in Section 2.2.1, under IEM(P), for $i, j \in [n]$ and $i < j$, the edge weight A_{ij} is distributed from a Bernoulli distribution with parameter P_{ij} independent of other edges. We first extend IEM to weighted independent edge model (WIEM).

Definition 2.3.1 (Weighted Independent Edge Model) *Consider a one-parameter family of distributions $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$. Let the graph parameters be a matrix $P \in \Theta^{n \times n} \subset \mathbb{R}^{n \times n}$. Then under a* weighted independent edge model (WIEM) *with respect to $\mathcal{F}$, for $i, j \in [n]$ and $i < j$, the edge weight between vertex i and vertex j is distributed according to $f_{P_{ij}}$ independently.*

To see that an IEM is a special case of WIEM, let $\mathcal{F}$ be the collection of Bernoulli distributions and let the graph parameters be a symmetric and hollow matrix $P \in$

$[0,1]^{n\times n}$. Note that the graphs considered in this paper are undirected without self-loops, thus the parameter matrix P needs to be symmetric and hollow. However, for convenience, we still define the parameters to be an n-by-n matrix while only $\binom{n}{2}$ of them are effective.

2.3.2 Weighted Random Dot Product Graph

As discussed in Section 2.2.2, the connectivity between two vertices in a graph generally depends on some hidden properties of the corresponding vertices. Such property is well captured by LPM as well as RDPG, which is a special case of LPM. In this section, we generalize RDPG to a weighted version so that it can model weighted graphs.

Definition 2.3.2 (Weighted Random Dot Product Graph) *Consider a collection of one-parameter distributions $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$. The* weighted random dot product graph (WRDPG) *with respect to $\mathcal{F}$ is defined as following: Let $X \in \mathbb{R}^{n\times d}$ be such that $X = [X_1, X_2, \ldots, X_n]^\top$, where $X_i \in \mathbb{R}^d$ for all $i \in [n]$. The matrix X is random and satisfies $\mathbb{P}(X_i^\top X_j \in \Theta) = 1$ for all $i, j \in [n]$. Conditioned on X, A_{ij} follows distribution $f_\theta \in \mathcal{F}$ with parameter $\theta = X_i^\top X_j$ independent of others for $i, j \in [n]$ and $i < j$.*

Under the WRDPG defined above, the parameter matrix $P = XX^T \in \mathbb{R}^{n\times n}$ is automatically symmetric because the link function is inner product. Moreover, to

have symmetric graphs without self-loops, only A_{ij} $(i < j)$ are sampled while leaving the diagonals of A to be all zeros.

After such extension, WRDPG inherits some properties from RDPG naturally, for example the positive defininteness of P, non-identifiability, etc. And similarly, since P is still the outer product of the latent positions X in WRDPG, it also motivates the low-rank estimator based on spectral decomposition for weighted graphs which will be discussed in later sections.

2.3.3 Weighted Stochastic Blockmodel

Definition 2.3.3 (Weighted Stochastic Blockmodel) *Consider a collection of one-parameter distributions $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$. A K-block* weighted stochastic blockmodel (WSBM) *with respect to $\mathcal{F}$ is defined as following: Let block probability matrix be $B \in \Theta^{K\times K}$, and let the vector of block memberships be $\tau \in [K]^n$, where for each $i \in [n]$, $\tau_i = k$ means vertex i is a member of block k. Then A_{ij} follows distribution $f_\theta \in \mathcal{F}$ with parameter $\theta = B_{ij}$ independent of others for $i, j \in [n]$ and $i < j$.*

WSBM can be defined in a similar way in the scenario when the block probability matrix B and block membership vector τ are random. As mentioned in Section 2.2.3, all analysis for unweighted graphs is based on RDPG setting. Likewise, in this section we will represent WSBM as WRDPG. Because of the structure of WRDPG, in order to consider WSBM as a WRDPG, the block probability matrix B needs to be positive

semidefinite. From now on, we will denote the sub-model of WSBM with positive semi-definite B as the WSBM.

Definition 2.3.4 (WSBM as WRDPG) *Consider a WRDPG with respect to $\mathcal{F}$ where each latent position X_i can take one of the K possible values $\nu_1, \ldots, \nu_K$ such that $\mathbb{P}(X_i = \nu_k) = \rho_k$ for $i \in [n]$ and $k \in [K]$, where the block proportion vector $\rho \in (0,1)^K$ with $\sum_{k=1}^{K} \rho_k = 1$. Then this is a WSBM with respect to $\mathcal{F}$ where $B = \nu^\top \nu$ with $\nu = [\nu_1, \ldots, \nu_K]^\top$. Similarly, we can define $\tau \in [K]^n$ such that $X_i = \nu_{\tau_i}$.*

As shown above, the WSBM can be regard as a WRDPG where all vertices in the same block have identical latent positions. In later sections when considering weighted graphs, we will always analyze WSBM in a WRDPG setting.

Chapter 3

A Law of Large Graphs

Estimation of the mean of a population based on samples is at the core of statistics. The sample mean, motivated by the law of large numbers and the central limit theorem, has its place as one of the most important statistics for this task. In modern settings, we take averages almost everywhere, from data in Euclidean space to more complex objects like images, shapes, and documents. In this chapter we consider the challenges of estimating a population mean based on a sample of graphs, for example the human brains as represented by their structural connectomes.

The mean of a population of graphs is a high dimensional object, consisting of $O(n^2)$ parameters for graphs with n vertices. When the number of samples m is much smaller than n^2, or even n, estimating such high dimensional estimands using naive unbiased methods often leads to inaccurate estimates with very high variance. Furthermore, using these estimates for subsequent inference tasks such as testing can

lead to low power and accuracy. By exploiting a bias-variance trade-off, it is often fruitful to develop estimators which have some bias but greatly reduced variance. When these estimators are biased towards low-dimensional structures which well approximate the full dimensional population mean, major improvements can be realized [Trunk, 1979].

In a striking result, Stein [1956] and James and Stein [1961] showed that even the arithmetic mean can be dominated by another procedure. In particular, James and Stein showed that the sample mean for a multivariate normal distribution with at least three dimensions has strictly higher risk than a procedure that introduces shrinkage, and can be strictly improved by carefully biasing the estimate towards any given fixed point. Twenty-seven years later, Gutmann [1982] proved that this phenomenon cannot occur when the sample spaces are finite, as is the case for graphs. However, while there must be some cases where the sample mean is preferred, this does not mean that other estimators should not be considered. In many situations where other structural information is hypothesized, other estimators may be preferable.

In complex data settings such as shape data, language data, or graph data, we also must take care in how we define the mean. For a population, we define the mean graph as the weighted adjacency matrix with weights given by the proportion of times the corresponding edge appears in the population. This definition naturally extends the definition of the mean for standard Euclidean data. As with real valued data, one may want to define the mean of a population of graphs to be a graph. This is

captured in the notion of the median graph [Jiang et al., 2001], however, this may be too restrictive for populations of graphs where there is high variation in which edges appear. As we will describe below, our definition of the mean graph is the expectation of the adjacency matrix.

This population mean is becoming an important object for statistical inference. For example, Ginestet et al. [2014] proposed a way to test if there is a difference between the distributions for two groups of networks. While hypothesis testing is the end goal of their work, estimation is a key intermediate step which may be improved by accounting for underlying structure in the mean matrix. Thus, improving the estimation procedures for the mean graph is not only important by itself, but also can be applied to help improve other statistical inference procedures.

To better illustrate the idea, we take the CoRR brain graphs with Desikan atlases as an example. The dataset contains 454 brain scans with 70 vertices. Each vertex represents a region defined by the Desikan atlases, while an edge exists between two vertices if there is at least one white-matter tract connecting the corresponding regions of the brain. More details about this dataset are given in Section **??**. By observing m graphs sampled from the 454 graphs, our goal is to estimate the mean graph of the population P, defined as the entry-wise mean of all the 454 graphs. We plot the population mean graph P on the left panel in Figure 3.1.

The element-wise sample mean is a reasonable estimator if we consider the general independent edge model (IEM) [Bollobás et al., 2007] introduced in Section 2.2.1

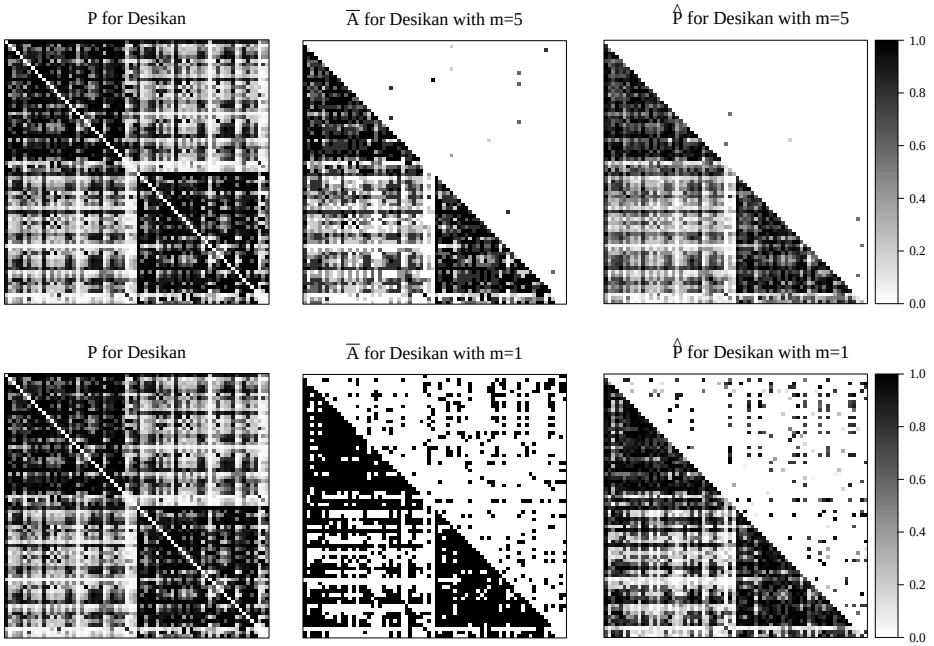


Figure 3.1: Heat maps of the population mean P, the sample mean $\bar{A}$, and the estimator $\widehat{P}$ based on sample sizes $m = 1$ and $m = 5$. The three heat maps in the upper level indicate the population mean for the 454 graphs (left), sample mean for the 5 sampled graphs (center), and $\widehat{P}$ for the same 5 sampled graphs with dimension $d = 11$ selected using the Zhu and Ghodsi method (right). Details about how to construct $\widehat{P}$ are discussed in Section 3.3.2. Darker pixels indicate a higher probability of an edge between the given vertices. By calculating the mean squared error based on this sample, we can see that $\widehat{P}$ (with mean squared error equals 0.015) outperforms $\bar{A}$ (with mean squared error equals 0.016), with a 3% relative improvement. In order to see where the improvements are clearly, in the upper triangular of the heat maps for $\bar{A}$ and $\widehat{P}$, we highlight the edges (18 edges highlighted for $\bar{A}$ and 6 for $\widehat{P}$) which have absolute estimation error larger than 0.4. In the lower level, we plot three heat maps in the similar way based on sample size $m = 1$. For this specific sampled graph, $\widehat{P}$ is calculated with dimension $d = 12$. A clearly smoothing effect can be seen in the heat map of $\widehat{P}$ (with mean squared error equals 0.049), which leads to a 53% relative improvement compared to $\bar{A}$ (with mean squared error equals 0.104). Similarly, we use the same absolute estimation error threshold 0.4 and highlight 504 edges for $\bar{A}$ and 234 edges for $\widehat{P}$.

without taking any additional structure into account. However, with only a small sample size, such as when the sample size is much less than the number of vertices, it does not perform very well. Now take a sample of size $m = 5$ in the CoRR dataset example with Desikan atlases mentioned above, then we calculate the entry-wise sample mean $\bar{A}$ and plot it in the top middle panel of Figure 3.1. Darker pixels indicate a higher probability of an edge between the given vertices. We can see $\bar{A}$ gives a fair estimate of P. However, there are still an amount of edges being estimated very inaccurately. In order to see these inaccurate estimates clearly, in the upper triangular of the heat maps for $\bar{A}$, we highlight the 18 edges which have absolute estimation error larger than 0.4 based on the same color scale. When the sample size is small, the performance of $\bar{A}$ degrades due to its high variances. Such phenomenon is most obvious when we decrease the sample size from $m = 5$ to $m = 1$. In the bottome middle panel of Figure 3.1, we plot the heat map of $\bar{A}$ based on sample size $m = 1$. Since there is only one observed graph, $\bar{A}$ is binary and thus very bumpy. Similarly, we use the same absolute estimation error threshold 0.4 and highlight 504 edges in the upper triangular. Intuitively, an estimator incorporating structure in the distribution of graphs, assuming the estimator is computationally tractable, is preferable to the entry-wise sample mean. In general, we do not have any knowledge about this structure so it can be hard to take advantage of in practice.

One of the most important structures in graphs is the community structure in which vertices are clustered into groups that share similar connectivity structure.

The stochastic blockmodel (SBM) [Holland et al., 1983] introduced in Section 2.2.3 is one model that captures this structural property and is widely used in modeling networks. From population mean P plotted in Figure 3.1, we can see the brain is a 2-block model at the highest level, representing the two hemispheres. More generally, the latent positions model (LPM) [**?**] introduced in Section 2.2.2, provides a way to parameterize the graph structure by latent positions associated with each vertex. Latent position models can capture strong community structure like the stochastic blockmodel, but may also allow for more variance within communities and other structures. One example of an LPM which captures this middle ground is the random dot product graph (RDPG) [**??**] introduced in Section 2.2.2 which motivates our estimator. It generalizes the positive semidefinite SBM by allowing for mixed membership and degree corrections.

Using estimates of the latent positions based on a truncated eigen-decomposition of the adjacency matrix, we propose an estimator which captures the low-rank structure of the mean graph for the RDPG model. Details about this estimator are discussed in Section 3.3.2. These estimates will improve performance since they will be biased towards the low-rank structure of the RDPG model and will have much lower overall variance than naive element-wise sample means. Here we consider the same random sample of size $m = 5$ based on the Desikan atlas in Figure 3.1 and plot the estimate $\widehat{P}$ in the top right panel. Note that compared to the sample mean $\bar{A}$ (with mean squared error equals 0.016), $\widehat{P}$ (with mean squared error equals 0.015) has a

finer gradient of values which in this case leads to a 3% relative improvement in estimation of the true probability matrix P. In order to see where the improvements are clearly, in the upper triangular of the heat map for $\widehat{P}$, we also highlight the 6 edges which have absolute estimation error larger than 0.4, where 18 edges are highlighted for $\bar{A}$ based on the same threshold. The smoothing effect is much more obvious when we decrease the sample size from $m = 5$ to $m = 1$. In the lower level of Figure3.1, we plot the heat map of $\widehat{P}$ based on sample size $m = 1$. From the figure, we can see that $\widehat{P}$ smooths the estimate, especially for edges across the two hemispheres, in the lower left and corresponding upper right block (which is not shown in the heat map). Based on the calculations, $\widehat{P}$ (with mean squared error equals 0.049) outperforms $\bar{A}$ (with mean squared error equals 0.104), with a 53% relative improvement. Similarly, we use the same absolute estimation error threshold 0.4 and highlight the 234 edges for $\widehat{P}$.

In this chapter, we show via theory, simulations, and real data analysis that the low-rank estimator frequently outperforms the element-wise sample mean, especially in small sample sizes.

In Section 3.1, we outline the model which we consider for our theorems and simulations, and in Section 3.3 we describe the entry-wise sample mean and introduce our specific low-rank estimator, which accounts for the unknown dimension and attempts to correct for other issues found in real world problems. Our main theoretical results are presented in Section 3.4. And then we present simulations results for the stochas-

tic blockmodel in Section 3.5, an investigation of a connectome dataset in Section 3.6, and a synthetic data analysis in Section 3.7.

3.1 Model

This chapter considers the scenario of having m unweighted graphs, represented as adjacency matrices, $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, each having n vertices with $A^{(t)} \in \{0,1\}^{n \times n}$ for $t \in [m]$. We assume there is a known correspondence for vertices across different graphs, so that vertex i in graph t corresponds to vertex i in graph t' for any $i \in [n]$, $t, t' \in [m]$. The graphs we consider are undirected and unweighted with no self-loops, so each $A^{(t)}$ is a binary symmetric matrix with zeros along the diagonal.

For the purpose of this paper, we also assume that the graphs are sampled independently and identically from some distribution. To this end, the mean graph we are trying to estimate is the expectation of each adjacency matrix.

Definition 3.1.1 (Mean Graph) *Suppose that* $A^{(1)}, \ldots, A^{(m)} \overset{iid}{\sim} \mathcal{G}$ *for some random graph distribution* $\mathcal{G}$*, with* $A^{(t)} \in \{0,1\}^{n \times n}$ *for* $t \in [m]$*. The* mean graph *is defined as* $\mathbb{E}[A^{(1)}]$*, where since the graphs are identically distributed* $\mathbb{E}[A^{(t)}] = \mathbb{E}[A^{(t')}]$ *for* $t, t' \in [m]$*.*

In this chapter, we consider the scenario that all m graphs follow the same SBM. Since the vertex correspondence is assumed across graphs, the block memberships τ_i are firstly drawn iid from a categorical distribution with block membership prob-

abilities given by $\rho \in [0,1]^K$ and this will keep the same for all m graphs to be sampled. Denote block probability matrix $B = \nu\nu^\top \in [0,1]^{K\times K}$. By Definition 3.1.1, the mean of the collection of graphs generated from this SBM is $P \in [0,1]^{n\times n}$, where $P_{ij} = B_{\tau_i,\tau_j}$. Then m graphs on n vertices $A^{(1)}, \cdots, A^{(m)}$ are sampled independently from the SBM conditioned on τ.

3.2 Methods

Before we start introducing our estimators for P, in this section we focus on several methods which are key components for constructing the estimators later.

3.2.1 Adjacency Spectral Embedding

We first introduce the adjacency spectral embedding (ASE), which is our most important tool for exploiting the low-rank property.

Definition 3.2.1 (Adjacency Spectral Embedding) *For a symmetric n-by-n matrix A, let its eigen-decomposition be $\widehat{U}\widehat{S}\widehat{U}^\top + \widetilde{U}\widetilde{S}\widetilde{U}^\top$, where $\widehat{S}$ is a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest d eigenvalues of A, and $\widehat{U}$ has columns given by the corresponding eigenvectors. Similarly, $\widetilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n-d$ eigenvalues of A, and $\widetilde{U}$ has the columns given by the corresponding eigenvectors. Then the d-dimensional* adjacency spectral embedding (ASE) *of A is*

defined as $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n \times d}$.

Consider the probability matrix P in an RDPG setting with latent positions $X \in \mathbb{R}^{n \times d}$, i.e. $P = XX^\top$. Then the d-dimensional ASE of P exactly recovers its latent positions X. Moreover, Sussman et al. [2014] showed that the ASE of the adjacency matrix A under RDPG gives good estimates of the latent vectors for each vertex under proper conditions.

3.2.2 Choosing Dimension

Often in dimensionality reduction techniques, the choice for dimension d, relies on analyzing the set of the ordered eigenvalues, looking for a "gap" or "elbow" in the scree-plot. Zhu and Ghodsi [2006] present an automated method for finding this gap in the scree-plot that takes only the ordered eigenvalues as an input and uses Gaussian mixture modeling to find these gaps. The mixture modeling results in multiple candidate dimensions or elbows, and our analysis indicated that underestimating the dimension is much more harmful than overestimating the dimension. For this reason, we used the 3rd elbow in the experiments performed for this work.

Universal Singular Value Thresholding (USVT) is a simple estimation procedure proposed in Chatterjee [2015] that can work for any matrix that has "a little bit of structure". In our setting, it selects the dimension d as the number of singular values that are greater than a constant c times $\sqrt{n/m}$. The specific constant c must be

selected carefully based on the mean and variance of the entries, and since again we found that overestimating the dimension was not overly harmful, we chose a relatively small value of $c = 0.7$.

Overall, selecting the appropriate dimension is a challenging task and numerous methods could be applied successfully depending on the setting. On the other hand, we have observed that in our setting, many dimensions will yield nearly optimal mean squared errors. Thus efforts to ensure the selected dimension is in the appropriate range are more important than finding the best dimension.

3.2.3 Graph Diagonal Augmentation

The graphs examined in this work have no self-loops and thus the diagonal entries of the adjacency matrix and the mean graph are all zero. However, when computing the low-rank approximation, these structural zeros lead to increased errors in the estimation of the mean graph. While this problem has been investigated in the single graph setting, with multiple graphs, the problem is exacerbated since the variance of the other entries is lower, so the relative impact of the bias in the diagonal entries is higher. Moreover, the sum of eigenvalues of the hollow matrix will be zero, leading to an indefinite matrix, which violates the positive semi-definite assumption. So it is important to remedy the situation that we don't observe the diagonal entries.

Marchette et al. [2011] proposed the simple method of imputing the diagonals to be equal to the average of the non-diagonal entries for the corresponding row. Earlier,

Scheinerman and Tucker [2010] proposed using an iterative method to impute the diagonal entries. In this work, we combine these two ideas by first using the row-average method (see Step 3 of Algorithm 1) and then using one step of the iterative method (see Step 6 of Algorithm 1), which will be discussed in Section 3.3. Note that when computing errors, we omit the diagonal entries since these are known to be zero.

3.3 Estimators

In this section, we present two estimators, the standard element-wise sample mean $\bar{A}$, and a low-rank estimator $\widehat{P}$. We describe the low-rank aspects of this estimator as well as further important details regarding diagonal augmentation and dimension estimation in this section.

3.3.1 Element-wise sample mean $\bar{A}$

The most natural estimator to consider is to take the average of the observed adjacency matrices which yields the element-wise sample mean. This estimator, defined as $\bar{A} = \frac{1}{m}\sum_{t=1}^{m} A^{(t)}$, is the maximum likelihood estimator (MLE) for the mean graph P if the graphs are sampled from an IEM distribution. It is unbiased so $\mathbb{E}[\bar{A}] = P$ with entry-wise variance $\mathrm{Var}(\bar{A}_{ij}) = P_{ij}(1 - P_{ij})/m$. Moreover, $\bar{A}$ is the uniformly minimum-variance unbiased estimator, so it has the smallest variance among all un-

biased estimators and enjoys the many asymptotic properties of the MLE as $m \to \infty$ for fixed n. However, if graphs with a large number of vertices are of interest, there are no useful asymptotic properties for $\bar{A}$ as the number of vertices n becomes large for fixed m.

Additionally, $\bar{A}$ doesn't exploit any graph structure. If the graphs are distributed according to an RDPG or SBM, then $\bar{A}$ is no longer the maximum likelihood estimator since it is not guaranteed to satisfy the properties of the mean graph for that model. The performance can be especially poor when the sample size m is small, such as when $m \ll n$. For example, when $m = 1$, $\bar{A}$ is simply the binary adjacency matrix $A^{(1)}$, which is an inaccurate estimate for an arbitrary P compared to estimates which exploit underlying structure, such as occurs for the RDPG.

3.3.2 Low-Rank Estimator $\widehat{P}$

Motivated by the low-rank structure of the RDPG mean matrix, we propose the estimator $\widehat{P}$ based on the spectral decomposition of $\bar{A}$ which yields a low rank approximation of $\bar{A}$. This estimator is similar to the estimator proposed by Chatterjee [2015] but additionally we propose adjustments to canonical low-rank methods which serve to improve the performance for the specific task of estimating the mean graph. Additionally, we consider an alternative dimension selection technique as discussed in Section 3.2.2. To summarize, our overall strategy to computer $\widehat{P}$ is described in Algorithm 1. The key component of this algorithm is the low-rank estimator. Details

of this vital step to compute the actual low-rank approximation is in Algorithm 2.

Algorithm 1 Algorithm to compute $\widehat{P}$

Require: Adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \{0,1\}^{n \times n}$

Ensure: Estimate $\widehat{P} \in [0,1]^{n \times n}$

1: Calculate the sample mean $\bar{A} = \frac{1}{m} \sum_{t=1}^{m} A^{(t)}$;

2: Calculate the scaled degree matrix $D^{(0)} = \text{diag}(\bar{A}\mathbf{1})/(n-1)$;

3: Select the dimension d based on the eigenvalues of $\bar{A} + D^{(0)}$; (see Section 3.2.2)

4: Set $\widetilde{P}^{(0)}$ to $\text{lowrank}_d(\bar{A} + D^{(0)})$; (see Algorithm 2)

5: Set $D^{(1)}$ to $\text{diag}(\widetilde{P}^{(0)})$, the diagonal matrix with diagonal matching $\widetilde{P}^{(0)}$;

6: Set $\widetilde{P}^{(1)}$ to $\text{lowrank}_d(\bar{A} + D^{(1)})$; (see Algorithm 2)

7: Set $\widehat{P}$ to $\widetilde{P}^{(1)}$ with values < 0 set to 0 and values > 1 set to 1.

For a given dimension d we consider the estimator $\text{lowrank}_d(\bar{A})$ defined as the best rank-d positive semidefinite approximation of $\bar{A}$. Since the graphs are symmetric, we can compute the eigen-decomposition of $\bar{A}$ as $\widehat{U}\widehat{S}\widehat{U}^\top + \widetilde{U}\widetilde{S}\widetilde{U}^\top$, where $\widehat{S}$ is a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest d eigenvalues of $\bar{A}$, and $\widehat{U}$ has columns given by the corresponding eigenvectors. Similarly, $\widetilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n-d$ eigenvalues of $\bar{A}$, and $\widetilde{U}$ has the columns given by the corresponding eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\bar{A}$ is given by $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n \times d}$. For an RDPG, the rows of $\widehat{X}$ are estimates of the latent vectors for each vertex [Sussman et al., 2014]. Using the adjacency

spectral embedding, we have the low-rank approximation of $\bar{A}$ to be $\widehat{X}\widehat{X}^\top = \widehat{U}\widehat{S}\widehat{U}^\top$. Algorithm 2 gives the steps to compute this low-rank approximation for a general symmetric matrix A.

Algorithm 2 Algorithm to compute the rank-d approximation of a matrix.

Require: Symmetric matrix $A \in \mathbb{R}^{n\times n}$ and dimension $d \leq n$.

Ensure: $\text{lowrank}_d(A) \in \mathbb{R}^{n\times n}$

1: Compute the algebraically largest d eigenvalues of A, $s_1 \geq s_2 \geq \ldots \geq s_d$ and corresponding unit-norm eigenvectors $u_1, u_2, \ldots, u_d \in \mathbb{R}^n$;

2: Set $\widehat{S}$ to the $d \times d$ diagonal matrix $\text{diag}(s_1, \ldots, s_d)$;

3: Set $\widehat{U} = [u_1, \ldots, u_d] \in \mathbb{R}^{n\times d}$;

4: Set $\text{lowrank}_d(A)$ to $\widehat{U}\widehat{S}\widehat{U}^\top$;

To compute our estimator $\widehat{P}$, we also need to specify what rank d to use and there are various ways of dealing with dimension selection. In this work, we use an elbow selection method proposed in Zhu and Ghodsi [2006] and the universal singular value thresholding (USVT) method [Chatterjee, 2015]. Details for these methods are discussed in Section 3.2.2.

Moreover, since the adjacency matrices are hollow, with zeros along the diagonal, there is a missing data problem that leads to inaccuracies if we compute $\widehat{P}$ based only on $\bar{A}$. To compensate for this issue, we use an iterative method developed in Scheinerman and Tucker [2010]. Details are discussed in Section 3.2.3.

Algorithm 1 gives the steps involved to compute the low-rank estimate $\widehat{P}$. For

convenience, here we consider Example 2.2.6 again. In Figure 3.2, the upper left figure shows the probability matrix P with $K = 5$ blocks and $n = 200$ vertices and the upper right figure shows an adjacency matrix A sampled under SBM(P), which repeats Figure 2.1. The bottom panels of Figure 3.2 demonstrate the two estimators $\widehat{P}$ and $\bar{A}$ for the stochastic blockmodel given by the upper left panel. The estimates are based on a sample of size $m = 3$ and in this instance visual inspection demonstrates that $\widehat{P}$ performs much better than $\bar{A}$. As we will see in the succeeding sections, this procedure will frequently yield improvements in estimation as compared to using the sample mean $\bar{A}$. While this is unsurprising for random dot product graphs, where we are able to show theoretical results to this effect, we also see this effect for connectome data and more general independent edge graphs. In the following sections, we explore this estimator in the context of the stochastic blockmodel discussed in Section 3.1.

3.4 Theoretical Results

To estimate the mean of a collection of graphs, we consider the two estimators from Section 3.3: the entry-wise sample mean $\bar{A}$ and the low-rank $\widehat{P}$ motivated by RDPG. We evaluate our estimators in terms of mean squared error, either $\text{MSE}(\widehat{P}_{ij}) = \mathbb{E}[\widehat{P}_{ij} - P_{ij}]^2$ or $\text{MSE}(\bar{A}) = \mathbb{E}[\bar{A}_{ij} - P_{ij}]^2$. While we can directly compare the difference in mean squared errors between the two estimators, it is frequently useful to consider the relative efficiency between two estimators.

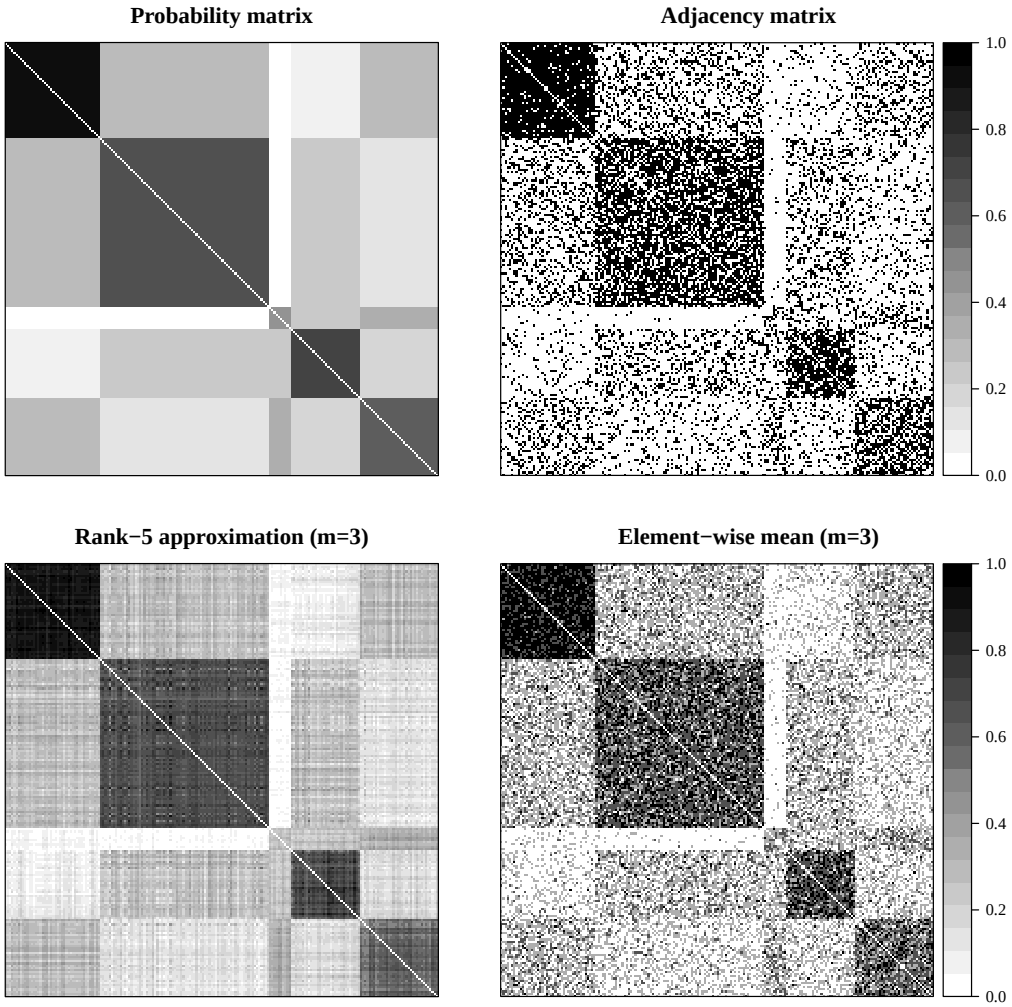


Figure 3.2: Example illustrating different estimates under the stochastic block-model. The top left figure shows the mean graph P with $K = 5$ blocks and $n = 200$ vertices and the top right figure shows an adjacency matrix A sampled according to the probabilities from P. While A is a noisy version of P, much of the structure of P is preserved in A, a property we will exploit in our estimation procedure. Based on three graphs sampled independently and identically according to the probability matrix P, we construct the element-wise mean $\bar{A}$, shown in the lower right panel (see Section 3.3.1). Finally, by taking a rank-5 approximation of $\bar{A}$ and thresholding the values to be between 0 and 1, we construct our proposed estimate $\widehat{P}$, shown in the lower left panel (see Section 3.3.2). By visual inspection, it is clear that the low-rank estimate $\widehat{P}$ more closely approximates the probability matrix P as compared to $\bar{A}$.

Definition 3.4.1 (Relative Efficiency) *For two estimators* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$*, the* relative efficiency (RE) *between two estimators are defined as*

$$\mathrm{RE}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\mathrm{MSE}(\widehat{\theta}_2)}{\mathrm{MSE}(\widehat{\theta}_1)}.$$

In our case, this is $\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = \frac{\mathrm{MSE}(\widehat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})}$, with values above 1 indicating $\bar{A}$ should be preferred while values below 1 indicate $\widehat{P}$ should be preferred. Relative efficiency is a useful metric for comparing estimators because it will frequently be invariant to the scale of the noise in the problem and hence is more easily comparable across different settings.

In this section, we analyze the performance of these two estimators under the SBM by computing the entry-wise relative efficiency. We also consider the *asymptotic relative efficiency (ARE)*, which is the limit of the relative efficiency as the number of vertices $n \to \infty$ but with the number of graphs m fixed, and the scaled relative efficiency, $n \cdot \mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ which in our case normalizes the relative efficiency so that the asymptotic scaled relative efficiency is non-zero and finite. Somewhat surprisingly, we will see that the asymptotic relative efficiency will not depend on this fixed sample size m.

For this asymptotic framework, we assume the block memberships τ_i are drawn iid from a categorical distribution with block membership probabilities given by $\rho \in [0,1]^K$. In particular, this implies that for each block k, the proportion $|\{i : \tau_i = k\}|/n$

of vertices in block k will converge to ρ_k as $n \to \infty$ by the law of large numbers. We will also assume that for a given n, the block membership probabilities are fixed for all graphs. Denote block probability matrix $B = \nu\nu^\top \in [0,1]^{K\times K}$. By definition, the mean of the collection of graphs generated from this SBM is $P \in [0,1]^{N\times N}$, where $P_{ij} = B_{\tau_i,\tau_j}$. After observing m graphs on n vertices $A^{(1)}, \cdots, A^{(m)}$ sampled independently from the SBM conditioned on τ, we can calculate the two estimators $\bar{A}$ and $\widehat{P}$.

Lemma 3.4.2 *For the above setting, for any $i \neq j$, if $\operatorname{rank}(B) = K = d$, we have for large enough n,*

$$\mathbb{E}[(\widehat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{mn} P_{ij}(1 - P_{ij}).$$

And

$$\lim_{n\to\infty} n \cdot \operatorname{Var}(\widehat{P}_{ij}) = \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{m} P_{ij}(1 - P_{ij}).$$

The first part of this lemma ensures that the estimator is asymptotically unbiased for P, and the second part gives the form of the asymptotic variance of $\widehat{P}$.

The proof of this lemma is given in Section 3.8 and is based on results for the variance of the adjacency spectral embedding from Athreya et al. [2016]. Here we provide an outline of the proof that leads to the approximate MSE of $\widehat{P}$ in the stochastic blockmodel case. The result depends on using the asymptotic results (see Theorem 3.8.1) for the distribution of eigenvectors from Athreya et al. [2016] which

extend to the multiple graph setting in a straightforward way.

Outline of Proof: The first key observation is that since $\bar{A}$ is computed from iid observations each with expectation P, $\bar{A}$ is unbiased for P and $\text{Var}(A_{ij}) = \frac{1}{m}P_{ij}(1 - P_{ij})$. The results of Athreya et al. [2016] provide a central limit theorem for estimates of the latent position in an RDPG model for a single graph. Details of this theorem are in Theorem 3.8.1. Since the variance of each entry is scaled by $1/m$ in $\bar{A}$, the analogous result for $\bar{A}$ is that the estimated latent positions will follow an approximately normal distribution with variance scaled by $1/m$ compared to the variance for a single graph.

Since $\widehat{P}_{ij} = \widehat{X}_i^\top \widehat{X}_j$ is a noisy version of the dot product of $\nu_s^\top \nu_t$ and each $\widehat{X}_i$ is approximately independent and normal, we can use common results for the variance of the inner product of two independent multivariate normals [Brown and Rutemiller, 1977]. After simplifications that occur in the stochastic blockmodel setting, we can derive that the variance of $\widehat{P}_{ij}$ converges to $\left(1/\rho_{\tau_i} + 1/\rho_{\tau_j}\right) P_{ij}(1 - P_{ij})/(n \cdot m)$ as $n \to \infty$. Since the variance of $\bar{A}_{ij}$ is $P_{ij}(1 - P_{ij})/m$, the relative efficiency between $\widehat{P}_{ij}$ and $\bar{A}_{ij}$ is approximately $(\rho_{\tau_i}^{-1} + \rho_{\tau_j}^{-1})/n$ when n is sufficiently large. ■

From Lemma 3.4.2, we can see that the MSE of $\widehat{P}_{ij}$ is of order $O(m^{-1}n^{-1})$ approximately. Similar to $\bar{A}$, the estimate will get better as the number of observations m increases. Furthermore, it also benefits from a larger graph because of the use of low-rank structure. That is, $\widehat{P}$ will perform better as the number of vertices of the graph n increases.

Moreover, since $\bar{A}_{ij}$ is the sample mean of m independent Bernoulli random variables with parameter P_{ij}, we have

$$\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2] = \frac{P_{ij}(1 - P_{ij})}{m}.$$

Based on this MSE result of $\bar{A}_{ij}$ and the MSE result of $\widehat{P}_{ij}$ given by Lemma 3.4.2, we can conclude the following theorem naturally.

Theorem 3.4.3 *In the same setting as in Lemma 3.4.2, for any $i \neq j$, if* $\operatorname{rank}(B) = K = d$*, then for large enough n, we have*

$$\operatorname{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{n}. \tag{3.1}$$

And the asymptotic relative efficiency (ARE) is

$$\operatorname{ARE}(\bar{A}_{ij}, \widehat{P}_{ij}) = \lim_{n\to\infty} \operatorname{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = 0.$$

Proof: Combine the MSE result of $\bar{A}_{ij}$

$$\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2] = \frac{P_{ij}(1 - P_{ij})}{m},$$

and Lemma 3.4.2, i.e. for large enough n,

$$\mathbb{E}[(\widehat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{mn} P_{ij}(1 - P_{ij}),$$

we have for large enough n,

$$\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = \frac{\mathrm{MSE}(\widehat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})} = \frac{\mathbb{E}[(\widehat{P}_{ij} - P_{ij})^2]}{\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2]} \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{n}.$$

And the ARE result follows directly by taking the limit of RE as $n \to \infty$. ■

This theorem indicates that under the SBM, $\widehat{P}$ is a much better estimate of the mean of the collection of graphs P than $\bar{A}$. Note that a relative efficiency less than 1 indicates that $\widehat{P}$ should be preferred over $\bar{A}$, so under the above assumptions, as $n \to \infty$, $\widehat{P}$ performs far better than $\bar{A}$. From the result, we see that the relative efficiency is of order $O(n^{-1})$ and $n \cdot \mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j}$ when $n \to \infty$. An important aspect of Theorem 3.4.3 is that the ARE does not depend on the number of graphs m, so the larger the graphs are, the better $\widehat{P}$ is relative to $\bar{A}$, regardless of m.

Note that the asymptotic result here is for number of vertices going to infinity with a fixed number of graphs. Such setting is very useful in certain circumstances, for example connectomics analysis since we anticipate the collection of larger and larger brain network which will also likely initially correspond to smaller sample sizes as the technology to scale these connectome collection techniques is developed.

The approximate formula Equation 3.1 indicates that the sizes of the blocks can greatly impact the relative efficiency. As an example, consider a 2-block SBM. If each of the blocks contain half the vertices, then for each pair of vertices, the relative efficiency is approximately $4/n$. If the first block gets larger, with $\rho_1 \to 1$, then the RE for estimating P_{ij} with $\tau_i = \tau_j = 1$ will tend to its minimum of $2/n$. On the other hand as $\rho_1 \to 1$, if $\tau_i = 1$ and $\tau_j = 2$, then since $\rho_2 = 1 - \rho_1$, the relative efficiency for estimating such an edge pair will be approximately 1 and the same will hold if $\tau_i = \tau_j = 2$. Note that the maximum value for the relative efficiency in a two-block model is achieved when $\rho_1 = 1/n$ and $\rho_2 = (n-1)/n$ in which case the relative efficiency is $n/(n-1) \approx 1$. (Note values of ρ_s below $1/n$ correspond to graphs where typically no vertices are in that block, so the effective minimum we can consider for ρ_s is $1/n$.)

To illustrate Equation 3.1 of Theorem 3.4.3, we consider a 2-block SBM with parameters

$$B = \begin{bmatrix} 0.42 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \tag{3.2}$$

so that $|\{i : \tau_i = 1\}| \approx |\{i : \tau_i = 2\}|$, especially for large n. Note that this simulation only focuses on the rank-2 setting primarily for the interpretability. When calculating $\widehat{P}$, we omit the dimension selection step from Algorithm 1 and instead use the true dimension $d = \text{rank}(B) = 2$. Figure 3.3 shows $2/\rho_1$ and $1/\rho_1 + 1/\rho_2$, the scaled asymptotic RE for pairs of vertices both in block one and pairs of vertices in different

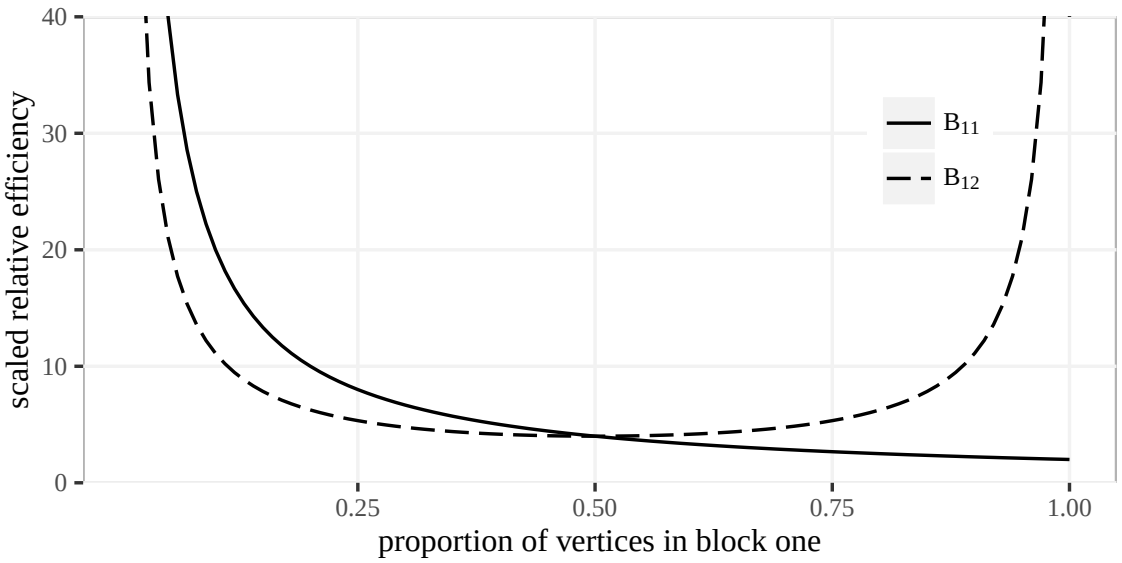


Figure 3.3: Asymptotic scaled relative efficiency $n \cdot \text{RE}(\bar{A}, \widehat{P})$ in a 2-block SBM. For each distinct pair of edge probabilities in a 2-block SBM specified in Eq. 3.2, the scaled relative efficiency only depends on the proportion of vertices in each block. We show the scaled asymptotic relative efficiency as ρ_1 changes from $0, 1$ for pairs of vertices where either both are in block one or one is in block one and one is in block two. These curves all intersect at a scaled relative efficiency of 4 when $\rho_1 = 1/2 = \rho_2$. Improvements using low-rank methods are greater for larger blocks, such as for B_{11} when ρ_1 is close to 1, while the improvements are smaller for block pairs with relatively few vertex pairs such as B_{11} when ρ_1 is small and B_{12} when ρ_1 is near 0 or 1. Note that the curve for B_{22} would be the same as that for B_{11} but reflected around the vertical line when $\rho_1 = 1/2$. Overall, $\widehat{P}$ performs best for large blocks while the improvements may be very minor for blocks with only a few vertices.

blocks, respectively, in the 2-block SBM we specified earlier. We vary ρ_1 between 0 and 1 to demonstrate how the number of pairs of vertices with the corresponding block memberships impacts the overall relative efficiency. For $n = 500$ and $m = 100$, estimates of the scaled RE based on simulations agree very closely with their corresponding theoretical values displayed in the figure. Note that when $\rho_1 = 0.5$, the scaled RE has value 4.0, which agrees with the result in Figure 3.4 for simulated

data.

If instead of assuming that the graphs follow an SBM distribution, we assume the graphs are distributed according to an RDPG distribution, similar gains in relative efficiency can be realized. While there is no compact analytical formula for the relative efficiency of $\widehat{P}$ versus $\bar{A}$ in the general RDPG case, using the same ideas as in Theorem 3.4.3, we can show that $\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = O(1/n)$.

Proposition 3.4.4 *Suppose that $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$ are independently and identically distributed from an RDPG distribution with common latent positions $X_1, \ldots, X_n$, which are independently and identically distributed from some distribution. As the number of vertices $n \to \infty$, it holds for any $i \neq j$ that*

$$\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = O(1/n).$$

where again the asymptotic relative efficiency in n does not depend on m.

The proof of this proposition closely follows the proofs of Lemma 3.4.2 and Theorem 3.4.3, and hence we omit it here.

Remark 3.4.5 *As we noted above, if the graphs are distributed according to an SBM or an RDPG, the relative efficiency is approximately invariant to the number of graphs m when n is large. If on the other hand, the graphs are generated according to a full-rank independent edge model, then the relative efficiency can change more dramatically as m changes. The reason for this is because for larger m, more of the eigenvectors of*

$\bar{A}$ will begin to concentrate around the eigenvectors of the mean graph. This leads to the fact that the optimal embedding dimension for estimating the mean will increase, making $\bar{A}$ and the low-rank approximation at the optimal dimension closer together. As a result, $\mathrm{RE}(\bar{A}, \widehat{P})$ *will increase as m increases for full-rank models. Indeed, for large m we could have* $\mathrm{RE}(\bar{A}, \widehat{P}) \geq 1$ *since we cannot guarantee that $\widehat{P}$ will choose the optimal dimension. The lack of gaps in the eigenvalues of the mean graph makes dimension reduction quite dangerous. In an extreme case, the low-rank assumption will be most violated when all eigenvalues of the mean graph are almost equal. This leads to a certain type of structure, which is close to a constant times the identity matrix. However we do not see such structure in connectomics. We will discuss this further in Section 3.6 when applying our estimator to the CoRR dataset.*

3.5 Finite Sample Toy Model Simulations

We first illustrate the theoretical results from Section 3.4 regarding the relative efficiency between $\bar{A}$ and $\widehat{P}$ via Monte Carlo simulation experiments in an idealized setting. These numerical simulations will also allow us to investigate the finite sample performance of the two estimators. Note that in Section 3.7, we will break the model assumptions slightly and run experiment in a more realistic setting.

Here, we consider the same 2-block SBM as in Equation 3.2. To be clear, we

restate the parameters here:

$$B = \begin{bmatrix} 0.42 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

Similarly, when calculating $\widehat{P}$, we omit the dimension selection step from Algorithm 1 and instead use the true dimension $d = \text{rank}(B) = 2$.

To investigate the finite sample relative efficiency, we first sample 1000 Monte Carlo replicates from the above SBM distribution with different numbers of vertices $N \in \{30, 50, 100, 250, 500, 1000\}$ and a fixed number of graphs $m = 100$. The relative efficiency $\text{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ can be estimated because P is known for this simulation. Since the relative efficiency only depends on the block memberships of the pair i, j, we estimate the relative efficiency for each block pair using

$$\widehat{\text{RE}}_{st}(\bar{A}, \widehat{P}) = \frac{\sum_{\tau_i=s,\tau_j=t,i\neq j} \widehat{\text{MSE}}(\widehat{P}_{ij})}{\sum_{\tau_i=s,\tau_j=t,i\neq j} \widehat{\text{MSE}}(\bar{A}_{ij})}$$

for $s, t \in \{1, 2\}$, where $\widehat{\text{MSE}}$ denotes the estimated mean squared error based on the Monte Carlo replicates. For the remaining simulations and real data analysis, we will always be considering estimated relative efficiency and estimated mean squared error rather than analytic results, and hence we will frequently omit that these are estimated values when it is clear from context.

In Figure 3.4, we plot the (estimated) relative efficiency (top panel) and the scaled

(estimated) relative efficiency (bottom panel), $n \cdot \widehat{\mathrm{RE}}_{st}(\bar{A}, \widehat{P})$. The different dashed lines denote the RE and scaled RE associated with different block pairs, either B_{11}, B_{12}, or B_{22}. As expected from Theorem 3.4.3, the top panel indicates that the relative efficiencies are all very close together and much less than 1, decreasing at the rate of $1/n$, indicating that $\widehat{P}$ is performing better than $\bar{A}$.

Based on Theorem 3.4.3, we also have that the scaled RE converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j} = 4$ as $n \to \infty$ for all pairs i, j. This is plotted as a solid line in the bottom panel. From the figure, we see that $n \cdot \widehat{\mathrm{RE}}_{st}(\bar{A}, \widehat{P})$ converges to scaled asymptotic RE quite rapidly. We omit error bars as the standard errors are very small for these estimates.

Remark 3.5.1 *An intriguing aspect of these finite sample results is that the scaled relative efficiencies behave differently for small graphs with fewer vertices. The estimates of the edge probabilities for pairs of vertices in different blocks are much better than the estimates for edges within each block. The reason for this is unclear and could be due to the actual values of the true probability, but it may also be due to the fact that there are approximately twice as many pairs of vertices in different blocks, $n^2/4$, than there are in the same block, $n^2/8 - n/4$. This could lead to an increase in effective sample size which may cause the larger differences displayed in the left parts of Figure 3.4. However, overall these differences are nearly indistinguishable for unscaled relative efficiency.*

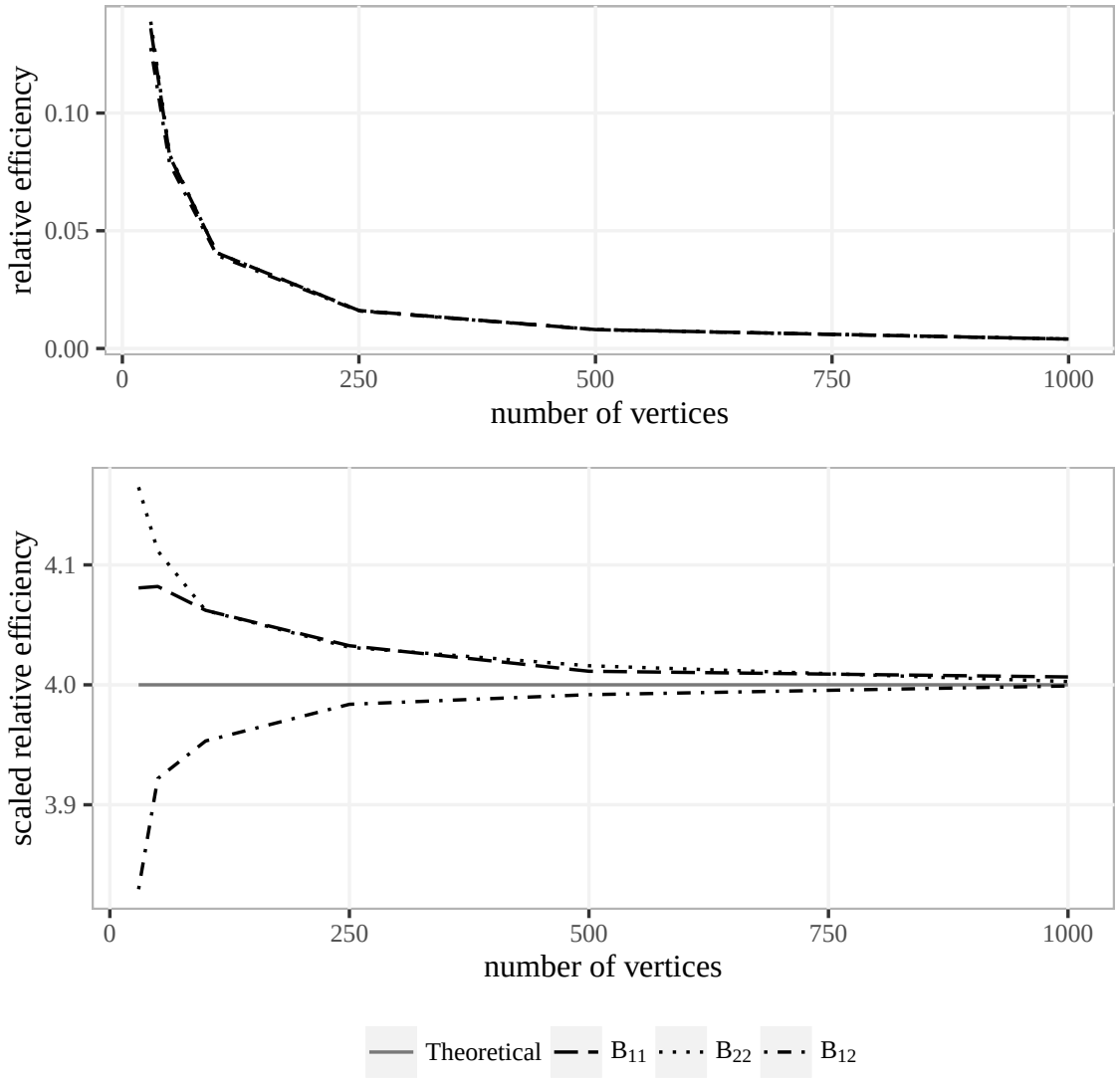


Figure 3.4: Finite sample relative efficiency based on simulations. The top panel shows the estimated relative efficiency $\widehat{\mathrm{RE}}(\bar{A}, \widehat{P})$ as a function of n for fixed $m = 100$ based on simulations of an SBM. For each value of n, we used 1000 Monte Carlo replicates of the SBM from Section 3.5 to estimate the RE. Each curve corresponds to an average across vertex pairs corresponding to the three distinct block probabilities B_{11}, B_{12}, and B_{22} in the two-block SBM. Recall that values below 1 indicate that $\widehat{P}$ is performing better than $\bar{A}$. The relative efficiencies are all very close so the lines are indistinguishable.
To distinguish the three curves, the bottom panel shows the corresponding scaled relative efficiencies, $n \cdot \widehat{\mathrm{RE}}(\bar{A}, \widehat{P})$. The solid horizontal line indicates the theoretical asymptotic scaled relative which is $1/\rho_s + 1/\rho_t = 4$, since $\rho_1 = \rho_2 = 4$. All the curves converge quickly to this theoretical limit.

3.6 CoRR Brain Graphs Experiment

In practice, graphs do not follow the independent edge model, let alone an RDPG or SBM, but the mean of a collection of graphs is still of interest for these cases. To demonstrate that the estimator $\widehat{P}$ is still useful in such cases, we tested its performance on structural connectomic data. The graphs are based on diffusion tensor MR images collected and available at the Consortium for Reliability and Reproducibility (CoRR) [Gorgolewski et al., 2015, Zuo et al., 2014]. The dataset contains 454 different brain scans, each of which was processed to yield an undirected, unweighted graph with no self-loops, using the pipeline described in Roncal et al. [2013] and **?**. The vertices of the graphs represent different regions in the brain defined according to an atlas. We used three atlases, the JHU atlas with 48 vertices, the Desikan atlas with 70 vertices, and the CPAC200 atlas with 200 vertices. An edge exists between two vertices whenever there is at least one white-matter tract connecting the corresponding two regions of the brain. Details of this dataset are provided in the following section.

3.6.1 Dataset Description

The original dataset is from the Emotion and Creativity One Year Retest Dataset provided by Qiu, Zhang and Wei from Southwest University available at the Consortium for Reliability and Reproducibility (CoRR) [Gorgolewski et al., 2015, Zuo

et al., 2014]. It is comprised of 235 subjects, all of whom were college students. Each subject underwent two sessions of anatomical, resting state DTI scans, spaced one year apart. Due to incomplete data, only 454 scans are available.

When deriving MR connectomes, Kiar et al. [2016] parcellate the brain into groups of voxels as defined by anatomical atlases. The atlases are defined either physiologically by neuroanatomists (Desikan and JHU), or are generated using an automated segmentation algorithm (CPAC200). Once the voxels in the original image space are grouped into regions, an edge is placed between two regions when there is at least one white-matter tract, derived using a tractography algorithm, connecting the corresponding two parts of the brain. The resulting graphs are undirected, unweighted, and have no self-loops.

3.6.2 Experiment Results

In order to evaluate the performance of the two estimators, we used a cross validation on the 454 graphs of each size. Specifically, for a given atlas, each Monte Carlo replicate corresponds to sampling m graphs out of the 454 and computing the low-rank estimator $\widehat{P}$ and the sample mean $\bar{A}$ using the m selected graphs. We then compared these estimates to the sample mean for the remaining $454 - m$ adjacency matrices. While we cannot interpret this mean graph as the probability matrix for an IEM distribution (see Section 3.7), the sample mean for the remaining graphs does give the proportion of times each pair of vertices are adjacent in the population from

which the graphs were sampled.

While in previous sections we evaluated the mean squared error for either an individual entry or for an entire block in the SBM, in this section and the next section we will focus on the overall error for estimating the mean graph. In particular we will use the average of the mean squared error across all pairs of vertices and we define $\text{MSE}(\bar{A}) = \binom{n}{2}^{-1} \sum_{i<j} \mathbb{E}[\bar{A}_{ij} - P_{ij}]$ and similarly for $\text{MSE}(\widehat{P})$, which are also used for the relative efficiency. As in the previous section, we will not use analytical evaluations of the MSE and instead estimate the MSE and relative efficiencies via Monte Carlo simulations.

We ran 1000 simulations on each of the three atlases for sample size $m = 5, 10$. For $m = 1$, we only have 454 different possibilities. So instead of running 1000 simulations, we looked through all 454 possible sample with size 1. As long as we determine which dimension to embed, the two estimates $\bar{A}$ and $\widehat{P}$ can be calculated based on the sample. In practice, we use algorithms like Zhu and Ghodsi's method or USVT discussed in Section 3.2.2 to select the dimension d. These methods are neither computationally advanced nor requiring sophisticated algorithms. We plot the estimated relative efficiencies between $\bar{A}$ and $\widehat{P}$ in Figure 3.5.

For each atlas and each sample size, we compare the Zhu and Ghodsi method [Zhu and Ghodsi, 2006] with the USVT method [Chatterjee, 2015] and note that both perform reasonably well relative to the full-dimensional $\bar{A}$. We omit confidence intervals for the estimated relative efficiencies since all confidence intervals have lengths less

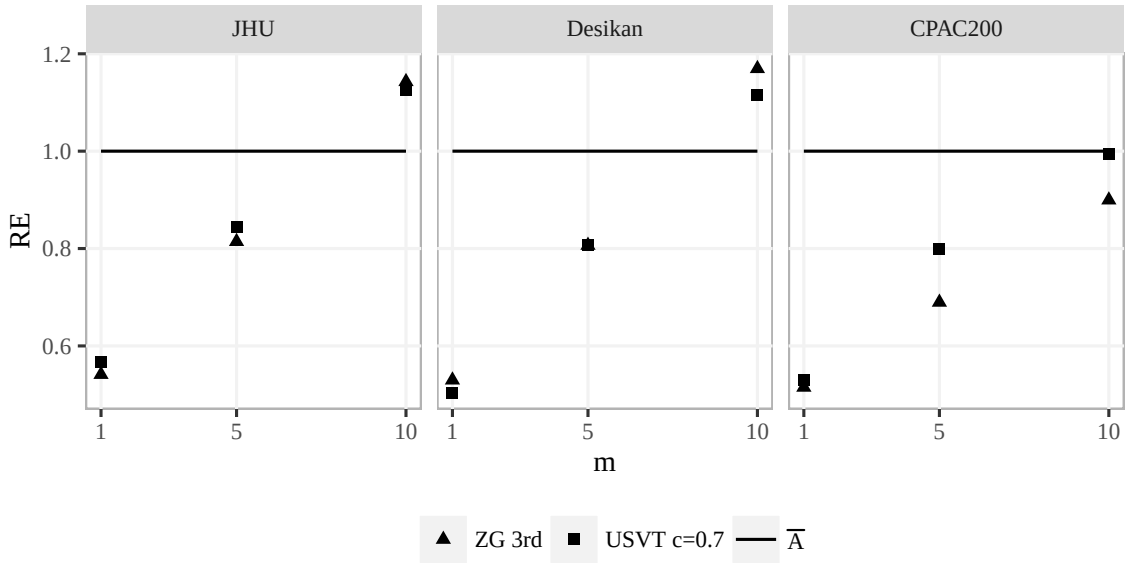


Figure 3.5: Relative efficiencies of $\bar{A}$ versus $\widehat{P}$ for the CoRR data set. For each atlas, JHU, Desikan, and CPAC 200, we sampled graphs which we used to compute $\bar{A}$ and $\widehat{P}$. We compared different sample sizes m and different dimension selection procedures, ZG and USVT. For each of the two methods for computing $\widehat{P}$, we estimated their relative efficiencies with respect to the sample mean $\bar{A}$. Confidence intervals all had lengths less than 0.015, and hence we omitted them for clarity. Overall, the relative efficiencies are greater for smaller sample sizes m and larger number of vertices n.

than 0.015, indicating that all relative efficiencies, aside from the relative efficiency for the CPAC200 atlas at $m = 10$, are very different from 1.

Again we can see that the largest improvements using $\widehat{P}$ occur when m is small and n is large, where the RE are smaller than 1. On the other hand, once $m = 10$, $\bar{A}$ tends to do nearly as well or better than $\widehat{P}$. Nonetheless, when applied to subgroups inference, such as all females between the age of 21 and 25, $\widehat{P}$ can be really helpful for better exploring differences between groups compared to $\bar{A}$ due to a small sample size of each subgroup. In addition, $\widehat{P}$ offers certain advantages, especially since low-rank estimates can often be more easily interpretable by considering the latent position representation which will be discussed in Section 3.6.4.

To further illustrate the differences between the two estimators, we considered a single random sample of size $m = 5$ based on the Desikan atlas. We calculated $\bar{A}$ and $\widehat{P}$, using Zhu and Ghodsi's 3rd elbow to select $d = 11$. In Figure 3.1, the estimates $\bar{A}$ and $\widehat{P}$ as well as the sample mean of 454 graphs (as a close estimate of P) are plotted in the upper level. Since the sample size is small, there are a lot of pairs of vertices with no edges or 5 edges in the 5 observations. This leads to the white and black pixels in the image corresponding to $\bar{A}$. On the other hand, $\widehat{P}$ has a finer gradient of values which in this case leads to a more accurate estimate. By calculating the mean squared error based on this sample, we can see that $\widehat{P}$ (with mean squared error equals 0.015) outperforms $\bar{A}$ (with mean squared error equals 0.016), with a 3% relative improvement. In order to see where the improvements are clearly, in the

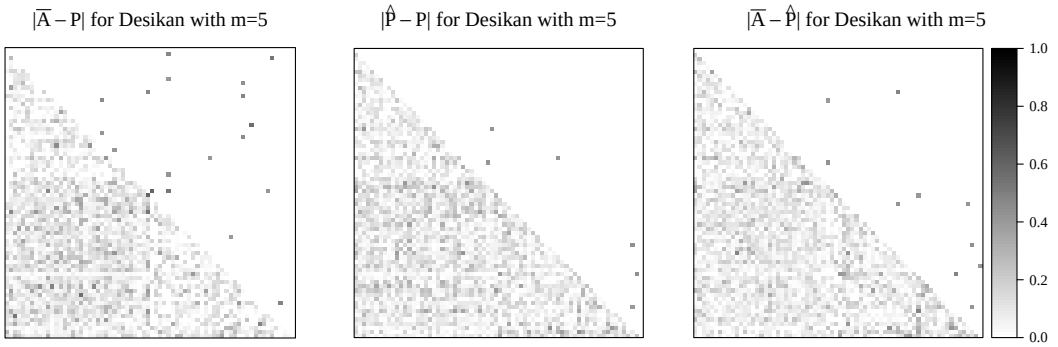


Figure 3.6: Heat plots of absolute estimation error for $\bar{A}$ and $\widehat{P}$ (lower triangle) and absolute errors above 0.4 (upper triangle). These heat plots show the absolute estimation error $|\bar{A} - P|$, $|\widehat{P} - P|$ and $|\bar{A} - \widehat{P}|$ for a sample of size $m = 5$ from the Desikan dataset. The embedding dimension for $\widehat{P}$ is $d = 11$ selected by the 3rd elbow of the ZG method. The lower triangular matrix shows the actual absolute difference, while the upper triangular matrix only highlights the edges with absolute differences larger than 0.4. The fact that 18 edges from $\bar{A}$ are highlighted and only 6 edges from $\widehat{P}$ are highlighted indicates that $\widehat{P}$ has fewer large outliers compared to $\bar{A}$.

upper triangular of the heat maps for $\bar{A}$ and $\widehat{P}$, we highlight the edges (18 edges highlighted for $\bar{A}$ and 6 for $\widehat{P}$) which have absolute estimation error larger than 0.4.

Moreover, for the same sample discussed above, Figure 3.6 shows the values for the absolute estimation error $|\bar{A} - P|$ and $|\widehat{P} - P|$. In addition, we include the absolute difference $|\bar{A} - \widehat{P}|$ to show the overall difference between the two estimates. The lower triangular sections show the actual absolute difference while the upper triangular matrix highlights the vertex pairs with absolute differences larger than 0.4. There are 18 edges from $\bar{A}$ and 6 edges from $\widehat{P}$ being highlighted in the figure, further indicating the superior performance of $\widehat{P}$. Note that approximately 13% of all pairs of vertices are adjacent in all 454 graphs and hence $\bar{A}$ will always have zero error for those pairs of vertices. Nonetheless, $\widehat{P}$ typically outperforms $\bar{A}$.

To investigate the difference in performance with respect to the geometry of the brain, in Figure 3.7 we plot the 50 edges with the largest differences $|\bar{A}_{ij} - P_{ij}| - |\widehat{P}_{ij} - P_{ij}|$ according to the location of the corresponding regions in the brain. Red edges indicate that $\widehat{P}$ overestimates P, while blue means that $\widehat{P}$ underestimates P. The edge width is determined by the estimation error for $\widehat{P}$, where pairs with larger estimation error are represented by thicker lines. We also highlight the five regions corresponding to vertices that contribute most to the difference, meaning the vertices i with the largest value of $\sum_j(|\bar{A}_{ij} - P_{ij}| - |\widehat{P}_{ij} - P_{ij}|)$. Notably, three of these top five regions form a contiguous group of regions. The top five regions are the inferior temporal, middle temporal, and transverse temporal regions in the left hemisphere and the parahippocampal and parsopercularis regions in the right hemisphere of the Desikan atlas.

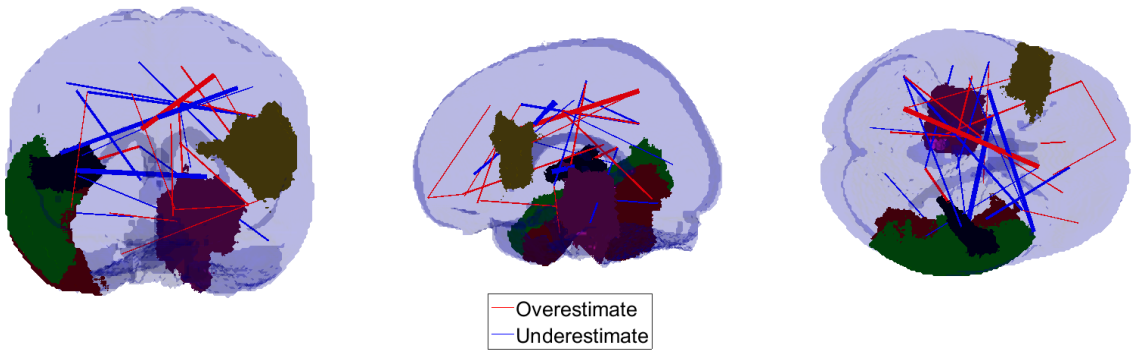


Figure 3.7: Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with the largest differences $|\bar{A}_{ij} - P_{ij}| - |\widehat{P}_{ij} - P_{ij}|$. Red edges indicate that $\widehat{P}$ overestimate P while blue means that $\widehat{P}$ underestimates P. The edge width is determined by the estimation error. Connections with larger estimation error are represented by thicker lines. This figure shows the regions and connections of the brain where $\widehat{P}$ outperforms $\bar{A}$ the most for estimating P.

3.6.3 Exploration of Dimension Selection Procedures

To further investigate the impact of the dimension selection procedures, we also considered all possible dimensions for $\widehat{P}$ by ranging d from 1 to n in order to investigate the impact of the dimension selection procedures. We plot $\widehat{\text{MSE}}$ of $\bar{A}$ and $\widehat{P}$ in Figure 3.8. The horizontal axis gives dimension d, which only impacts $\widehat{P}$, which is why estimated MSE of $\bar{A}$ is shown as flat.

When d is small, $\widehat{P}$ underestimates the dimension and throws away important information, which leads to relatively poor performance. When $d = n$, $\widehat{P}$ is equal to $\bar{A}$, so that the curve for $\widehat{\text{MSE}}$ for $\widehat{P}$ ends at $\widehat{\text{MSE}}(\bar{A})$.

In the figure, we denote the 3rd elbow found by the Zhu and Ghodsi method by a triangle, and denote the dimension selected by USVT with threshold 0.7 by a square. Both dimension selection algorithms tend to select dimensions which nearly minimize the mean squared error.

When m is 1 or 5, $\bar{A}$ has large variance which leads to large $\widehat{\text{MSE}}$. Meanwhile, $\widehat{P}$ reduces the variance by taking advantages of inherent low-rank structure of the mean graph. Such smoothing effect is especially obvious while we only have 1 observation. When $m = 1$, all weights of the graph are either 0 or 1, leading to a very bumpy estimate $\bar{A}$. In this case, $\widehat{P}$ smooths the connectomes estimate and improves the performance. Additionally, we see that there is a large range of dimensions where the

performance for $\widehat{P}$ is superior to $\bar{A}$. With a larger m, the performance of $\bar{A}$ improves so that its performance is frequently superior but nearly identical to $\widehat{P}$.

3.6.4 Interpretability of Low-rank Methods

Low-rank methods can often be more easily interpreted in a vertexy way. In particular, in the RDPG model, by representing a low-rank matrix in terms of the latent positions, where each vertex is represented as a vector in $\mathbb{R}^d$ and the entries of the matrix are given by the inner products of these vectors, one can analyze and visualize the geometry of these vectors in order to interpret how each vertex is behaving in the context of the larger graph. Now we take the CoRR brain graphs with Desikan atlases as an example. By embedding the mean graph P which is the average of all 454 graphs, we get the estimated latent positions $\widehat{X} \in \mathbb{R}^{n \times d}$, where $n = 70$ is the number of vertices and $d = 8$ is the dimension selected by the Zhu and Ghodsi's method. We color the brain using the first 4 dimensions of $\widehat{X}$ as in Figure 3.9 respectively. By the figure of the second dimension, we can see a clear distinction of the left and right hemisphere as conveyed in the second dimension. Additionally, such a representation allows the use of techniques from multivariate analysis to further study the estimated population mean.

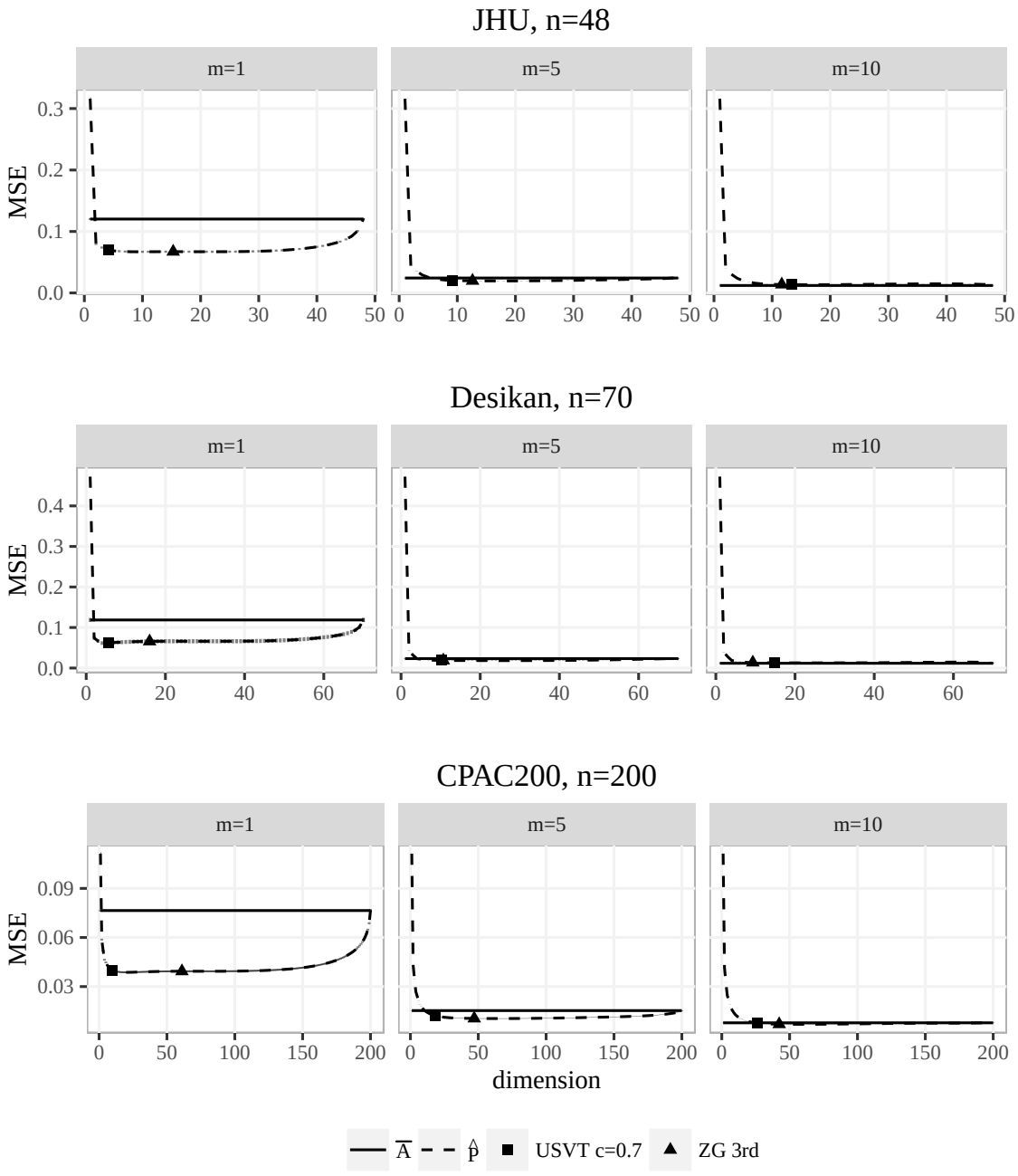
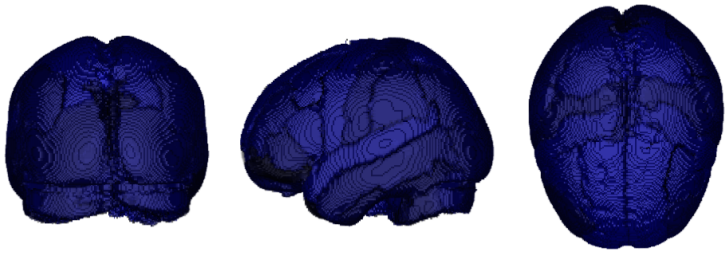
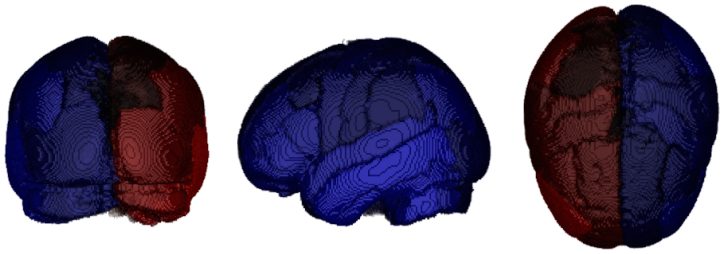


Figure 3.8: Comparison of $\widehat{\text{MSE}}$ of $\widehat{P}$ and $\bar{A}$ for three atlases at three sample sizes for the CoRR data. These plots show the MSE for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for three dataset (JHU, Desikan, and CPAC200) while embedding the graphs into different dimensions and with different sample sizes m. The average dimensions chosen by the 3rd elbow of Zhu and Ghodsi is denoted by a triangle and those chosen by USVT with threshold equaling 0.7 is denoted by a square. Vertical intervals, visible mainly in the $n = 48, 70$ and $m = 1$ plots, represent the 95% confidence interval for the mean squared errors.

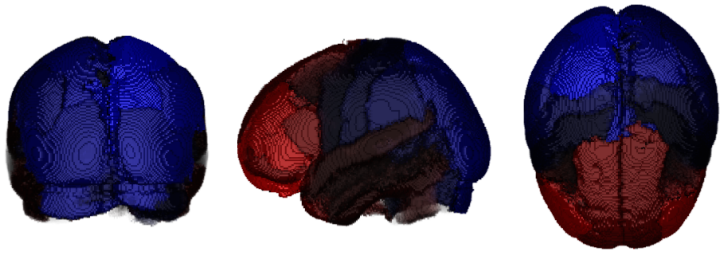
(a) 1st dimension



(b) 2nd dimension



(c) 3rd dimension



(d) 4th dimension

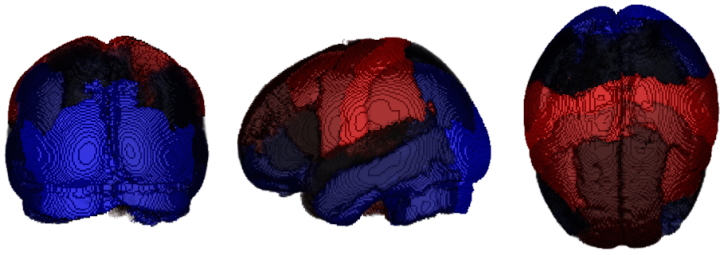


Figure 3.9: Brain plots colored by the first 4 dimensions of $\widehat{X}$ for the Desikan atlas respectively. We plot the brain using the first 4 dimensions of $\widehat{X}$. From the figure, we can see the embeddings have their own neuro-meaning, for example there is a clear distinction of the left and right hemisphere as conveyed in the second dimension. Also, the first dimension provides an average level of the entire brain.

3.6.5 Challenges of the CoRR Dataset

While our estimator $\widehat{P}$ performs well when the sample size m is small and the number of vertices n is large, the CoRR dataset itself does not strictly adhere to the low-rank assumptions of our theory.

As discussed in Remark 3.4.5, we first check whether the dataset has the low-rank property or not. In Figure 3.10, we plot the relative error $\|\text{lowrank}_d(P) - P\|_F^2 / \|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) as solid curves. The rate at which this curve tends to zero provides an indication of the relative performance of using $\widehat{P}_d$ as compared to $\bar{A}$ when m is large. For all three atlases, while these error rates do tend to zero relatively quickly, substantial errors remain for any low-rank approximation. This can be compared to the dashed lines which show how these errors would behave if P was truly low-rank. As can be expected, the CoRR dataset is actually high-rank, which violates the low-rank assumption.

In addition, by plotting the histograms of the eigenvalues of P in Figure 3.11, we see that there are a bunch of negative eigenvalues, which indicates that the positive semi-definiteness is also violated in this real data experiment. This makes it even harder for $\widehat{P}$ to outperform $\bar{A}$.

Two other parts of this dataset provide challenges for our low-rank methods. First, there are a large number of negative eigenvalues which $\widehat{P}$ will not capture. We can adapt our low-rank methods by including large negative eigenvalues as well however we found that for low sample sizes excluding negative eigenvalues improved

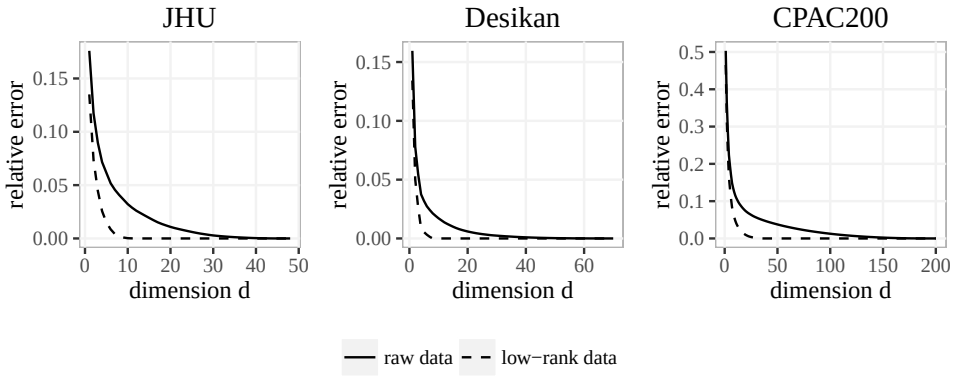


Figure 3.10: Relative error of the rank-d approximation of the population mean. The solid curves show the relative error $\|\text{lowrank}_d(P) - P\|_F^2 / \|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) for three different atlases. The relative error decays relatively slowly when d is close to n, which indicates that P is not low-rank. Also, if P actually has the low-rank property, the relative error plot will look like the dashed curves, where we revised P to be low-rank by only keeping a few large eigenvalues.

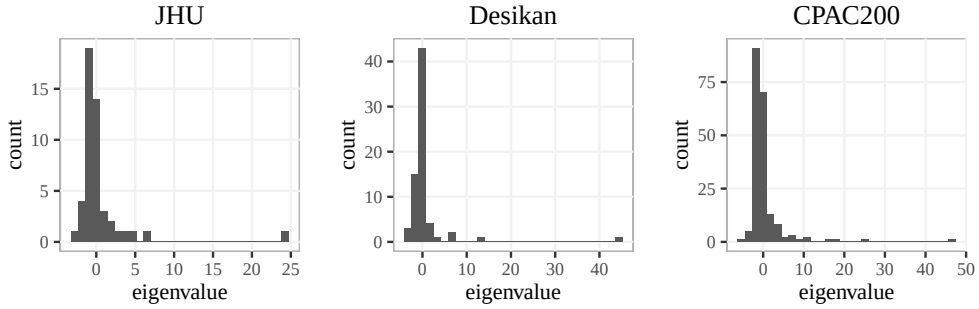


Figure 3.11: Histogram of the population mean. These figures show the histograms of the eigenvalues of the mean graph of all 454 graphs with diagonal augmentation. A bunch of negative eigenvalues indicate that P is not positive semi-definite.

performance. Second, approximately 12.8% of the entries of P are exactly equal to 1. For these edges, $\bar{A}$ will always have zero error, while $\widehat{P}$ will necessarily give a less accurate estimate.

Moreover, by the histogram of P, i.e. the entry-wise mean of all 454 graphs based on the Desikan atlas, as in Figure 3.12, we can clearly see more edge probabilities are concentrated on both sides, i.e. close to 0 or 1. In particular, 12.8% of the edges has probability equals to 1 exactly. For these edges, MLE $\bar{A}$ always recover the probability 1 exactly even with only 1 observation, while $\widehat{P}$ will give a less accurate estimate because of the smoothing effect. So the CoRR dataset we consider is highly preferable to $\bar{A}$ compared to $\widehat{P}$. However, even in this situation, $\widehat{P}$ still outperforms $\bar{A}$ when the sample size is relatively small.

Despite all these challenges, our results show that when the sample size is relatively small, such as $m = 1$ or $m = 5$, and for the atlases with a larger number of vertices, $\widehat{P}$ still gives a better estimate than $\bar{A}$ for the CoRR dataset. Importantly, this improvement is robust to the embedding dimension provided the dimension is not underestimated.

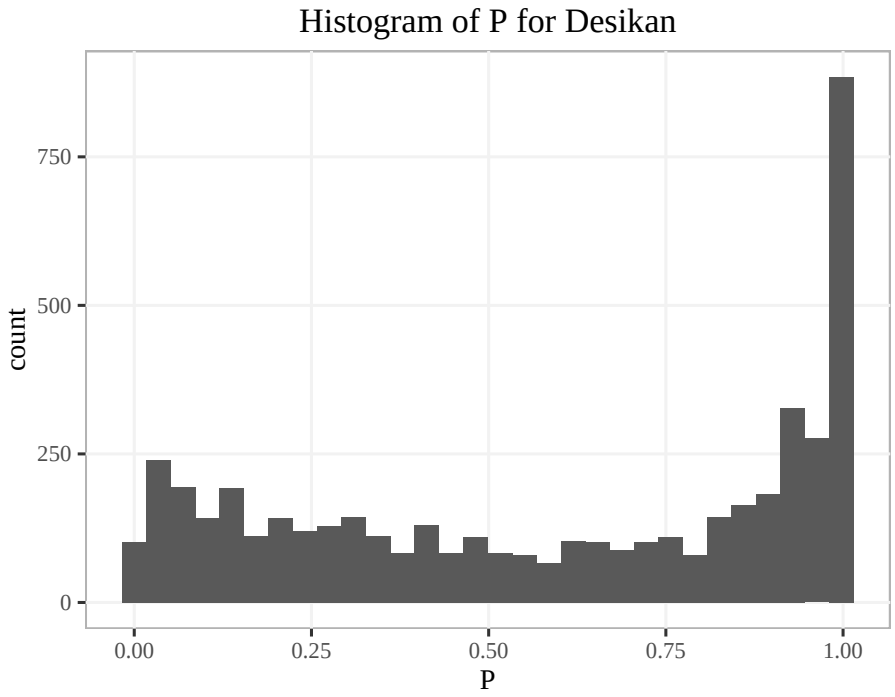


Figure 3.12: Histogram of P for Desikan atlas. This figure shows the histogram of P, i.e. the entry-wise mean of all 454 graphs based on the Desikan atlas. More edge probabilities of P are concentrated on both sides, i.e. close to 0 or 1. In particular, 12.8% of the edges has probability equals to 1 exactly.

3.6.6 Lobe Structure behind the Low-rank Methods

In previous sections, we have shown that how the low-rank methods help us improve the accuracy of estimation while providing convenient interpretations simultaneously. Certainly there is more behind it. In this section, we focus on the lobe structure in particular.

The lobes of the brain is an anatomical classification, which has been shown to be related to different brain functions. Basically, there are 4 different lobes for each hemisphere, i.e. Frontal, Parietal, Occipital, and Temporal. For the Desikan atlas, there are 70 different regions (35 regions for each hemisphere). Each region belongs to one lobe. However, 8 regions of the Desikan atlas (Unknown, Banks of Superior Temporal Sulcus, and Corpus Callosum, for both hemispheres) do not have obvious lobe assignment. So we cluster them into a new lobe category named "other" to resolve this issue.

Generally, the regions within the same lobe should be more similar compared to the regions across the lobes. In order to see whether the embedded latent positions X preserve this property or not, we propose a test statistics T to be the average differences between vertices within the same lobe minus the average differences between

vertices across different lobes, i.e.

$$T(X,l) = \frac{\sum_{i\neq j, l(i)=l(j)} \|X_i - X_j\|_2}{\sum_{i\neq j, l(i)=l(j)} 1} - \frac{\sum_{i\neq j, l(i)\neq l(j)} \|X_i - X_j\|_2}{\sum_{i\neq j, l(i)\neq l(j)} 1},$$

where $l(i)$ represents the lobe assignment for vertex i. If the latent positions X and the lobe assignment l are independent, then we expect $T(X,l)$ to be close to zero. A small test statistic $T(X,l)$ indicates that latent positions of the regions within the same lobe are closer compared to the ones across the lobes, which is evidence that the low-rank methods preserves the lobe structure.

However, the anatomical geometry might contribute to the dependence between X and l and we do not want to be distracted by this factor. So instead of testing the dependence between X and l, we are more interested in the following hypothesis test:

H_0: X and l are conditionally independent given anatomical geometry.

H_A: X and l are conditionally dependent given anatomical geometry.

In this situation, we can focus on how much of the lobe structure is really captured by the low-rank methods without affected by the inherent spatial relationship. Note that this test is significantly underpowered compared to the test on a unconditional independence. To test under the anatomical geometry conditions, we permute the lobe assignment l in a way that the number of regions in each lobe remain the same and the regions within the same lobe are still spatially connected. In order to permute under such constraints, we define a flip to be a swap of two pairs of vertices which

keeps the number of regions in each lobe, and do it several times.

As mentioned before, there are 10 lobes and 70 regions based on the Desikan atlas. We say two regions are adjacent if they share a common boundary. We denote such spatial adjacency by an adjacency matrix S for the 70 regions, where $S_{ij} = 1$ means region i and region j are contain a pair of voxels, v_i and v_j, which are spatially adjacent. If this is true, then we say region j is a neighbor of region i. We denote the lobe i.d. for region i by l_i.

Now we define a uniform 1-flip to be:

1. Select a pair of adjacent regions (region i_1 and region j_1) across the boundary of lobes uniformly, i.e. $S_{i_1 j_1} = 1$ and $l(i_1) \neq l(j_1)$;

2. Uniformly select another pair of adjacent regions (region i_2 and region j_2 where $i_1 \neq i_2$ and $j_1 \neq j_2$) across the same boundary of lobes uniformly, i.e. $S_{i_2 j_2} = 1$ and $l(i_1) = l(i_2)$ and $l(j_1) = l(j_2)$;

3. Reassign region j_1 to lobe l_{i_1} and reassign region i_2 to lobe l_{j_2}.

By the definition, after a uniform 1-flip, the number of regions in each lobe stays the same, where only two regions are changed to a different lobe.

Then we can define a uniform k-flip naturally as sequentially performing uniform 1-flip k times. Note that after a uniform k-flip, the number of regions in each lobe still keeps the same.

In the permutation test, we apply a uniform k-flip and calculate the test statistic

$T(X, l)$ based on the lobe assignment after flipping. The p-value is computed as the proportion of uniform k-flips with a T value smaller than the T value for the true lobe assignments.

For a fixed number of flips, we ran 1000 simulations and calculate the test statistics $T(X, l)$ after the permutation. We vary the number of flips from 1 to 10 and the results are shown in Figure 3.13. In this violin plot, we use dashed line to represent the $T(X, l)$ based on the true lobe assignment. As the number of flips increases, $T(X, l)$ converges to the expected value under the assumption that X and l are independent in a constrained way mentioned above. And we can see that the $T(X, l)$ according to the true lobe assignment move away from the 95% confidence interval. We also calculate the corresponding p-values and labeled them in the figure. We can see that when the number of flips is larger than 7, the p-value is less than 0.05, which suggests that the latent positions based on the low-rank methods preserves the lobe structure regardless of the anatomical geometry.

3.7 Synthetic Data Analysis for Full Rank IEM

While the theory we have developed is based on the assumption that the mean graph is low rank, as we have seen in Section 3.6, $\widehat{P}$ often performs well even when this assumption is false. To further illuminate this point, we performed a synthetic

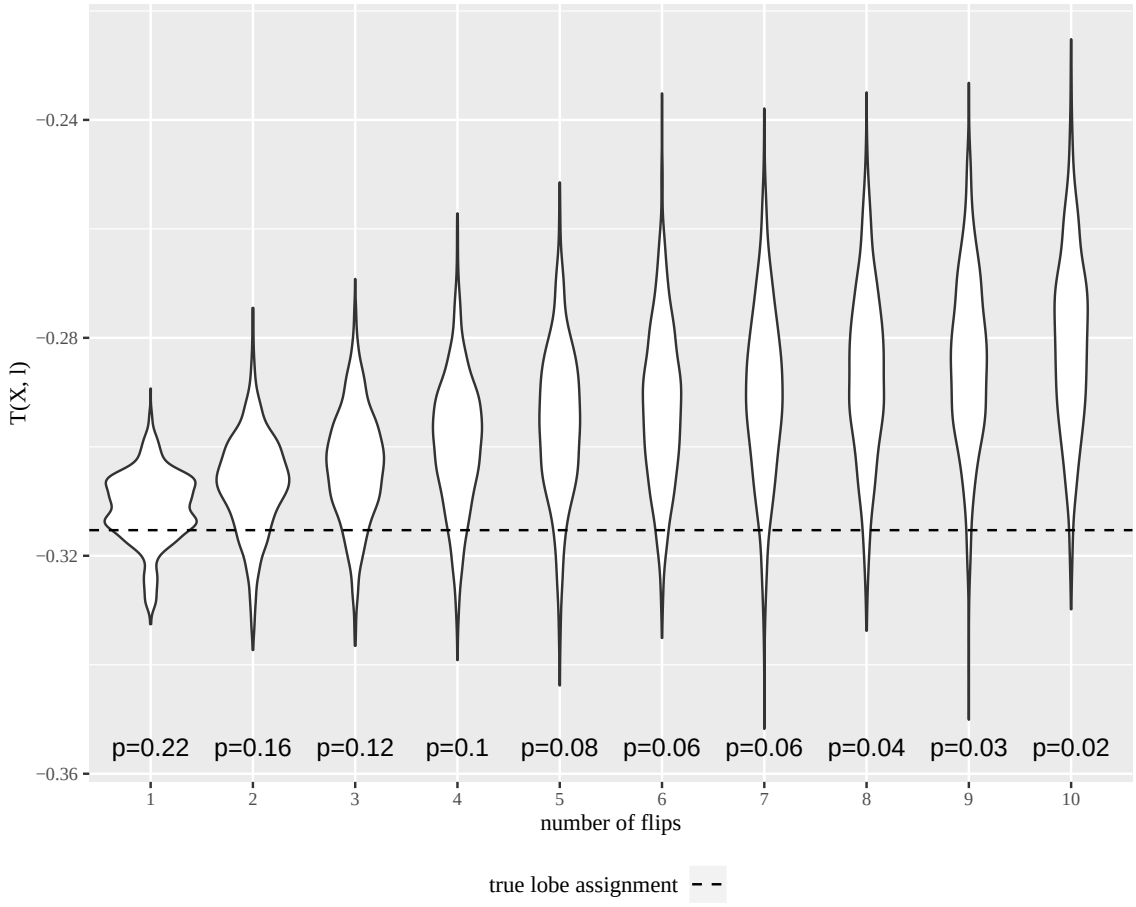


Figure 3.13: Violin plot of the permutation test. We run 1000 simulations for each number of flips. Dashed line represents the situation based on true lobe assignment. As the number of flips increases, $T(X, l)$ converges to the expected value under the assumption that X and l are independent in a constrained way mentioned above. And we can see that the $T(X, l)$ according to the true lobe assignment move away from the 95% confidence interval.

data analysis under a full-rank independent edge model where we used the sample mean of the 454 graphs in the Desikan dataset as the probability matrix P. As in the previous section, we simulated datasets from the full rank IEM distribution with probability matrix P of size $m = 1, 5$, and 10 and used $\bar{A}$ and $\widehat{P}$, where we varied the rank of $\widehat{P}$ from 1 to 70.

Figure 3.14 shows the resulting estimated MSE for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for simulated data based on the full rank probability matrix P shown in the left panel of Figure 3.1. We see that the results are very similar to those presented in Section 3.6, though overall $\widehat{P}$ performs even better than in the real data experiments. When m is small, $\widehat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including those selected by the Zhu and Ghodsi method. On the other hand, when m is large enough, both estimators perform well with the decision between the two being less conclusive. This simulation again shows the robustness of $\widehat{P}$ to deviations from the RDPG model, specifically if the probability matrix is full-rank.

We also note that the finite-sample relative efficiency in these cases shows is even more favorable to $\widehat{P}$, with relative efficiencies lower than $1/3$ for $m = 1$, than for the real data, where relative efficiencies were at best around $1/2$ for $m = 1$. From this observation, we can postulate that the degradation in the performance of $\widehat{P}$ in real data can at least partially be attributed to the fact that the independent edge assumption does not hold for real data.

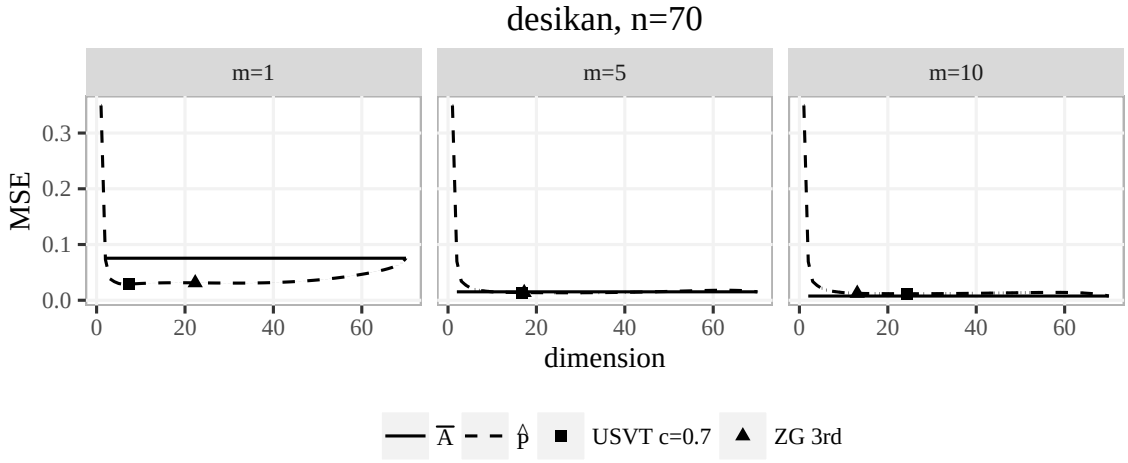


Figure 3.14: Comparison of $\widehat{P}$ and $\bar{A}$ for synthetic data analysis. As in Figure 3.4, this figure shows $\widehat{\text{MSE}}$ for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for simulated data with different sample sizes m based on the sample mean for the Desikan dataset. Again, the average of dimensions selected by the USVT method (square) and the ZG method (triangle) tend to nearly approximate the optimal dimension. Overall, we see that the structure of these plots well approximates the structure for the real data indicating that performance for the independent edge model will tend to translate in structure to non-independent edge scenarios. On the other hand, the relative efficiency $\widehat{\text{RE}}(\bar{A}, \widehat{P})$ is lower for this synthetic data analysis than for the CoRR data.

3.8 Appendix: Proofs for Theory Results

Here we present the proofs of the results in Section 3.4. To keep the ideas clear and concise, we leave out some details which are only slight changes to previous works. We assume the block memberships τ_i are drawn iid from a categorical distribution with block membership probabilities given by $\rho \in [0,1]^K$ where $\sum_i \rho_i = 1$. We will also assume that for a given n, the block memberships are fixed for all graphs.

Denote matrix of between-block edge probabilities by $B = \nu\nu^\top \in [0,1]^{K\times K}$ which we assume has rank K and is positive definite. By definition, the mean of the collection of graphs generated from this SBM is P, where $P_{ij} = B_{\tau_i,\tau_j}$.

We observe m graphs on n vertices $A^{(1)}, \cdots, A^{(m)}$ sampled independently from the SBM conditioned on τ. Define $\bar{A} = \frac{1}{m}\sum_{t=1}^m A^{(t)}$. Let $\widehat{U}\widehat{S}\widehat{U}^\top$ be the best rank-d positive semidefinite approximation of $\bar{A}$, then we define $\widehat{P} = \widehat{X}\widehat{X}^\top$, where $\widehat{X} = \widehat{U}\widehat{S}^{1/2}$.

The proofs presented here will rely on a central limit theorem developed in Athreya et al. [2016]. We modify the theorem slightly to account for the multiple graph setting and present it in the special case of the stochastic blockmodel.

Theorem 3.8.1 (Corrolary of Theorem 1 in Athreya et al. [2016]) *In the setting above, let $X = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{K\times d}$ have row i equal to $X_i = \nu_{\tau_i}$ (recall that τ_i are drawn from $[K]$ according to the probabilities ρ). Then there exists an orthogonal matrix W such that for each row i and j and any $z \in \mathbb{R}^d$, conditioned on $\tau_i = s$ and*

$\tau_j = t,$

$$\mathbb{P}\left\{\sqrt{n}(W\widehat{X}_i - \nu_s) \leq z, \sqrt{n}(W\widehat{X}_j - \nu_t) \leq z'\right\} = \Phi(z, \Sigma(\nu_s)/m)\Phi(z', \Sigma(\nu_t)/m) + o(1) \tag{3.3}$$

where $\Sigma(x) = \Delta^{-1}\mathbb{E}[X_j X_j^\top (x^\top X_j - (x^\top X_j)^2)]\Delta^{-1}$ *and* $\Delta = \mathbb{E}[X_1 X_1^T]$ *is the second moment matrix, with all expectations taken unconditionally. The function* Φ *is the cumulative distribution function for a multivariate normal with mean zero and the specified covariance, and* $o(1)$ *denotes a function that tends to zero as* $n \to \infty$.

The proof of this result follows very closely the proof of the result in the original paper with only slight modifications for the multiple graph setting.

We now prove a technical lemma which yields the simplified form for the variance under the stochastic blockmodel.

Lemma 3.8.2 *In the same setting as Theorem 3.4.3, for any* $1 \leq s, t \leq K$, *we have*

$$\nu_s^\top \Sigma(\nu_t)\nu_s \ = \frac{1}{\rho_s}\nu_s^\top \nu_t \ (1 - \nu_s^\top \nu_t).$$

Proof: Under the stochastic blockmodel with parameters (B, ρ), we have $X_i \overset{iid}{\sim} \sum_{k=1}^K \rho_k \delta_{\nu_k}$, where $\nu = [\nu_1, \cdots, \nu_K]^\top$ satisfies $B = \nu\nu^\top$. Without loss of generality, we could assume that $\nu = US$ where $U = [u_1, \cdots, u_K]^\top$ is orthonormal in columns and S is a diagonal matrix. Here we can conclude that $\nu_s^\top = u_s^\top S$. Defining $R =$

$\mathrm{diag}(\rho_1, \cdots, \rho_K)$, we have

$$\Delta = \mathbb{E}[X_1 X_1^\top] = \sum_{k=1}^{K} \rho_k \nu_k \nu_k^\top = \nu^\top R \nu = S U^\top R U S.$$

Thus

$$\begin{aligned}
\nu_s^\top \Sigma(\nu_t)\nu_s =& \nu_s^\top \Delta^{-1} \sum_{k=1}^{K} \rho_k \nu_k \nu_k^\top (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k)\Delta^{-1}\nu_s \\
=& \sum_{k=1}^{K} \rho_k (\nu_s^\top \Delta^{-1}\nu_k)(\nu_k^\top \Delta^{-1}\nu_s)(\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
=& \sum_{k=1}^{K} \rho_k (u_s^\top U^\top R^{-1} U u_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
=& \sum_{k=1}^{K} \rho_k (e_s^\top R^{-1} e_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
=& \sum_{k=1}^{K} \rho_k \delta_{sk} \rho_s^{-2} (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
=& \frac{1}{\rho_s} \nu_t^\top \nu_s (1 - \nu_t^\top \nu_s)
\end{aligned}$$

■

Lemma 3.8.3 (Lemma 3.4.2) *In the same setting as above, for any i, j, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$, we have*

$$\lim_{n\to\infty} n \cdot \mathrm{Var}(\widehat{P}_{ij}) = \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{m} P_{ij}(1 - P_{ij}).$$

And for n large enough, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$, we have

$$\mathbb{E}[(\widehat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{mn} P_{ij}(1 - P_{ij}).$$

Proof: Conditioned on $X_i = \nu_k$, we have by Theorem 3.8.1,

$$\mathbb{E}[W\widehat{X}_i] = \nu_k + o(1)$$

and

$$n \cdot \mathrm{Cov}(W\widehat{X}_i, W_n\widehat{X}_i) = \Sigma(\nu_k)/m.$$

Also, Corollary 3 in Athreya et al. [2016] says $\widehat{X}_i$ and $\widehat{X}_j$ are asymptotically independent. Thus, conditioning on $X_i = \nu_s$ and $X_j = \nu_t$, we have $\lim_{n\to\infty} \mathbb{E}[\widehat{X}_i^\top \widehat{X}_j] = \lim_{n\to\infty} \mathbb{E}[(W_n\widehat{X}_i)^\top]\mathbb{E}[W_n\widehat{X}_j] = \nu_s^\top \nu_t = P_{ij}$.

Since $\widehat{P}_{ij} = \widehat{X}_i^\top \widehat{X}_j$ is a noisy version of the dot product of $\nu_s^\top \nu_t$, combined with Lemma 3.8.2 and the results above, by Equation 5 in [Brown and Rutemiller, 1977], conditioning on $X_i = \nu_s$ and $X_j = \nu_t$, we have

$$\mathbb{E}[\widehat{X}_i^\top \widehat{X}_j] = \mathbb{E}[(W_n\widehat{X}_i)^\top]\mathbb{E}[W_n\widehat{X}_j] = \nu_s^\top \nu_t + o(1) = P_{ij} + o(1)$$

and

$$\begin{aligned}
& n \cdot \mathrm{Var}(\widehat{P}_{ij}) \\
=& \frac{1}{m}\left(\nu_s^\top \Sigma(\nu_t)\nu_s + \nu_t^\top \Sigma(\nu_s)\nu_t^\top\right) + \frac{1}{m^2 n}\left(tr(\Sigma(\nu_s)\Sigma(\nu_t))\right) + o(1) \\
=& \frac{1}{m}\left(\nu_s^\top \Sigma(\nu_t)\nu_s + \nu_t^\top \Sigma(\nu_s)\nu_t^\top\right) + o(1) \\
=& \frac{1/\rho_s + 1/\rho_t}{m} P_{ij}(1 - P_{ij}) + o(1).
\end{aligned}$$

Since $\widehat{P}_{ij} = \widehat{X}_i^\top \widehat{X}_j$ is asymptotically unbiased for P_{ij}, when n is large enough, we have

$$\mathbb{E}[(\widehat{P}_{ij} - P_{ij})^2] = \mathrm{Var}(\widehat{P}_{ij}) \approx \frac{1/\rho_s + 1/\rho_t}{mn} P_{ij}(1 - P_{ij}) + o(1).$$

■

Chapter 4

Robust Generalizations of the Law of Large Graphs

While Chapter 3 makes an effort to estimate the mean of a collection of unweighted graphs, we shift our focus to weighted graphs under a more general setting in this chapter. In the general parametric framework, $G \sim f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, and selecting a principled and productive estimator $\widehat{\theta}$ for the unknown graph parameter θ given a sample of graphs $\{G^{(1)}, \cdots, G^{(m)}\}$ is one of the most foundational and essential tasks, facilitating subsequent inference. For example, Ginestet et al. [2014] proposes a method to test for a difference between the networks of two groups of subjects in functional neuroimaging; while hypothesis testing is the ultimate goal, estimation is a key intermediate step. Note that this setting is more general since in Chapter 3, estimating the mean of a collection of unweighted graphs is equivalent

to estimating θ when $\mathcal{F}$ are Bernoulli distributions. We propose a widely-applicable, robust, low-rank estimation procedure for a collection of weighted graphs.

Consider for illustration the connectome data set made available through the Consortium for Reliability and Reproducibility[1] and investigated in Section 4.6 below. We have $m = 114$ brain graphs, each having $n = 70$ vertices representing different anatomical regions; the (errorfully observed) weight of an edge between two vertices represents the number of white-matter tracts connecting the corresponding two regions of the brain, as measured by diffusion tensor magnetic resonance imaging. Our goal in this situation is to estimate the average number of white-matter tracts between different regions of the brain. A more accurate estimate can lead to a better understanding of brain connectivity and hence functionality. Also, better estimates will improve performance on other tasks, such as diagnosis of brain disease.

The maximum likelihood estimate (MLE) – the edge-wise sample mean, without taking any graph structure into account, as in the (weighted extension of) the independent edge graph model (IEM) [Bollobás et al., 2007] (described in Section 2.3.1) – is a natural candidate for our estimation problem. However, the MLE suffers from at least two major deficiencies in our setting: high variance and non-robustness.

In our high dimensional setting (a large number of vertices, n), the edge-wise MLE leads to estimates with unacceptably high variance unless the sample size (the number of graphs, m) is exceedingly large. However, if the graphs can be assumed

[1] `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/`

to be (approximately) low-rank, then by biasing towards low-rank structure, more elaborate estimators can have greatly reduced variance and win the bias-variance tradeoff, as discussed in Chapter 3. For our connectome data in Section 4.6 we observe this approximate low-rank property. Tang et al. [2016] develops an estimator based on a low-rank approximation and proves that this new estimator outperforms the edge-wise MLE, decreasing the overall asymptotic variance dramatically by smoothing towards the low-rank structure, which is discussed in Chapter 3.

The second edge-wise MLE deficiency in our setting derives from the edge observations being subject to contamination. That is, the weights attributed to edges are possibly observed with noise. The sample mean is notoriously un-robust to outliers; thus, under the possibility of contamination, it is wise to use robust methods, such as the MLqE [Ferrari and Yang, 2010, **?**] considered in this paper.

To address these two deficiencies simultaneously, we propose an estimation methodology which is a natural extension of [Tang et al., 2016] to gross error contamination. Our proposed estimator both inherits MLqE robustness and wins the bias-variance tradeoff by taking advantage of low-rank structure.

We organize the chapter as follows. In Section 4.1, we define the gross error contamination model we will consider based on WSBM in a WRDPG setting. In Section 4.2, we present our estimation methodology in terms of two estimators designed to address the two edge-wise MLE deficiencies described above, and we construct our final estimator by combining the two estimators. In Section 4.3, we prove that

our estimator is superior, under appropriate conditions, and this result is generalized in Section 4.4. In Section 4.5 and Section 4.6, we illustrate the performance of our estimator through experimental results on simulated and real data.

4.1 Contamination Model

In this chapter, we are in the scenario where m weighted graphs on n vertices are given in the adjacency matrices form $\{A^{(t)}\}(t = 1, \ldots, m)$. Again, the graphs are undirected without self-loops, i.e. each $A^{(t)}$ is symmetric with zeros along the diagonal. Moreover, we assume the vertex correspondence is known across different graphs, so that vertex i of the t_1-th graph corresponds to vertex i of the t_2-th graph for any $i \in [n]$, $t_1, t_2 \in [m]$.

In practice, we can hardly get data accurately. So there will always be noise in the observations, which deviates from our general model assumptions. In order to incorporate this effect, a contamination model, the gross error model [Bickel and Doksum, 2007, Mah and Tamhane, 1982], is considered in this work.

Generally in a gross error model, we observe good measurement $G^* \sim f_P \in \mathcal{F}$ most of the time, while there are a few wild values $G^{**} \sim h_C \in \mathcal{H}$ when the gross errors occur. Here P and C represent the respective parameter matrices of the two distribution families. As to the graphs, one way to generalize from the gross error model is to contaminate the entire graph with some small probability $\epsilon \in (0, 1)$, that

is $G \sim (1-\epsilon)f_P + \epsilon h_C$. However, since all the models we consider are subsets of the WIEM, it is more natural to consider the contaminations with respect to each edge, i.e. for $1 \leq i < j \leq n$, $G_{ij} \sim (1-\epsilon)f_{P_{ij}} + \epsilon h_{C_{ij}}$ with $f \in \mathcal{F}$ and $h \in \mathcal{H}$, where both $\mathcal{F}$ and $\mathcal{H}$ are one-parameter distribution families.

In this chapter, we assume that when gross errors occur, the weights of the edges are also from the same one-parameter family $\mathcal{F}$. Moreover, we also assume that the connectivity follows the WSBM as a WRDPG. Thus, similar to the uncontaminated distribution $f_{P_{ij}}$ with $P_{ij} = B_{\tau_i,\tau_j}$ where B is the block probability matrix and τ is the block assignments, the contamination distribution $f_{C_{ij}}$ with $C_{ij} = B'_{\tau'_i,\tau'_j}$ also have the block structure, where B' is the block probability matrix and τ' is the block assignments. For clarity, we will introduce the sampling procedure when the contamination has the same block structure, i.e. $\tau = \tau'$. However, it is not required in our theories.

To generate m graphs under this contamination model with known vertex correspondence, we first sample τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs as in Section 3.1. Then m symmetric and hollow graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A^{(t)}_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ for $t \in [m]$, $i, j \in [n]$ and $i < j$, where $P_{ij} = B_{\tau_i,\tau_j}$ and $C_{ij} = B'_{\tau_i,\tau_j}$. Here ϵ is the probability of an edge to be contaminated, P is the parameter matrix as in Section 2.3.3, and C is the parameter matrix for contaminations.

4.2 Estimators

Under any model introduced in Section 3.1, our goal is always to estimate the parameter matrix P based on the m observations $A^{(1)}, \ldots, A^{(m)}$. Especially when under the contamination model, although there are other parameters like ϵ and C, our goal is still to estimate the uncontaminated parameter matrix P. In this section, we present four estimators as in Figure 4.1, i.e. the standard entry-wise MLE $\widehat{P}^{(1)}$, the low-rank approximation of the entry-wise MLE $\widetilde{P}^{(1)}$, the entry-wise robust estimator MLqE $\widehat{P}^{(q)}$, and the low-rank approximation of the entry-wise MLqE $\widetilde{P}^{(q)}$. Since the observed graphs are symmetric and hollow with a symmetric parameter matrix of the model, we do not care about the estimate of the diagonal of P. However, the estimate itself should be at least symmetric.

4.2.1 Entry-wise Maximum Likelihood Estimator $\widehat{P}^{(1)}$

Under WIEM, the most natural estimator is the MLE, which happens to be the element-wise MLE $\widehat{P}^{(1)}$ in this case. Note that this is $\bar{A}$ in Chapter 3. Moreover, when $\mathcal{F}$ is a one-parameter exponential family, for instance Bernoulli distributions or Exponential distributions, the entry-wise MLE $\widehat{P}^{(1)}$ is uniformly minimum-variance unbiased estimator, i.e. it has the smallest variance among all unbiased estimators. In addition, it satisfies many good asymptotic properties as the number of graphs

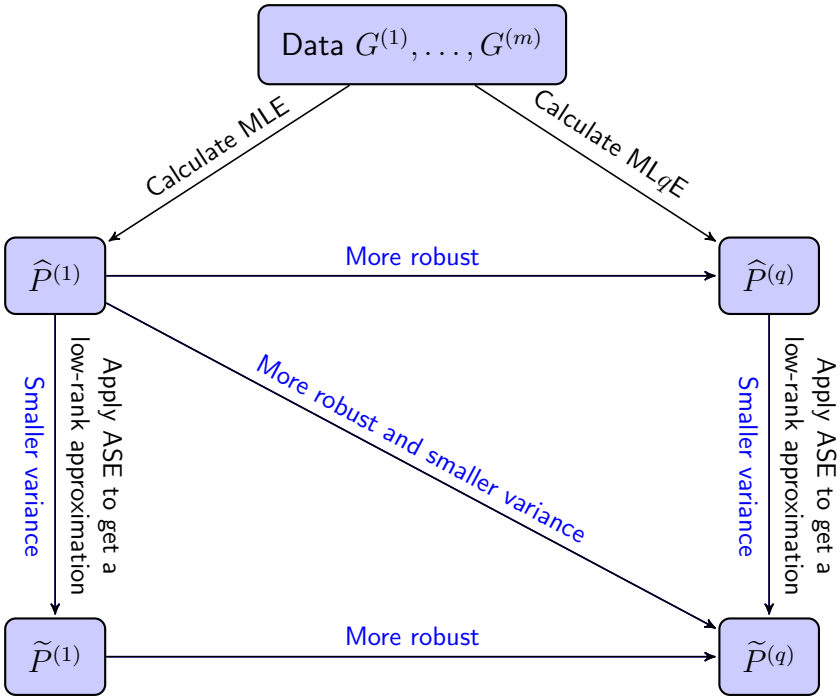


Figure 4.1: Roadmap among the data and four estimators.

m goes to infinity. However, in high dimensional situations like this, the entry-wise MLE often leads to inaccurate estimates with very high variance when the sample size m is small. Also, it does not exploit any graph structure. The performance will not get any better when the number of vertices in each graph n increases since it is an entry-wise estimator. Moreover, if the graphs are actually distributed under a WRDPG or a WSBM, then the entry-wise MLE is no longer the MLE any more and the performance can be very poor.

4.2.2 Estimator $\widetilde{P}^{(1)}$ Based on Adjacency Spectral Embedding of $\widehat{P}^{(1)}$

Motivated by the low-rank structure of the parameter matrix P in WRDPG, we consider the estimator $\widetilde{P}^{(1)}$ proposed by Tang et al. [2016] based on the spectral decomposition of $\widehat{P}^{(1)}$, i.e. $\widehat{P}$ in Chapter 3. Dimension selection technique discussed in Section 3.2.2 and the diagonal augmentation procedure introduced in in Section 3.2.3 will also be used in this section. The construction procedure of $\widetilde{P}^{(1)}$ consists of several steps, which will be introduced respectively in the following subsections.

4.2.2.1 Rank-d Approximation

Given a dimension d, we consider $\widetilde{P}^{(1)} = \text{lowrank}_d(\widehat{P}^{(1)})$ as the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$. To find such best approximation, first calculate

the eigen-decomposition of the symmetric matrix $\widehat{P}^{(1)} = \widehat{U}\widehat{S}\widehat{U}^\top + \widetilde{U}\widetilde{S}\widetilde{U}^\top$, where $\widehat{S}$ is the diagonal matrix with the largest d eigenvalues of $\widehat{P}^{(1)}$, and $\widehat{U}$ has the corresponding eigenvectors as each column. Similarly, $\widetilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n-d$ eigenvalues of $\widehat{P}^{(1)}$, and $\widetilde{U}$ has the columns given by the corresponding eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\widehat{P}^{(1)}$ is given by $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n\times d}$, which follows Definition 3.2.1. Based on the ASE result, we have the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$ to be $\widetilde{P}^{(1)} = \widehat{X}\widehat{X}^\top = \widehat{U}\widehat{S}\widehat{U}^\top$. In the RDPG setting, Sussman et al. [2014] proved that each row of $\widehat{X}$ can accurately estimate the the latent position for each vertex up to an orthogonal transformation. We will analyze its performance under the contaminated WRDPG setting in Section 4.3.

Here, we restate the algorithm in [Tang et al., 2016] (also mentioned in Chapter 3) to give the detailed steps of computing this low-rank approximation of a general n-by-n symmetric matrix A in Algorithm 3.

Algorithm 3 Algorithm to compute the rank-d approximation of a matrix.

Require: Symmetric matrix $A \in \mathbb{R}^{n\times n}$ and dimension $d \leq n$.

Ensure: $\text{lowrank}_d(A) \in \mathbb{R}^{n\times n}$

1: Compute the algebraically largest d eigenvalues of A, $s_1 \geq s_2 \geq \ldots \geq s_d$ and corresponding unit-norm eigenvectors $u_1, u_2, \ldots, u_d \in \mathbb{R}^n$;

2: Set $\widehat{S}$ to the $d \times d$ diagonal matrix $\text{diag}(s_1, \ldots, s_d)$;

3: Set $\widehat{U} = [u_1, \ldots, u_d] \in \mathbb{R}^{n\times d}$;

4: Set $\text{lowrank}_d(A)$ to $\widehat{U}\widehat{S}\widehat{U}^\top$;

By combining the key parts introduced above, we give the detailed description for calculating the estimator $\widetilde{P}^{(1)}$ with dimension selection method in Algorithm 4.

Algorithm 4 Algorithm to compute $\widetilde{P}^{(1)}$

Require: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \mathbb{R}^{n\times n}$

Ensure: Estimate $\widetilde{P}^{(1)} \in \mathbb{R}^{n\times n}$

1: Calculate the entry-wise MLE $\widehat{P}^{(1)}$;

2: Select the dimension d based on the eigenvalues of $\widehat{P}^{(1)}$; (see Section 3.2.2)

3: Set Q to $\text{lowrank}_d(\widehat{P}^{(1)})$; (see Algorithm 3)

4: Set $\widetilde{P}^{(1)}$ with each entry $\widetilde{P}^{(1)}_{ij} = \max(Q_{ij}, 0)$.

4.2.3 Entry-wise Maximum Lq-likelihood Estimator $\widehat{P}^{(q)}$

The MLE is asymptotically efficient, i.e. when sample size is large enough, the MLE is at least as accurate as any other estimator. However, when the sample size is moderate, robust estimators always outperforms MLE in terms of mean squared error by winning the bias-variance tradeoff. Moreover, under contamination models, robust estimators can even beat MLE asymptotically since they are designed to be not unduly affected by the outliers. And now we are going to consider one robust estimator, i.e. the maximum Lq-likelihood estimator (MLqE) proposed by Ferrari and Yang [2010].

Definition 4.2.1 (MLqE) *Let $X_1, \ldots, X_m$ be sampled from $f_{\theta_0} \in \mathcal{F} = \{f_\theta, \theta \in \Theta\}$, $\theta_0 \in \Theta$. Then the* maximum Lq-likelihood estimate *($q > 0$) of θ_0 based on the parametric model $\mathcal{F}$ is defined as*

$$\widehat{\theta}_{\mathrm{ML}q\mathrm{E}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{m} L_q[f_\theta(X_i)],$$

where $L_q(u) = (u^{1-q} - 1)/(1 - q)$.

Note that $L_q(u) \to \log(u)$ when $q \to 1$. Thus MLqE is a generalization of MLE. Moreover, define

$$U_\theta(x) = \nabla_\theta \log f_\theta(x)$$

and

$$U_\theta^\star(x;q) = U_\theta(x) f_\theta(x)^{1-q}.$$

Then the MLqE $\widehat{\theta}_{\mathrm{ML}q\mathrm{E}}$ can also be seen as a solution to the equation

$$\sum_{i=1}^{m} U_\theta^\star(X_i;q) = 0.$$

This form interprets $\widehat{\theta}_{\mathrm{ML}q\mathrm{E}}$ as a solution to the weighted likelihood equation. The weights $f_\theta(x)^{1-q}$ are proportional to the $(1-q)$th power of the corresponding probability. Specifically, when $0 < q < 1$, the MLqE puts less weight on the data points which do not fit the current distribution well. Equal weights happens when $q = 1$ and lead to the MLE.

Under the WIEM, we can calculate the robust entry-wise MLqE $\widehat{P}^{(q)}$ based on the adjacency matrices $A^{(1)}, \ldots, A^{(m)}$. Note that $\widehat{P}^{(1)}$, the entry-wise MLE, is a special case of entry-wise MLqE $\widehat{P}^{(q)}$ when $q = 1$. That is what the superscriptions q and 1 mean. There is also a bias-variance tradeoff in selecting the parameter q. Qin and Priebe [2017] proposed a way to select q in general. In this work, we do not focus on how to select q.

4.2.4 Estimator $\widetilde{P}^{(q)}$ Based on Adjacency Spectral Embedding $\widehat{P}^{(q)}$

Intuitively, the low-rank structure of the parameter matrix P in WRDPG should be preserved more or less in the entry-wise MLqE $\widehat{P}^{(q)}$. Thus, in order to take advantage of such low-rank structure as well as the robustness, we apply the similar idea here as in building $\widetilde{P}^{(1)}$, i.e. enforce a low-rank approximation on the entry-wise MLqE matrix $\widehat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$. As in Algorithm 4, we apply the same dimension selection method and diagonal augmentation procedure. The only change is to substitute $\widehat{P}^{(1)}$ by $\widehat{P}^{(q)}$. The details of the algorithm is shown in Algorithm 5.

Algorithm 5 Algorithm to compute $\widetilde{P}^{(q)}$

Require: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \mathbb{R}^{n\times n}$

Ensure: Estimate $\widetilde{P}^{(q)} \in \mathbb{R}^{n\times n}$

1: Calculate the entry-wise MLqE $\widehat{P}^{(q)}$;

2: Select the dimension d based on the eigenvalues of $\widehat{P}^{(q)}$; (see Section 3.2.2)

3: Set Q to $\text{lowrank}_d(\widehat{P}^{(q)})$; (see Algorithm 3)

4: Set $\widetilde{P}^{(q)}$ with each entry $\widetilde{P}^{(q)}_{ij} = \max(Q_{ij}, 0)$.

4.3 Theoretical Results

In this section, for illustrative purpose, we are going to present theoretical results when the contamination model introduced in Section 4.1 is with respect to exponential distributions. That is $\mathcal{F} = \{f_\theta(x) = \frac{1}{\theta}e^{-x/\theta}, \theta \in [0, R] \subset \mathbb{R}\}$, where $R > 0$ is a constant. The results can be extended to a general situation with proper assumptions, which will be discussed in Section 4.4.

For clarity, we restate the model settings discussed in Section 4.1. Consider the SBM with parameter B and ρ. First sample the block membership τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs. Conditioned on this τ we sampled, the probability matrix P then satisfies $P_{ij} = B_{\tau_i,\tau_j}$. In this section, we assume the contamination has the same block membership τ, thus the contamination matrix $C \in \mathbb{R}^{n\times n}$ has the same block structure as P. Note that this is not necessary for the result. Different block structure can lead to the same result since the rank is still finite. Denote ϵ as the probability of an edge to be contaminated. Then m symmetric graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A_{ij}^{(t)} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ for $t \in [m]$, $i, j \in [n]$ and $i < j$.

Under such setting, we now analyze the performance of all four estimators introduced in Section 4.2 based on m adjacency matrices for estimating the probability matrix P in terms of the mean squared error. When comparing two estimators, we mainly focus on both asymptotic bias and asymptotic variance. Note that all the

results in this section are entry-wise, which can easily lead to a result of the total MSE for the entire matrix.

We only present the main results in this section. The proofs are given in Section 4.7.

4.3.1 $\widehat{P}^{(1)}$ vs. $\widehat{P}^{(q)}$

We first compare the performance between the entry-wise MLE $\widehat{P}^{(1)}$ and the entry-wise MLqE $\widehat{P}^{(q)}$. Without using the graphs structure, the asymptotic results for these two estimators are in terms of the number of graphs m, not the number of vertices n within each graph.

Theorem 4.3.1 *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

$$\lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(q)}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|,$$

for $1 \le i, j \le n$ and $i \neq j$. Moreover, without any assumption on the contaminated model, for $1 \le i, j \le n$,

$$\mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(1/m).$$

And thus

$$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0.$$

Theorem 4.3.1 shows that the entry-wise MLqE $\widehat{P}^{(q)}$ has smaller bias for estimating P asymptotically compared to the entry-wise MLE $\widehat{P}^{(1)}$. Although we put restrictions on the parameter matrix C in the statement of the theorem, the result still holds provided that $\epsilon(C_{ij}-P_{ij}) > (1-q)P_{ij}$. This condition only requires the contamination of the model is large enough (either large contamination parameter matrix, or more likely to encounter an outlier). From a different perspective, it also requires $\widehat{P}^{(q)}$ to be robust enough with respect to the contamination. Thus besides the current condition for C, equivalently, we can also replace it by the assumption of a large enough ϵ or a small enough q.

Theorem 4.3.1 also indicates that both estimators have variances converge to zero as the number of graphs m goes to infinity, following the asymptotic properties of minimum contrast estimates. Thus the bias term will dominate in the comparison in terms of MSE.

As a result, $\widehat{P}^{(q)}$ reduces the bias while keeping variance the same asymptotically compared to $\widehat{P}^{(1)}$. Thus in terms of MSE, $\widehat{P}^{(q)}$ is a better estimator than $\widehat{P}^{(1)}$ when the number of graphs m is large with enough contamination.

4.3.2 $\widehat{P}^{(1)}$ vs. $\widetilde{P}^{(1)}$

We next analyze the effect of the ASE procedure applied to the entry-wise MLE $\widehat{P}^{(1)}$ under the contamination model, so that we can compare the performance between $\widehat{P}^{(1)}$ and $\widetilde{P}^{(1)}$.

Before proceeding to the comparison between the two estimators, we first recall the definition of the asymptotic relative efficiency (ARE) [Serfling, 2011], which is a very important and useful criterion to compare two estimators. Note that the original definition is for unbiased estimators. Here we adapt the definition to estimators with the same asymptotic bias.

Definition 4.3.2 (Asymptotic Relative Efficiency) *For any parameter θ of a distribution f, and for estimators $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ such that $E[\widehat{\theta}^{(1)}] = E[\widehat{\theta}^{(2)}] = \theta'$, $n \cdot \mathrm{Var}(\widehat{\theta}^{(1)}) \to V_1(f)$ and $n \cdot \mathrm{Var}(\widehat{\theta}^{(2)}) \to V_2(f)$, the* Asymptotic Relative Efficiency (ARE) *of $\widehat{\theta}^{(2)}$ to $\widehat{\theta}^{(1)}$ is given by*

$$\mathrm{ARE}(\widehat{\theta}^{(2)}, \widehat{\theta}^{(1)}) = \frac{V_1(f)}{V_2(f)}.$$

By the definition above, if $\mathrm{ARE}(\widehat{\theta}^{(2)}, \widehat{\theta}^{(1)}) < 1$, then $\widehat{\theta}^{(1)}$ has a smaller variance in its sampling distribution and thus is more efficient compared to $\widehat{\theta}^{(2)}$. Combine with the fact that both estimators have the same asymptotic bias, $\widehat{\theta}^{(1)}$ is a better estimate in this case.

To compare $\widehat{P}^{(1)}$ and $\widetilde{P}^{(1)}$, we will first show they have the same entry-wise asymp-

totic bias under proper conditions, and then use the ARE criterion to compare the performance in the following theorem.

Theorem 4.3.3 *Assuming that $m = O(n^b)$ for any $b > 0$, then*

$$\lim_{n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(1)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} \text{Bias}(\widehat{P}_{ij}^{(1)}).$$

In addition, for $1 \leq i, j \leq n$ and $i \neq j$,

$$\text{Var}(\widetilde{P}_{ij}^{(1)}) = O(m^{-1}n^{-1}(\log n)^3), \text{Var}(\widehat{P}_{ij}^{(1)}) = O(m^{-1}).$$

And thus

$$\frac{\text{Var}(\widetilde{P}_{ij}^{(1)})}{\text{Var}(\widehat{P}_{ij}^{(1)})} = O(n^{-1}(\log n)^3),$$

$$\text{ARE}(\widehat{P}_{ij}^{(1)}, \widetilde{P}_{ij}^{(1)}) = 0.$$

Theorem 4.3.3 says that when m is a constant, or m is going to infinity with order $m = O(n^b)$ for any $b > 0$, i.e. m is fixed or it grows not faster than any polynomial with respect to n, the ASE procedure applied to $\widehat{P}^{(1)}$ will not affect the asymptotic bias for estimating P. Combined with the fact that the ratio of the variances of two estimators is of order $O(n^{-1}(\log n)^3)$, we have that ARE goes to 0 when $n \to \infty$. Thus $\widetilde{P}_{ij}^{(1)}$ is much better than $\widehat{P}_{ij}^{(1)}$ for a large n. We emphasize that the order of the ratio of the variances does not depend on m.

As a result, the ASE procedure applied to the entry-wise MLE $\widehat{P}^{(1)}$ helps reduce

the variance while keeping the bias unchanged asymptotically, leading to a better estimate $\widetilde{P}^{(1)}$ for P in terms of MSE.

4.3.3 $\widehat{P}^{(q)}$ vs. $\widetilde{P}^{(q)}$

We now proceed to analyze the effect of the ASE procedure applied to the entry-wise MLqE $\widehat{P}^{(q)}$ under the contamination model in order to compare the performance between $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$. Similarly, we first show that the two estimators have the same entry-wise asymptotic bias under proper conditions, and then use the ARE criterion to compare the performance in the following theorem.

Theorem 4.3.4 *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has the same entry-wise asymptotic bias as MLqE, i.e.*

$$\lim_{n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} \text{Bias}(\widehat{P}_{ij}^{(q)}).$$

In addition, for $1 \leq i, j \leq n$ and $i \neq j$,

$$\text{Var}(\widetilde{P}_{ij}^{(q)}) = O(n^{-1}(\log n)^3), \text{Var}(\widehat{P}_{ij}^{(q)}) = O(m^{-1}).$$

And thus

$$\frac{\text{Var}(\widetilde{P}_{ij}^{(q)})}{\text{Var}(\widehat{P}_{ij}^{(q)})} = O(mn^{-1}(\log n)^3).$$

Moreover, if $m = o(n(\log n)^{-3})$*, then*

$$\mathrm{ARE}(\widehat{P}_{ij}^{(q)}, \widetilde{P}_{ij}^{(q)}) = 0.$$

The proof for Theorem 4.3.4 is almost the same as the proof for Theorem 4.3.3. But unlike the results for MLE, we are missing the term m^{-1} in the variance bound $\mathrm{Var}(\widetilde{P}^{(q)}) = O(n^{-1}(\log n)^3)$ due to the structure of maximum Lq likelihood equation. As a result, while the ASE procedure still does not affect the asymptotic bias, the ARE has an extra term m. This leads to a slight difference in the comparison. Specifically, when m is fixed, the order of the ARE is $O(n^{-1}(\log n)^3)$, which will goes to 0 as $n \to \infty$. Even if m also increases as n increases, as long as it grows in the order of $o(n(\log n)^{-3})$, the ARE still goes to 0.

Thus the ASE procedure applied to the entry-wise MLqE $\widehat{P}^{(q)}$ also helps reduce the variance while keeping the bias asymptotically, leading to a better estimate $\widetilde{P}^{(q)}$ for P in terms of MSE.

4.3.4 $\widetilde{P}^{(1)}$ vs. $\widetilde{P}^{(q)}$

To finish the last piece, we compare the performance between $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$ by combining the previous results.

Theorem 4.3.5 *For sufficiently large* C *and any* $1 \le i, j \le n$*, if* $m = O(n^b)$ *for any*

$b > 0$, then

$$\lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(1)}) > \lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)})$$

Moreover, if $m = O(n(\log n)^{-3})$, then

$$\lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(q)}) = 0.$$

Theorem 4.3.5 is a direct result of Theorem 4.3.1, Theorem 4.3.3, and Theorem 4.3.4. It concludes that $\widetilde{P}^{(q)}$ inherits the robustness from the entry-wise MLqE $\widehat{P}^{(q)}$ and has a smaller asymptotic bias compared to $\widetilde{P}^{(1)}$ while both estimates have variance goes to 0 as $m \to \infty$. Thus in summary, $\widetilde{P}^{(q)}$ is the best among all four estimators.

4.3.5 Summary

We summarize all the four estimators and their relationship in Figure 4.2. From top to bottom of the figure, we apply ASE to construct low-rank approximations which preserve the asymptotic bias and reduce the asymptotic variance. From left to right, we underweight the outliers to construct robust estimators. So with enough contaminations, whenever the number of graphs m is large enough, the bias term which dominates the MSE will be improved.

In conclusion, when contamination is relatively large as well as m and n, $\widetilde{P}^{(q)}$ is the best among the four estimators.

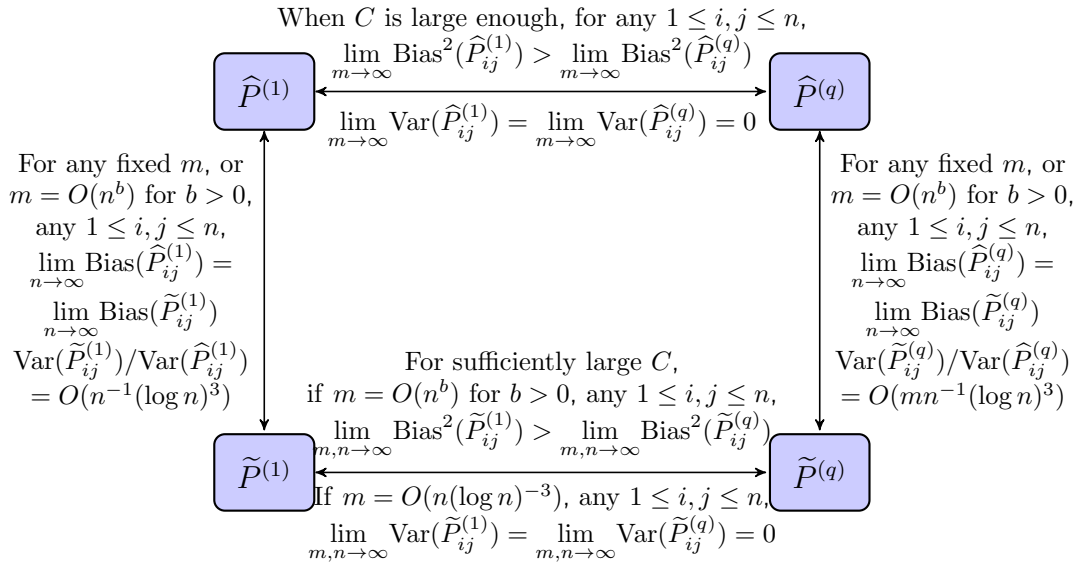


Figure 4.2: Relationship among four estimators.

4.4 Extensions

Results in Section 4.3 are presented in the setting of exponential distributions with MLqE estimator. However, the results can be generalized to a broader class of distribution families, and even a different entry-wise robust estimator (denoted as $\widehat{P}^{(R)}$) other than MLqE provided that the following conditions are satisfied:

1. Let $A_{ij} \overset{ind}{\sim} (1-\epsilon) f_{P_{ij}} + \epsilon f_{C_{ij}}$, then $E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \le \text{const}^k \cdot k!$, where $\widehat{P}^{(1)}$ is the entry-wise MLE as defined before;

2. There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,

$$\lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(R)}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|;$$

3. $\widehat{P}_{ij}^{(R)} \leq \text{const} \cdot \widehat{P}_{ij}^{(1)}$;

4. $\text{Var}(\widehat{P}_{ij}^{(R)}) = O(m^{-1})$, where m is the number of observations.

Condition 1 is to ensure that observations will not deviate too far away from the expectation, so that the concentration inequalities hold; Condition 2 is discussed in Section 4.3.1. It requires the contamination of the model to be large enough (a restriction on the distribution) and $\widehat{P}$ to be sufficiently robust with respect to the contamination (a condition on the estimator); By taking advantage of Condition 1 which controls $\widehat{P}^{(1)}$, Condition 3 reuses Condition 1 to bound an arbitrary $\widehat{P}^{(R)}$; Condition 4 is to ensure that the variance of $\widehat{P}_{ij}^{(R)}$ is comparable to the variance of the entry-wise MLE $\widehat{P}_{ij}^{(1)}$, which is of order $O(m^{-1})$. Nevertheless, should the variance be bigger, similar but weaker results can still be derived.

As an example of the distribution satisfying the above four conditions other than the exponential distribution mentioned in Section 4.3, we sketch the Poisson distribution as following. Poisson distribution is a commonly used distribution for nonnegative graphs with integer values. Lemma 4.7.34 verifies Condition 1. Intuitively, since exponential distribution has a fatter tail compare to Poisson, we should have the bound for central moment of Poisson directly from the results for exponential distribution. Condition 2 is satisfied when it is the robust MLqE. For Condition 3, $\widehat{P}_{ij}^{(R)}/\widehat{P}_{ij}^{(1)}$ is maximized when there are m data $x_1, \cdots, x_m$ with $0 \leq x_1 = \cdots = x_k \leq \bar{x} \leq x_{k+1} = \cdots = x_m \leq m\bar{x}/(m-k)$. In order to have ML$q$E larger than MLE $\bar{x}$, we need the weights of the first m data to be smaller than the

weights of the rest $m-k$ data. So $e^{-\bar{x}} < \bar{x}^{x_m} e^{-\bar{x}}/x_m!$. Then $x_m! < \bar{x}^{x_m}$. By the lower bound in Stirling's formula, we have $x_m < e\bar{x}$ when $x_m > 0$. Note that if $x_m = 0$ then MLE equals MLqE since all data equals zero. Thus MLqE is bounded by $e\bar{x}$. As a result, $\widehat{P}_{ij} \leq e\widehat{P}_{ij}^{(1)}$ and Condition 3 is satisfied. At last, Condition 4 follows directly from theory of minimum contrast estimators.

In summary, all theorems in Section 4.3 hold for the Poisson distribution. This section provides a general way to extend the theory to proper models and robust estimators.

4.5 Simulations

In this section, we first illustrate the theoretical comparison among the four estimators discussed in Section 4.3 via various Monte Carlo simulation experiments in an idealized setting.

4.5.1 Simulation Setting

Here we consider the 2-block SBM with respect to the exponential distributions parameterized by

$$B = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

Let the contamination also be a 2-block SBM with the same structure parameterized by

$$B' = \begin{bmatrix} 9 & 6 \\ 6 & 13 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

With these parameters specified, we sample graphs according to Section 4.1.

For the ease of presentation, in the simulation, we assume the true dimension $d = \text{rank}(B) = 2$ is known. So we ignore the dimension selection step in Algorithm 4 and Algorithm 5.

As suggested in [Tang et al., 2016], in this experiment we are going to combine both ideas by first using Marchette's row-averaging method and then another one-step Scheinerman's iterative method.

4.5.2 Simulation Results

In order to see how the performance of the four estimators varies with respect to the contaminations, we first run 1000 Monte Carlo replicates based on the contaminated SBM specified in Section 4.5.1 with a fixed number of vertices $n = 100$ and a fixed number of graphs $m = 20$ while varying the contamination probability ϵ from 0 to 0.4. Given each sample, four estimators can be calculated following Algorithm 4 and Algorithm 5. Since we are not focusing on how to select the parameter q in the MLqE estimator, we are going to use a fixed $q = 0.9$ throughout this paper. Then the MSE of each estimator can be estimated since the probability matrix P is known

in this simulation.

The results are given in Figure 4.3. Different colors represent the simulated MSE associated with four different estimators. Firstly, we see MLE $\widehat{P}^{(1)}$ is the best estimator when there is little or no contamination (i.e. ϵ is small or $\epsilon = 0$); however it degrades dramatically as contamination increases. On the contrary, the MLqE $\widehat{P}^{(q)}$ is slightly less efficient than the MLE $\widehat{P}^{(1)}$ when the contamination is small, but is much more robust under a large contamination compared to the MLE. Next we see that even with a relative small number of vertices $n = 100$, the ASE procedure which takes advantage of the low rank structure already helps improve the performance of $\widehat{P}^{(1)}$ and let $\widetilde{P}^{(1)}$ win the bias-variance tradeoff. Since the MLqE $\widehat{P}^{(q)}$ preserves the low rank structure of the original graph more or less, the ASE procedure also helps and makes $\widetilde{P}^{(q)}$ a better estimate. Although both $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$ take advantage of the low-rank structure and has reduced variances, $\widetilde{P}^{(q)}$ constructed based on MLqE inherits the robustness from MLqE in addition. So when the contamination is large enough, $\widetilde{P}^{(q)}$ outperforms $\widetilde{P}^{(1)}$ and degrades slower.

Figure 4.4 shows additional simulation results by varying the parameter q in MLqE with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different types of lines represent the simulated MSE associated with four different estimators. From the figure, we can see that the ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators for a wide range of q. Moreover, within a large range of q, the MLqE wins the bias-

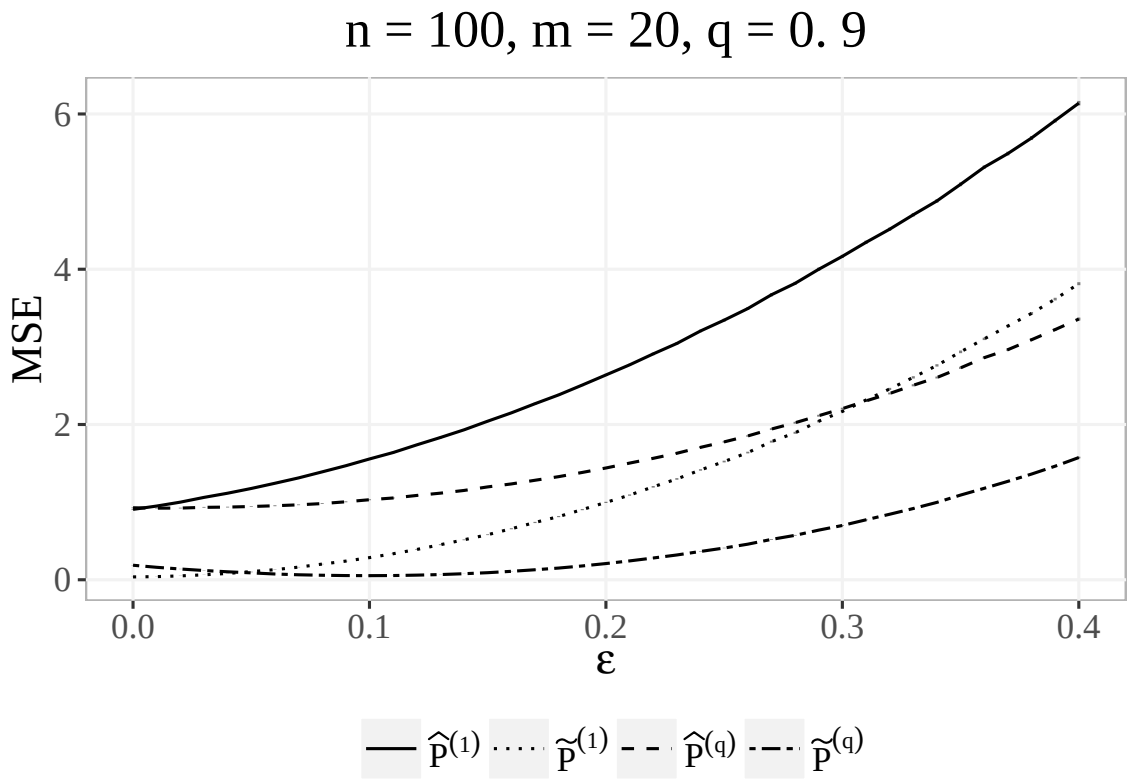


Figure 4.3: Mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 20$ based on 1000 Monte Carlo replicates. And we use $q = 0.9$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. 1. MLE $\widehat{P}^{(1)}$ vs MLqE $\widehat{P}^{(q)}$: MLE outperforms a little bit when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination increases; 2. MLE $\widehat{P}^{(1)}$ vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$: ASE procedure takes the low rank structure into account and $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff; 3. MLqE $\widehat{P}^{(q)}$ vs ASE $\circ$ MLqE $\widetilde{P}^{(q)}$: MLqE preserves the low rank structure of the original graph more or less, so ASE procedure still helps and $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff; 4. ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$: When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

variance tradeoff and shows the robustness property compare to the MLE. And as q goes to 1, MLqE goes to the MLE as expected.

By comparing the performance of the four estimators based on different setting, we demonstrate the theoretical results in Section 4.3.

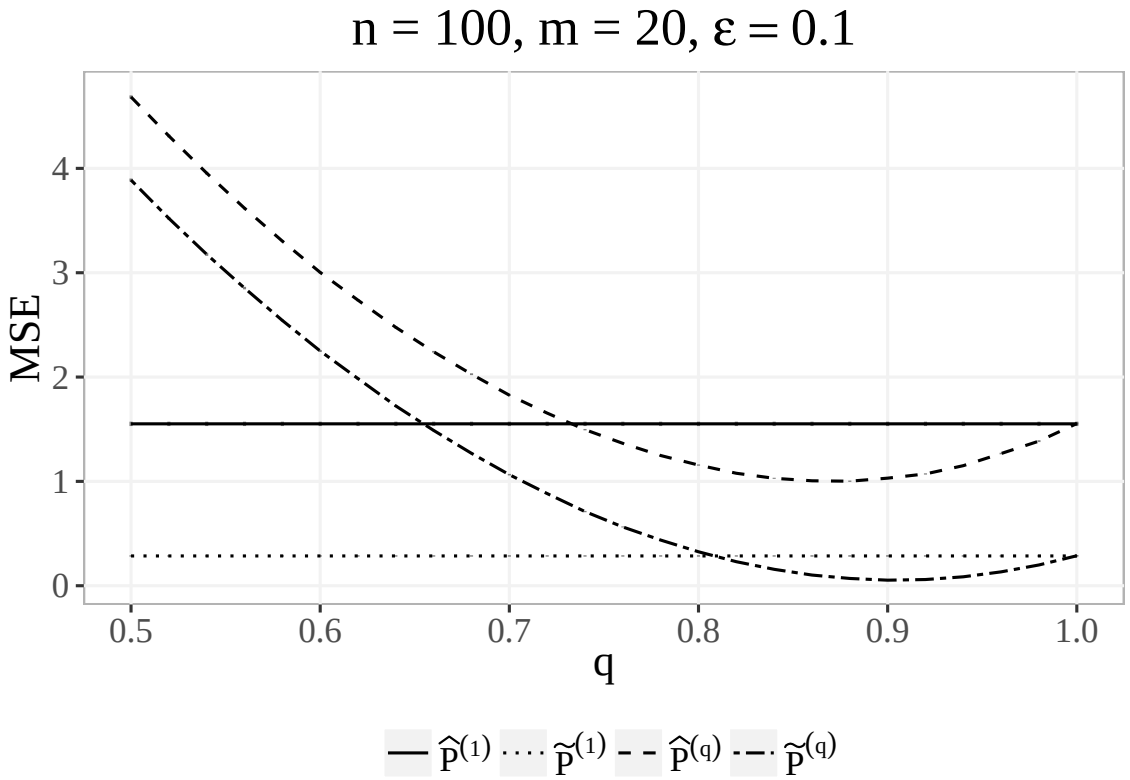


Figure 4.4: Mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different types of lines represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators independent of the selection of q; 2. Within a proper range of q, MLqE wins the bias-variance tradeoff and shows the robustness property compare to the MLE. Also as q goes to 1, MLqE goes to the MLE as expected.

4.6 CoRR Brain Graphs Experiment

We now compare the four estimators on a structural connectomic data. The graphs in this dataset are based on diffusion tensor MR images. There are 114 different brain scans, each of which was processed to yield an undirected, weighted graph with no self-loops, using the m2g pipeline described in [Kiar et al., 2016]. The

vertices of the graphs represent different regions in the brain defined according to an atlas. We used the Desikan atlas with 70 vertices in this expereiment. The weight of an edge between two vertices represents the number of white-matter tract connecting the corresponding two regions of the brain.

Generally, we do not expect the graphs to perfectly follow an RDPG, or not even IEM. Before we calculate those estimators, we will perform some exploratory analysis to check whether the dataset could possibly have a low-rank structure. Indeed, without a low-rank structure, we will not expect the ASE procedure to improve the bias-variance tradeoff because of a potential high bias. In the left panel of Figure 4.5, we plot the eigenvalues of the mean graph of all 114 graphs (with diagonal augmentation) in decreasing algebraic order for the Desikan atlases based on the m2g pipeline. The eigenvalues first decrease dramatically and then stay around zero for a large range of dimensions. In addition, we also plot the histograms in the right panel of Figure 4.5. From the figures we can see many eigenvalues are concentrated around zero. So the information is mostly contained in the first few dimensions. Such quasi low-rank property provides an opportunity to win the bias-variance tradeoff by applying ASE procedure.

We now discuss an important issue with respect to this current dataset. To compare the four estimators, we need a notion of the MSE, which requires the true parameter matrix P. However, unlike simulation experiment in Section 4.5.1, P is definitely not obtainable in practice since the 114 graphs themselves are also a sample

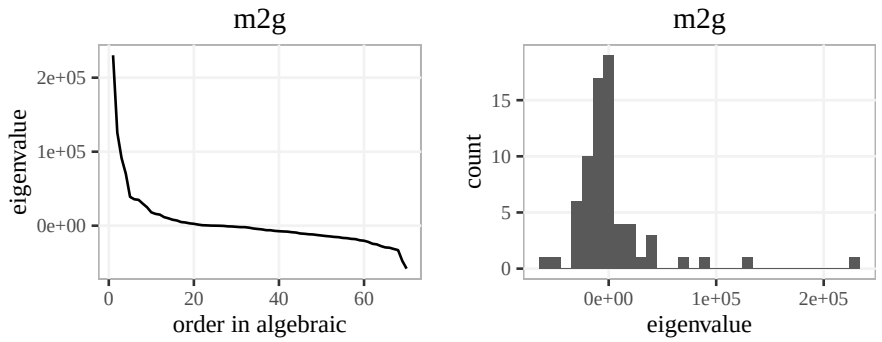


Figure 4.5: Screeplot and the histogram of the eigenvalues of the mean of 114 graphs based on m2g pipeline. The screeplot in the left panel shows the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation in decreasing algebraic order for the Desikan atlas. The right panel shows the histogram of the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation. Many eigenvalues are around zero, which lead to a quasi low-rank structure.

from the population. We address this issue by finding a surrogate estimate for P and use it to calculate the MSE is a feasible way in this experiment. Recently, Kiar et al. [2016] proposed a better pipeline ndmg2 compared to m2g. Then the MLE derived from the 114 graphs in ndmg2 should be a relative more accurate estimate of the actual probability matrix P for the population. And we are going to use this as P when calculating the MSE. However, such P generally has full rank, which breaks the low-rank assumptions. So this setting makes it hard for $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$ to improve and is favorable to the $\widehat{P}^{(1)}$ and $\widehat{P}^{(q)}$. Thus any improvement is conservative. Moreover, it is still possible that the 114 graphs from ndmg2 contain outliers. Thus by using the MLE as P, the performance of MLqE related estimators $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$ are underestimated.

In this experiment, we build the four estimates based on the samples with size m

from the m2g dataset, while using the MLE of all 114 graphs from the ndmg2 dataset as the probability matrix P. Note that diagonal augmentation procedure introduced in Section 3.2.3 is also applied here to compensate for the unnecessary bias. We run 100 simulations on this dataset for different sample sizes $m = 2, 5, 10$. Specifically, in each Monte Carlo replicate, we sample m graphs out of the 114 from the m2g dataset and compute the four estimates based on the m sampled graphs. Once again for simplicity, we set q to be 0.9 without further exploiting. However, the results are consistent for many choices of q. We then compare these estimates to the MLE of all 114 graphs in the ndmg2 dataset. For those two low-rank estimators $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$, we apply ASE for all possible dimensions, i.e. d ranges from 1 to n. The MSE results are shown in Figure 4.6.

When d is small, ASE procedure underestimates the dimension and fails to get important information, which leads to poor performance. In this work, we use Zhu and Ghodsi's method discussed in Section 3.2.2 to select the dimension d. We denote the selected dimensions by square and circle in the figure. We can see the algorithm does a pretty good job for selecting the dimension to embed. More importantly, there is a wide range of dimensions which could lead to a better performance when applying ASE. Although the P we are estimating is actually a high-rank matrix, ASE procedure still wins the bias-variance tradeoff and improves the performance while being suppressed in this setting.

Also, the robust estimator $\widehat{P}^{(q)}$ performs relatively better than $\widehat{P}^{(1)}$ in this exper-

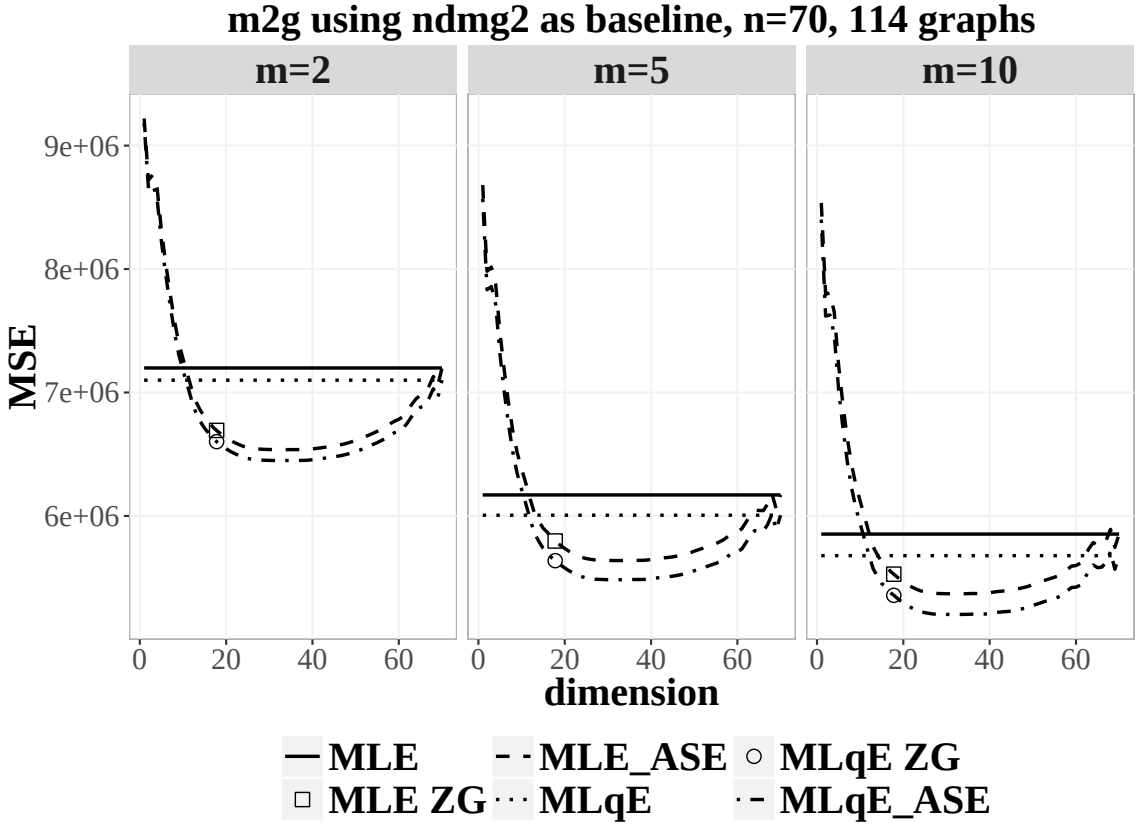


Figure 4.6: Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes. The x-axis represents the dimensions to embed while y-axis is the MSE of each estimator. 1. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs MLqE $\widehat{P}^{(q)}$ (horizontal dotted line): MLqE outperforms MLE since in practice observations are always contaminated and robust estimators are preferred; 2. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$ (dashed line): $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff when being embedded into a proper dimension; 3. MLqE $\widehat{P}^{(q)}$ (horizontal dotted line) vs ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ (dashed dotted line): $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff when being embedded into a proper dimension; 4. ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ (dashed dotted line) vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$ (dashed line): $\widetilde{P}^{(q)}$ is better, since it inherits the robustness from MLqE. The square and circle represent the dimensions selected by the Zhu and Ghodsi method. We can see it does a pretty good job. And more importantly, a wide range of dimensions could lead to an improvement.

iment, even though P still contains outliers. This strongly indicates that there are many outliers in the original graphs from m2g pipeline. And $\widetilde{P}^{(q)}$ successfully inherits the robustness from MLqE and outperforms $\widetilde{P}^{(1)}$.

For all three sample sizes ($m = 2, 5, 10$), $\widetilde{P}^{(q)}$ estimates P most accurately while the target is preferable to the other three estimators more or less. So it should provide a even better estimate for the true but unknown P.

4.7 Appendix: Proofs for Theory Results

4.7.1 Outline of the Proofs

Firstly, in Section 4.7.2, we prove in Lemma 4.7.4 that when the contamination is large enough, the robust estimator $\widehat{P}^{(q)}$ has smaller asymptotic bias compared to $\widehat{P}^{(1)}$. By the results of minimum contrast estimator, we also show in Lemma 4.7.8 that both estimators have variances go to zero as the number of graphs m goes to infinity.

In Section 4.7.3, we mainly analyze the properties of the ASE procedure. We first prove Theorem 4.7.9, which provides an upper bound for the 2-norm of the difference between the estimator $\widehat{P}^{(1)}$ and its expectation $H_{ij}^{(1)} = E[\widehat{P}_{ij}^{(1)}]$. Lemma 4.7.11 shows that $U^\top \widehat{U}$ can be approximated by an orthogonal matrix $W^* = W_1 W_2^\top$, where U and $\widehat{U}$ are the eigenspaces with respect to the largest d eigenvalues of $H_{ij}^{(1)}$ and $\widehat{P}^{(1)}$ respectively. More conveniently, Lemma 4.7.12 indicates that we can change the order

of W^* in the matrix multiplications accordingly without affecting the result much. With these tool results, in Lemma 4.7.13 we give an upper bound of $\|\widehat{Z} - ZW\|_F$, which controls the error of the $\widehat{Z}$ for estimating the true latent positions Z up to rotation. With the extent of Lemma 4.7.13, we then give a bound of the $(2 \to \infty)$-norm of the $\widehat{Z} - ZW$, i.e. $\max_i \|\widehat{Z}_i - WZ_i\|_2$ in Theorem 4.7.14.

In Section 4.7.4, we give a bound of the estimation error $\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right|$ in Lemma 4.7.15 based on the results in Section 4.7.3. In order to bound the variance of our estimator $\widetilde{P}^{(1)}$, all results in this section will be based on a truncated version of $\widetilde{P}^{(1)}$ defined in Definition 4.7.16. This is just for technical reasons and will not affect the estimation procedure in practice, which is discussed in details in Remark 4.7.17. Then we can bound the expectation (Lemma 4.7.18) and variance (Theorem 4.7.19) of the truncated $\widetilde{P}^{(1)}$ by carefully choosing a breakpoint a and analyzing separately. And as a direct result, we have the bound for the relative efficiency between $\widehat{P}_{ij}^{(1)}$ and $\widetilde{P}_{ij}^{(1)}$ in Theorem 4.7.20.

In Section 4.7.5, we compare the performance between $\widetilde{P}^{(q)}$ and $\widehat{P}^{(q)}$. The results in this section are proved in a similar way as in Section 4.7.3 and Section 4.7.4. However, since the MLqE in a mixture distribution model generally does not have a closed form, we explore the relationship between MLE and MLqE to give a relaxed bound which is used when MLqE is hard to analyze.

In Section 4.7.6, we compare the performance between $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$ by combining all the previous results.

In Section 4.7.7, we provide proofs for all supplementary results mentioned in the manuscript.

Here we will first define the notation "with high probability", which is used through out the entire proofs.

Definition 4.7.1 *We say a bound holds with high probability, if there exits a constant $n_0(c)$ such that if $n > n_0$, then for any η satisfying $n^{-c} < \eta < 1/2$, the bound holds with probability greater than $1 - \eta$.*

4.7.2 $\widehat{P}^{(q)}$ vs. $\widehat{P}^{(1)}$

Lemma 4.7.2 *Consider the model $X_1, \cdots, X_m \overset{iid}{\sim} \text{Exp}(P)$ with $m \geq 2$ and $E[X_1] = P$. Given any data $x = (x_1, \cdots, x_m)$ such that $x_{(1)} > 0$ and not all x_i's are the same, then no matter how the data is sampled, we have*

- *There exists at least one solution to the MLq equation;*
- *All the solutions to the MLq equation are less than the MLE.*

Thus the MLqE $\widehat{P}^{(q)}$, the root closest to the MLE, is well defined.

Proof: The MLE is

$$\widehat{P}^{(1)}(x) = \bar{x}.$$

Consider the continuous function $g(\theta, x) = \sum_{i=1}^{m} e^{-\frac{(1-q)x_i}{\theta}}(x_i - \theta)$. Then the ML$q$ equation is $g(\theta, x) = 0$.

Let $x_{(1)} \leq \cdots \leq x_{(l)} \leq \bar{x} \leq x_{(l+1)} \leq \cdots \leq x_{(m)}$. Define $s_i = \bar{x} - x_{(i)}$ for $1 \leq i \leq l$, and $t_i = x_{(l+i)} - \bar{x}$ for $1 \leq i \leq m-l$. Note that $\sum_{i=1}^{l} s_i = \sum_{i=1}^{m-l} t_i$. Then for any $\theta \geq \bar{x}$, we have

$$\begin{aligned}
g(\theta, x) &= \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}} (x_{(i)} - \theta) = \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}} (x_{(i)} - \bar{x} + \bar{x} - \theta) \\
&= -\sum_{i=1}^{l} e^{-\frac{(1-q)x_{(i)}}{\theta}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i + \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}} (\bar{x} - \theta) \\
&\leq -\sum_{i=1}^{l} e^{-\frac{(1-q)x_{(i)}}{\theta}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -e^{-\frac{(1-q)x_{(l+1)}}{\theta}} \sum_{i=1}^{l} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -e^{-\frac{(1-q)x_{(l+1)}}{\theta}} \sum_{i=1}^{m-l} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -\sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&= 0,
\end{aligned}$$

and equality holds if and only if all x_i's are the same, which is excluded by the assumption. Thus $g(\theta, x) < 0$ for any $\theta \geq \bar{x}$.

Denote any solution to the MLq equation to be $\widehat{P}^{(q)}(x)$, then we also know:

- $g(\widehat{P}^{(q)}(x), x) = 0$;
- $\lim_{\theta \to 0^+} g(\theta, x) = 0$;
- $g(\theta, x) > 0$ when $\theta < x_{(1)}$;

Thus there exists at least one solution to the MLq equation. And all solutions to the MLq equation are between $x_{(1)}$ and $\bar{x}$, i.e. less than the MLE. ■

Lemma 4.7.3 *Consider an exponential distribution model while the data is actually sampled from the contaminated model* $X, X_1, \cdots, X_m \overset{iid}{\sim} (1-\epsilon)\text{Exp}(P) + \epsilon\text{Exp}(C)$. *Denote such contaminated distribution as* F. *Then there exists exactly one real solution* $\theta(F)$ *of the population version of MLq equation, i.e.* $E_F[e^{-\frac{(1-q)X}{\theta(F)}}(X-\theta(F))] = 0$. *Moreover,* $\theta(F) < E_F[\bar{X}] = (1-\epsilon)P + \epsilon C$.

Proof: For the MLE, i.e. $\bar{X}$, we have $E[\bar{X}] = (1-\epsilon)P + \epsilon C$. According to Equation (3.2) in [**?**], $\theta(F)$ satisfies

$$\frac{\epsilon C}{(C(1-q)+\theta)^2} - \frac{\epsilon}{C(1-q)+\theta} + \frac{(1-\epsilon)P}{(P(1-q)+\theta)^2} - \frac{(1-\epsilon)}{P(1-q)+\theta} = 0,$$

i.e.

$$\frac{\epsilon(\theta - Cq)}{(C(1-q)+\theta)^2} = \frac{(1-\epsilon)(Pq-\theta)}{(P(1-q)+\theta)^2}.$$

Define $h(\theta) = (C(1-q)+\theta)^2(1-\epsilon)(Pq-\theta) - (P(1-q)+\theta)^2\epsilon(\theta - Cq)$. Then $\lim_{\theta\to\infty} h(\theta) = -\infty$, $h(0) > 0$, and $h(Cq) < 0$. Consider q as the variable and solve the equation $h(E[\bar{X}]) = 0$, we have three roots and one of them is $q = 1$ obviously. The other two roots are

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon + C(1-\epsilon))} \pm \sqrt{\frac{\epsilon(1-\epsilon)(C-P)^2\left(\epsilon(1-\epsilon)(C-P)^4 - 4P^2C^2\right)}{4P^2C^2(P\epsilon + C(1-\epsilon))^2}}.$$

To prove the roots are greater or equal to 1, we need to show

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon+C(1-\epsilon))}-\sqrt{\frac{\epsilon(1-\epsilon)(C-P)^2\left(\epsilon(1-\epsilon)(C-P)^4-4P^2C^2\right)}{4P^2C^2(P\epsilon+C(1-\epsilon))^2}}>1.$$

For the first part,

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon+C(1-\epsilon))}>1+\frac{(P-C)^2\epsilon(1-\epsilon)(P+C)}{2PC(P\epsilon+C(1-\epsilon))}.$$

To prove the roots are greater or equal to 1, we just need to show

$$(P-C)^4\epsilon^2(1-\epsilon)^2(P+C)^2\geq\epsilon^2(1-\epsilon)^2(C-P)^6.$$

Then it is sufficient to show that

$$(P+C)^2\geq(C-P)^2,$$

which is true. Combined with the fact that when $q=0$, $h(E[\bar{X}])<0$, we have for any $0<q<1$, $h(E[\bar{X}])<0$.

The equation $h(\theta)=0$ is a cubic polynomial, so it has at most three real roots. In addition, by calculating we know there is only one real root, while the other two are complex roots. Combined with the fact that $h(Pq)>0$, we have for any $0<q<1$, the only real root of the population version of MLq equation is less than

$E[\bar{X}] = (1-\epsilon)P + \epsilon C.$ ■

Lemma 4.7.4 (Theorem 4.3.1) *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

$$\lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(q)}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|,$$

for $1 \le i, j \le n$ and $i \ne j$.

Proof: For the MLE $\widehat{P}_{ij}^{(1)} = \bar{A}_{ij}$,

$$E[\widehat{P}_{ij}^{(1)}] = E[\bar{A}_{ij}] = \frac{1}{m}\sum_{t=1}^{m} E[A_{ij}^{(t)}] = E[A_{ij}^{(1)}] = (1-\epsilon)P_{ij} + \epsilon C_{ij}.$$

As shown in Lemma 4.7.3, $\theta(F)$ satisfies

$$\frac{\epsilon(\theta(F) - C_{ij}q)}{(C_{ij}(1-q) + \theta(F))^2} = \frac{(1-\epsilon)(P_{ij}q - \theta(F))}{(P_{ij}(1-q) + \theta(F))^2}.$$

Thus $\theta(F) - C_{ij}q$ and $\theta(F) - P_{ij}q$ should have different signs. Combined with $C_{ij} > P_{ij}$, we have

$$qP_{ij} < \theta(F).$$

To have a smaller asymptotic bias in absolute value, combined with Lemma 4.7.7, we need

$$|\theta(F) - P_{ij}| < \epsilon(C_{ij} - P_{ij}).$$

Based on Lemma 4.7.2, we need

$$qP_{ij} > P_{ij} - \epsilon(C_{ij} - P_{ij}),$$

i.e.

$$C_{ij} > P_{ij} + \frac{(1-q)P_{ij}}{\epsilon} = C_0(P_{ij}, \epsilon, q).$$

■

Lemma 4.7.5 *The MLqE based on the model to be exponential distribution* $\text{Exp}(P)$ *while the data is actually sampled from the contaminated distribution* $(1-\epsilon)\text{Exp}(P) + \epsilon\text{Exp}(C)$ *is a minimum contrast estimator.*

Proof: Consider the contaminated distribution $F(x) = (1-\epsilon)f(x;P) + \epsilon f(x;C)$, where $f(x)$ represents the pdf of exponential distribution. By Lemma 4.7.3, we know there is a one-to-one correspondence between the uncontaminated parameter P and the only real solution $\theta(F)$ of the population version of MLq equation, i.e. $E_F[e^{-\frac{(1-q)X}{\theta(F)}}(X - \theta(F))] = 0$. Let $r(\theta(F)) = P$. Then we can define $\rho(x;\theta) = \frac{f(x;r(\theta))^{1-q}}{1-q}$, where $q \in (0,1)$ is a constant. By reparameterizing $\rho(x;\theta)$ to $\widetilde{\rho}(x;r)$ such that $\widetilde{\rho}(x;r(\theta)) = \rho(x;\theta)$, we can use the proof of Lemma 4.7.3 directly to prove that $D(\theta_0,\theta) = E_{\theta_0}[\rho(X,\theta)]$ is uniquely minimized at θ_0. Thus the MLqE is a minimum contrast estimator. ■

Lemma 4.7.6 *Uniform convergence of the MLq equation, i.e.*

$$\sup_{\theta\in[0,R]}\left|\frac{1}{m}\sum_{i=1}^{m}e^{-\frac{(1-q)X_i}{\theta}}(X_i-\theta)-E_F[e^{-\frac{(1-q)X}{\theta}}(X-\theta)]\right|\overset{a.s.}{\to}0.$$

Proof: Define $g(x,\theta)=e^{-\frac{(1-q)x}{\theta}}(x-\theta)$ and $d(x)=e^{-\frac{(1-q)x}{R}}(x+R)$. Then $E_F[d(X)]<\infty$ and $g(x,\theta)\leq d(x)$ for all $\theta\in[0,R]$. Combined with the fact that $[0,R]$ is compact and the function $g(x,\theta)$ is continuous at each θ for all $x>0$ and measurable function of x at each θ, we have the uniform convergence by Lemma 2.4 in [Newey and McFadden, 1994]. ■

Lemma 4.7.7 $\widehat{P}_{ij}^{(q)}\overset{P}{\to}\theta(F_{ij})$ *as* $m\to\infty$, *where* F_{ij} *is the contaminated distribution* $(1-\epsilon)\text{Exp}(P_{ij})+\epsilon\text{Exp}(C_{ij})$, *and* $\theta(F_{ij})$ *is defined in Lemma 4.7.3.*

Proof: By the proof of Lemma 4.7.3, we have

$$\inf\{D(\theta_0,\theta):|\theta-\theta_0|\geq\epsilon\}>D(\theta_0,\theta_0)$$

for every $\epsilon>0$. Combined with Lemma 4.7.6, we know the MLq is consistent based on Theorem 5.2.3 in [Bickel and Doksum, 2007]. ■

Lemma 4.7.8 (Theorem 4.3.1) *For* $1\leq i,j\leq n$,

$$\text{Var}(\widehat{P}_{ij}^{(1)})=\text{Var}(\widehat{P}_{ij}^{(q)})=O(1/m).$$

And thus

$$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0.$$

Proof: Both MLE and MLqE are minimum constrast estimators. By consistency (shown in Lemma 4.7.7) and other regularity conditions, we know the variances are both of order $1/m$ based on Theorem 5.4.2 in [Bickel and Doksum, 2007]. ■

4.7.3 ASE Procedure of $\widehat{P}^{(1)}$

Theorem 4.7.9 *Let P and C be two n-by-n symmetric matrices satisfying element-wise conditions $0 < P_{ij} \leq C_{ij} \leq R$ for some constant $R > 0$. For $0 < \epsilon < 1$, we define m symmetric and hollow matrices as*

$$A^{(t)} \overset{iid}{\sim} (1-\epsilon)\mathrm{Exp}(P) + \epsilon\mathrm{Exp}(C),$$

for $1 \leq t \leq m$. Let $\widehat{P}^{(1)}$ be the element-wise MLE based on exponential distribution with m observations. Define $H_{ij}^{(1)} = E[\widehat{P}_{ij}^{(1)}] = (1-\epsilon)P_{ij} + \epsilon C_{ij}$, then for any constant $c > 0$, there exists another constant $n_0(c)$, independent of n, P, C and ϵ, such that if $n > n_0$, then for all η satisfying $n^{-c} \leq \eta \leq 1/2$,

$$P\left(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \leq 4R\sqrt{n\ln(n/\eta)/m}\right) \geq 1-\eta.$$

Remark: This is an extended version of Theorem 3.1 in [Oliveira, 2009].

Proof: Let $\{e_i\}_{i=1}^n$ be the canonical basis for $\mathbb{R}^n$. For each $1 \leq i, j \leq n$, define a corresponding matrix G_{ij}:

$$G_{ij} \equiv \begin{cases} e_i e_j^T + e_j e_i^T, & i \neq j; \\ e_i e_i^T, & i = j. \end{cases}$$

Thus

$$\widehat{P}^{(1)} = \sum_{1 \leq i < j \leq n} \widehat{P}_{ij}^{(1)} G_{ij} = \frac{1}{m} \sum_{t=1}^{m} \sum_{1 \leq i < j \leq n} A_{ij}^{(t)} G_{ij}$$

and

$$H^{(1)} = \sum_{1 \leq i < j \leq n} H_{ij}^{(1)} G_{ij}.$$

Then we have $\widehat{P}^{(1)} - H^{(1)} = \frac{1}{m} \sum_{1 \leq t \leq m, 1 \leq i < j \leq n} X_{ij}^{(t)}$, where $X_{ij}^{(t)} = \left(A_{ij}^{(t)} - H_{ij}^{(1)} \right) G_{ij}$ for $1 \leq t \leq m$ and $1 \leq i < j \leq n$.

First bound the k-th moment of X_{ij} for $1 \leq i < j \leq n$ as following:

$$\begin{aligned} E[(A_{ij}^{(t)} - H_{ij}^{(1)})^k] \leq & (1-\epsilon) \cdot \exp(-H_{ij}/P_{ij}) P_{ij}^k \Gamma(1+k, -H_{ij}/P_{ij}) \\ & + \epsilon \cdot \exp(-H_{ij}/C_{ij}) C_{ij}^k \Gamma(1+k, -H_{ij}/C_{ij}) \\ \leq & \left((1-\epsilon) \cdot \exp(-H_{ij}/P_{ij}) P_{ij}^k + \epsilon \cdot \exp(-H_{ij}/C_{ij}) C_{ij}^k \right) k! \\ \leq & \left((1-\epsilon) \cdot P_{ij}^k + \epsilon \cdot C_{ij}^k \right) k! \\ \leq & R^k k!, \end{aligned} \tag{4.1}$$

Combined with

$$G_{ij}^k \equiv \begin{cases} e_i e_i^T + e_j e_j^T, & \text{k is even;} \\ e_i e_j^T + e_j e_i^T, & \text{k is odd,} \end{cases}$$

thus we have

1. When k is even,

$$E[(X_{ij}^{(t)})^k] = E[(A_{ij}^{(t)} - H_{ij}^{(1)})^k] G_{ij}^2 \preceq k! R^k G_{ij}^2;$$

2. When k is odd,

$$E[(X_{ij}^{(t)})^k] = E[(A_{ij}^{(t)} - H_{ij}^{(1)})^k] G_{ij} \preceq k! R^k G_{ij}^2.$$

So

$$E[(X_{ij}^{(t)})^k] \preceq k! R^k G_{ij}^2.$$

Let

$$\sigma^2 := \left\| \sum_{1 \le t \le m, 1 \le i < j \le n} (\sqrt{2} R G_{ij})^2 \right\|_2 = 2R^2 m \|(n-1)I\|_2 = 2R^2 m(n-1).$$

Notice that random matrices $X_{ij}^{(t)}$ are independent, self-adjoint and have mean zero,

apply Theorem 6.2 in [Tropp, 2012] we have

$$\begin{aligned} P\left(\lambda_{\max}(\widehat{P}^{(1)} - H^{(1)}) \geq t\right) &= P\left(\lambda_{\max}\left(\frac{1}{m}\sum_{1\leq t\leq m, 1\leq i<j\leq n} X_{ij}^{(t)}\right) \geq t\right) \\ &= P\left(\lambda_{\max}\left(\sum_{1\leq t\leq m, 1\leq i<j\leq n} X_{ij}^{(t)}\right) \geq mt\right) \\ &\leq n\exp\left(-\frac{(mt)^2/2}{\sigma^2 + Rmt}\right) \\ &\leq n\exp\left(-\frac{mt^2/2}{2R^2n + Rt}\right). \end{aligned}$$

Now consider $Y_{ij}^{(t)} \equiv \left(H_{ij}^{(1)} - A_{ij}^{(t)}\right) G_{ij}$, for $1 \leq t \leq m$ and $1 \leq i < j \leq n$. Then we have $H^{(1)} - \widehat{P}^{(1)} = \frac{1}{m}\sum_{1\leq t\leq m, 1\leq i<j\leq n} Y_{ij}^{(t)}$. Since

$$E[(H^{(1)} - \widehat{P}^{(1)})^k] = (-1)^k E[(\widehat{P}^{(1)} - H^{(1)})^k],$$

1. When k is even,

$$E[(Y_{ij}^{(t)})^k] = E[(\widehat{P}^{(1)} - H^{(1)})^k]G_{ij}^2 \preceq k!R^kG_{ij}^2;$$

2. When k is odd,

$$E[Y_{ij}^k] = -E[(\widehat{P}^{(1)} - H^{(1)})^k]G_{ij} \preceq k!R^kG_{ij}^2.$$

Thus by similar arguments,

$$P\left(\lambda_{\min}(\widehat{P}^{(1)} - H^{(1)}) \leq -t\right) = P\left(\lambda_{\max}(H^{(1)} - \widehat{P}^{(1)}) \geq t\right) \leq n \exp\left(-\frac{mt^2/2}{2R^2 n + Rt}\right).$$

Therefore we have

$$P\left(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \geq t\right) \leq n \exp\left(-\frac{mt^2/2}{2R^2 n + Rt}\right).$$

Now let $c > 0$ be given and assume $n^{-c} \leq \eta \leq 1/2$. Then there exists a $n_0(c)$ independent of n, P, C and ϵ such that whenever $n > n_0(c)$,

$$t = 4R\sqrt{n \ln(n/\eta)/m} \leq 6Rn.$$

Plugging this t into the equation above, we get

$$P(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \geq 4R\sqrt{n \ln(n/\eta)/m}) \leq n \exp\left(-\frac{t^2}{16R^2 n}\right) = \eta.$$

■

Define $H^{(1)} = E[\widehat{P}^{(1)}] = (1-\epsilon)P + \epsilon C$, where $P = XX^T$, $X \in \mathbb{R}^{n \times d}$, $C = YY^T$, $Y \in \mathbb{R}^{n \times d'}$. Let $d^{(1)} = \text{rank}(H^{(1)})$ be the dimension in which we are going to embed $\widehat{P}^{(1)}$. Then we can define $H^{(1)} = ZZ^T$ where $Z \in \mathbb{R}^{n \times d^{(1)}}$. Since $H^{(1)} =$

$[\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y][\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y]^T$, we have $d^{(1)} \leq d + d'$.

For simplicity, from now on, we will use $\widehat{P}$ to represent $\widehat{P}^{(1)}$, use H to represent $H^{(1)}$ and use k to represent the dimension $d^{(1)}$ we are going to embed. Assume $H = USU^T = ZZ^T$, where $Z = [Z_1, \cdots, Z_n]^T$ is a n-by-k matrix. Then our estimate for Z up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^T$ is the rank-k spectral decomposition of $|\widehat{P}| = (\widehat{P}^T\widehat{P})^{1/2}$.

Furthermore, we assume that the second moment matrix $E[Z_1 Z_1^T]$ is rank k and has distinct eigenvalues $\lambda_i(E[Z_1 Z_1^T])$. In particular, we assume that there exists $\delta > 0$ such that

$$\delta < \lambda_k(E[Z_1 Z_1^T])$$

Lemma 4.7.10 *Under the above assumptions, $\lambda_i(H) = \Theta(n)$ with high probability when $i \leq k$, i.e. the largest k eigenvalues of H is of order n. Moreover, we have $\|S\|_2 = \Theta(n)$ and $\|\widehat{S}\|_2 = \Theta(n)$ with high probability.*

Remark: This is an extended version of Proposition 4.3 in [Sussman et al., 2014].

Proof: Note that $\lambda_i(H) = \lambda_i(ZZ^T) = \lambda_i(Z^T Z)$ when $i \leq k$. Since each entry of $Z^T Z$ is a sum of n independent random variables each in $[0, R]$, i.e. $(Z^T Z)_{ij} = \sum_{l=1}^n Z_{li} Z_{lj}$. By Hoeffding's inequality,

$$P(|(Z^T Z - nE[Z_1 Z_1^T])_{ij}| \geq t) \leq 2\exp(-\frac{2t^2}{nR^2}).$$

Now let $c > 0$ and assume $n^{-c} \leq \eta \leq 1/2$. Let

$$t = R\sqrt{n \ln(\sqrt{2/\eta})},$$

we have

$$P\left(|(Z^T Z - nE[Z_1 Z_1^T])_{ij}| \geq R\sqrt{n \ln(\sqrt{2/\eta})}\right) \leq \eta.$$

By the union bound, we have

$$P\left(\|Z^T Z - nE[Z_1 Z_1^T]\|_F \geq kR\sqrt{n \ln(\sqrt{2/\eta})}\right) \leq k^2\eta.$$

Then by Weyl's Theorem [Horn and Johnson, 2012], we have

$$|\lambda_i(H) - n\lambda_i(E[Z_1 Z_1^T])| \leq \|Z^T Z - nE[Z_1 Z_1^T]\|_2 = O(\sqrt{n \log(1/\eta)})$$

with probability at least $1 - k^2\eta$. Thus $\lambda_i(H) = S_{ii} = \Theta(n)$ with probability at least $1 - \frac{2k^2}{n^2}$ when $i \leq k$. Moreover,

$$\|H\|_2 - \|H - \widehat{P}\|_2 \leq \|\widehat{S}\|_2 \leq \|\widehat{P} - H\|_2 + \|H\|_2.$$

Combined with Theorem 4.7.9, with high probability we have $\|\widehat{S}\|_2 = \Theta(n)$. ∎

Lemma 4.7.11 *Let $W_1 \Sigma W_2^T$ be the singular value decomposition of $U^T \widehat{U}$. Then for*

sufficiently large n,

$$\|U^T\widehat{U} - W_1W_2^T\|_F = O(m^{-1}n^{-1}\log n)$$

with high probability.

Proof: Let $\sigma_1, \cdots, \sigma_k$ denote the singular values of $U^T\widehat{U}$. Then $\sigma_i = \cos(\theta_i)$ where the θ_i are the principal angles between the subspaces spanned by $\widehat{U}$ and U. Furthermore, by the Davis-Kahan $\sin(\Theta)$ theorem [Davis and Kahan, 1970], combined with Theorem 4.7.9 and Lemma 4.7.10,

$$\begin{aligned} &\|\widehat{U}\widehat{U}^T - UU^T\|_2 = \max_i |\sin(\theta_i)| \\ \leq &\frac{\|\widehat{P} - H\|_2}{\lambda_k(H)} \leq \frac{C\sqrt{n\log n/m}}{n} \\ = &O(m^{-1/2}n^{-1/2}\sqrt{\log n}) \end{aligned} \tag{4.2}$$

for sufficiently large n with high probability. Here $\lambda_k(H)$ denotes the k-th largest

eigenvalue of H. Thus with high probability,

$$\begin{aligned}\|U^T\widehat{U} - W_1W_2^T\|_F = \|\Sigma - I\|_F &= \sqrt{\sum_{i=1}^{k}(1-\sigma_i)^2} \\ &\leq \sum_{i=1}^{k}(1-\sigma_i) \leq \sum_{i=1}^{k}(1-\sigma_i^2) \\ &= \sum_{i=1}^{k}\sin^2(\theta_i) \leq k\|\widehat{U}\widehat{U}^T - UU^T\|_2^2 \\ &= O(m^{-1}n^{-1}\log n).\end{aligned}$$

■

We will denote the orthogonal matrix $W_1W_2^T$ by W^*.

Lemma 4.7.12 *For sufficiently large n,*

$$\|W^*\widehat{S} - SW^*\|_F = O(m^{-1/2}\log n),$$

$$\|W^*\widehat{S}^{1/2} - S^{1/2}W^*\|_F = O(m^{-1/2}n^{-1/2}\log n)$$

and

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(m^{-1/2}n^{-3/2}\log n)$$

with high probability.

Proof: By Proposition 2.1 in [Rohe et al., 2011] and Equation (4.2), we have for

some orthogonal matrix W,

$$\begin{aligned}\|\widehat{U} - UW\|_F^2 \leq \frac{2\|\widehat{U}\widehat{U}^T - UU^T\|_F^2}{\delta^2} \leq \frac{8k^2\|\widehat{U}\widehat{U}^T - UU^T\|_2^2}{\delta^2}\\ =O(m^{-1}n^{-1}\log n),\end{aligned}$$

with high probability. Let $Q = \widehat{U} - UU^T\widehat{U}$. And Q is the residual after projecting $\widehat{U}$ orthogonally onto the column space of U, we have

$$\|Q\|_F = \|\widehat{U} - UU^T\widehat{U}\|_F \leq \|\widehat{U} - UT\|_F = O(m^{-1/2}n^{-1/2}\sqrt{\log n}). \tag{4.3}$$

for all $k \times k$ matrices T with high probability. Then

$$\begin{aligned}W^*\widehat{S} =&(W^* - U^T\widehat{U})\widehat{S} + U^T\widehat{U}\widehat{S} = (W^* - U^T\widehat{U})\widehat{S} + U^T\widehat{P}\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)\widehat{U} + U^TH\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + U^TH\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + SU^T\widehat{U}.\end{aligned}$$

Combined with Theorem 4.7.9, Lemma 4.7.10, Lemma 4.7.11, we have

$$\begin{aligned}
&\|W^*\widehat{S} - SW^*\|_F \\
=&\|(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + S(U^T\widehat{U} - W^*)\|_F \\
\leq&\|W^* - U^T\widehat{U}\|_F(\|\widehat{S}\|_2 + \|S\|_2) + \|U^T\|_F\|\widehat{P} - H\|_2\|Q\|_F + \|U^T(\widehat{P} - H)U\|_F \\
\leq&O(m^{-1}\log n) + O(m^{-1/2}\log n) + \|U^T(\widehat{P} - H)U\|_F
\end{aligned}$$

with high probability. And we know $U^T(\widehat{P} - H)U$ is a $k \times k$ matrix with ij-th entry to be

$$u_i^T(\widehat{P} - H)u_j = \sum_{s=1}^{n}\sum_{t=1}^{n}(\widehat{P}_{st} - H_{st})u_{is}u_{jt} = 2\sum_{s<t}(\widehat{P}_{st} - H_{st})u_{is}u_{jt}$$

where u_i and u_j are the i-th and j-th columns of U. Thus, conditioned on H, U is fixed and $u_i^T(\widehat{P} - H)u_j$ is a sum of independent mean 0 random variables.

By Equation (4.1), we have

$$\begin{aligned}
&E\left[\left((A_{st}^{(t')} - H_{st})u_{is}u_{jt}\right)^k\right] \\
\leq&k!R^k u_{is}^k u_{jt}^k \\
\leq&\frac{k!}{2}R^{k-2}(\sqrt{2}u_{is}u_{jt}R)^2.
\end{aligned}$$

Also we have

$$\sigma^2 := \left|\sum_{t',s<t} 2R^2u_{is}^2u_{jt}^2\right| \leq mR^2,$$

then by Theorem 6.2 in [Tropp, 2012], we have

$$P\left(\left|2\sum_{s<t}(\widehat{P}_{st}-H_{st})u_{is}u_{jt}\right|\geq t\right)\leq \exp\left(\frac{-mt^2/8}{R^2+Rt/2}\right).$$

Let $t = cRm^{-1/2}\log n$ for any $c > 0$, we have

$$P\left(\left|2\sum_{s<t}(\widehat{P}_{st}-H_{st})u_{is}u_{jt}\right|\geq Cm^{-1/2}\log n\right)\leq n^{-c}.$$

Thus each entry of $U^T(\widehat{P}-H)U$ is of order $O(m^{-1/2}\log n)$ with high probability and

$$\|U^T(\widehat{P}-H)U\|_F = O(m^{-1/2}\log n) \tag{4.4}$$

with high probability. Hence

$$\|W^*\widehat{S}-SW^*\|_F = O(m^{-1/2}\log n)$$

with high probability. Also, since

$$W^*_{ij}(\lambda_j^{1/2}(\widehat{P})-\lambda_i^{1/2}(H)) = W^*_{ij}\frac{\lambda_j(\widehat{P})-\lambda_i(H)}{\lambda_j^{1/2}(\widehat{P})+\lambda_i^{1/2}(H)}$$

and the eigenvalues $\lambda_j^{1/2}(\widehat{P})$ and $\lambda_i^{1/2}(H)$ are both of order $\Theta(\sqrt{n})$, we have

$$\|W^*\widehat{S}^{1/2}-S^{1/2}W^*\|_F = O(m^{-1/2}n^{-1/2}\log n)$$

with high probability. Similarly, since

$$W^*_{ij}(\lambda_j^{-1/2}(\widehat{P}) - \lambda_i^{-1/2}(H)) = W^*_{ij}\frac{\lambda_i(H) - \lambda_j(\widehat{P})}{(\lambda_j^{-1/2}(\widehat{P}) + \lambda_i^{-1/2}(H))\lambda_j(\widehat{P})\lambda_i(H)}$$

and the eigenvalues $\lambda_j(\widehat{P})$ and $\lambda_i(H)$ are both of order $\Theta(n)$, with high probability we have

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(m^{-1/2}n^{-3/2}\log n).$$

■

Lemma 4.7.13 *There exists a rotation matrix W such that for sufficiently large n,*

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P} - H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: Let $Q_1 = UU^T\widehat{U} - UW^*$, $Q_2 = W^*\widehat{S}^{1/2} - S^{1/2}W^*$ and $Q_3 = \widehat{U} - UW^* =$

$\widehat{U} - UU^T\widehat{U} + Q_1 = Q + Q_1$. Then since $UU^TH = H$ and $\widehat{U}\widehat{S}^{1/2} = \widehat{P}\widehat{U}\widehat{S}^{-1/2}$,

$$\begin{aligned}
\widehat{Z} - US^{1/2}W^* =& \widehat{U}\widehat{S}^{1/2} - UW^*\widehat{S}^{1/2} + U(W^*\widehat{S}^{1/2} - S^{1/2}W^*) \\
=& (\widehat{U} - UU^T\widehat{U})\widehat{S}^{1/2} + Q_1\widehat{S}^{1/2} + UQ_2 \\
=& (\widehat{P} - H)\widehat{U}\widehat{S}^{-1/2} - UU^T(\widehat{P} - H)\widehat{U}\widehat{S}^{-1/2} + Q_1\widehat{S}^{1/2} + UQ_2 \\
=& (\widehat{P} - H)UW^*\widehat{S}^{-1/2} - UU^T(\widehat{P} - H)UW^*\widehat{S}^{-1/2} \\
& + (I - UU^T)(\widehat{P} - H)Q_3\widehat{S}^{-1/2} + Q_1\widehat{S}^{1/2} + UQ_2.
\end{aligned}$$

By Lemma 4.7.11, with high probability,

$$\|Q_1\|_F \leq \|U\|_F\|U^T\widehat{U} - W^*\|_F = O(m^{-1}n^{-1}\log n).$$

By Lemma 4.7.12, with high probability,

$$\|Q_2\|_F = O(m^{-1/2}n^{-1/2}\log n).$$

By Equation (4.3), with high probability,

$$\|Q_3\|_F \leq \|Q\|_F + \|Q_1\|_F = O(m^{-1/2}n^{-1/2}(\log n)^{1/2}).$$

By Equation (4.4), with high probability,

$$\|UU^T(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F \leq \|U^T(\widehat{P}-H)U\|_F\|\widehat{S}^{-1/2}\|_2 = O(m^{-1}n^{-1/2}\log n).$$

By Lemma 4.7.12, with high probability,

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(m^{-1/2}n^{-3/2}\log n).$$

Therefore, with high probability,

$$\begin{aligned}
&\|\widehat{Z} - US^{1/2}W^*\|_F \\
=&\|(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F + O(m^{-1}n^{-1/2}\log n) + \|I-UU^T\|_2\|\widehat{P}-H\|_2 O(m^{-1/2}n^{-1}(\log n)^{1/2}) \\
&+ O(m^{-1}n^{-1/2}\log n) + O(m^{-1/2}n^{-1/2}\log n) \\
=&\|(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}\log n) \\
\leq&\|(\widehat{P}-H)US^{-1/2}W^*\|_F + \|(\widehat{P}-H)U(W^*\widehat{S}^{-1/2} - S^{-1/2}W^*)\|_F + O(m^{-1/2}n^{-1/2}\log n) \\
=&\|(\widehat{P}-H)US^{-1/2}\|_F + O(m^{-1}n^{-1}(\log n)^{3/2}) + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\
=&\|(\widehat{P}-H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}).
\end{aligned}$$

Note that $Z = US^{1/2}W$ for some orthogonal matrix W. As W^* is also orthogonal, therefore $Z\widetilde{W} = US^{1/2}W^*$ for some orthogonal $\widetilde{W}$, which completes the proof. ■

Theorem 4.7.14 *There exists a rotation matrix W such that for sufficiently large*

n,

$$\max_i \|\widehat{Z}_i - WZ_i\|_2 = O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: By Lemma 4.7.13, we have

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P} - H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability and similarly we could have the bound for each column vector with high probability that

$$\begin{aligned}\max_i \|\widehat{Z}_i - WZ_i\|_2 \leq & \frac{1}{\lambda_k^{1/2}(H)} \max_i \|((\widehat{P} - H)U)_i\|_2 + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ \leq & \frac{k^{1/2}}{\lambda_k^{1/2}(H)} \max_j \|(\widehat{P} - H)u_j\|_\infty + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})\end{aligned}$$

where $((\widehat{P} - H)U)_i$ represents the i-th row of $(\widehat{P} - H)U$ and u_j denotes the j-th column of U. Now given i and j, the i-th element of the vector $(\widehat{P} - H)u_j$ is of the form

$$\sum_{s=1}^{n}(\widehat{P}_{is} - H_{is})u_{js} = \sum_{s\neq i}(\widehat{P}_{is} - H_{is})u_{js}.$$

Thus, conditioned on H, the i-th element of the vector $(\widehat{P} - H)u_j$ is a sum of inde-

pendent mean 0 random variables. By Equation (4.1), we have

$$\begin{aligned} & E\left[\left((A_{is}^{(t)}-H_{is})u_{js}\right)^k\right] \\ \leq & k!R^k u_{js}^k \\ \leq & \frac{k!}{2}R^{k-2}(\sqrt{2}Ru_{js})^2. \end{aligned}$$

Also we have

$$\sigma^2 := |\sum_{t,s\neq i} 2R^2u_{js}^2| \leq 2R^2m,$$

then by Theorem 6.2 in [Tropp, 2012], we have

$$P\left(\left|\sum_{s\neq i}(\widehat{P}_{is}-H_{is})u_{js}\right| \geq t\right) \leq \exp\left(\frac{-mt^2/2}{2R^2+Rt}\right).$$

Let $t = 3cRm^{-1/2}\log n$, we have

$$P\left(\left|\sum_{s\neq i}(\widehat{P}_{is}-H_{is})u_{js}\right| \geq 3cRm^{-1/2}\log n\right) \leq n^{-c},$$

i.e. it is of order $O(m^{-1/2}\log n)$ with high probability. Taking the union bound over all i and j, with high probability we have,

$$\begin{aligned} \max_i \|\widehat{Z}_i - WZ_i\|_2 &\leq \frac{Ck^{1/2}}{\lambda_k^{1/2}(H)}m^{-1/2}(\log n)^{3/2} + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ &= O(m^{-1/2}n^{-1/2}(\log n)^{3/2}). \end{aligned}$$

■

4.7.4 $\widetilde{P}^{(1)}$ vs. $\widehat{P}^{(1)}$

Lemma 4.7.15 $\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right| = O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$ *with high probability.*

Proof: Let W be the rotation matrix in Theorem 4.7.14, then

$$\begin{aligned}\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right| =& \left|\widehat{Z}_i^T\widehat{Z}_j - \widehat{Z}_i^T W Z_j + \widehat{Z}_i^T W Z_j - (WZ_i)^T W Z_j\right| \\ \leq& \left|\widehat{Z}_i^T(\widehat{Z}_j - WZ_j) + (\widehat{Z}_i^T - (WZ_i)^T)WZ_j\right| \\ \leq& \|\widehat{Z}_i\|_2\|\widehat{Z}_j - WZ_j\|_2 + \|Z_j\|_2\|\widehat{Z}_i^T - (WZ_i)^T\|_2.\end{aligned}$$

Since $\|Z_i\|_2^2 = Z_i^T Z_i = H_{ii}^{(1)} = E[\widehat{P}_{ii}^{(1)}] = (1-\epsilon)P_{ij} + \epsilon C_{ij} \leq R$, we have $\|Z_i\|_2 = O(1)$. Combined with Theorem 4.7.14,

$$\begin{aligned}\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right| =& (\|\widehat{Z}_i\|_2 + \|Z_j\|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ \leq& (\|\widehat{Z}_i - WZ_i\|_2 + \|WZ_i\|_2 + \|Z_j\|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ =& O(m^{-1/2}n^{-1/2}(\log n)^{3/2})\end{aligned}$$

with high probability. ■

Definition 4.7.16 *Define* $\widetilde{P}_{ij}^{(1)} = (\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}}$, *our estimator for* P_{ij}, *to be a projection of* $\widehat{Z}_i^T\widehat{Z}_j$ *onto* $[0, \min(\widehat{P}_{ij}^{(1)}, R)]$.

Remark 4.7.17 *The truncation step above to construct estimator is only for technical reasons. Since the constant R could be arbitrarily large, we do not need this truncation step in practice. Note that Theorem 4.3.3 still holds with this modified estimator. And all our simulation and real data experiment do not contain this truncation procedure.*

Lemma 4.7.18 (Theorem 4.3.3 Part 1) *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLE has the same entry-wise asymptotic bias as MLE, i.e.*

$$\lim_{n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(1)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} \text{Bias}(\widehat{P}_{ij}^{(1)}).$$

Proof: Fix some $a > 0$, we have

$$\begin{aligned} & E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|] \\ =& E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij} \le a\}] + E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij} > a\}] \end{aligned}$$

For the first term, we have

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}] \\
\le &E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}\mathbb{I}\{\text{Lemma 4.7.15 holds}\}] \\
&+ E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}\mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\le &E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 holds}] \\
&+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le &O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\
&+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - \widehat{P}_{ij}|\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
&+ n^{-c}E[|\widehat{P}_{ij} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le &O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + n^{-c}E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
&+ n^{-c}E[(\widehat{P}_{ij} + R)\mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le &O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + an^{-c} + (a+R)n^{-c} \\
\le &O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a+R).
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij} > a\}] = E[\left(\frac{1}{m}\sum_{1\le t\le m} A_{ij}^{(t)}\right)\mathbb{I}\{\widehat{P}_{ij} > a\}] \\
=&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\mathbb{I}\{\widehat{P}_{ij} > a\}] \le \frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\mathbb{I}\{\max_{1\le s\le m} A_{ij}^{(s)} > a\}] \\
\le&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\left(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)} > a\}\right)] = E[A_{ij}^{(1)}\left(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)} > a\}\right)] \\
=&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)} > a\})] + (m-1)E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(2)} > a\})] \\
=&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)} > a\})] + (m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)} > a),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&E[(\widehat{P}_{ij} + R)\mathbb{I}\{\widehat{P}_{ij} > a\}] \\
=&E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij} > a\}] + R\cdot P(\widehat{P}_{ij} > a) \\
\le&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)} > a\})] + (m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)} > a) + R\cdot m\cdot P(A_{ij}^{(1)} > a).
\end{aligned}
$$

Thus for the second term,

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq &E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}| \mathbb{I}\{\widehat{P}_{ij} > a\}] + E[|\widehat{P}_{ij} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq &E[\widehat{P}_{ij} \mathbb{I}\{\widehat{P}_{ij} > a\}] + E[(\widehat{P}_{ij} + R) \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq &2E[A_{ij}^{(1)} \mathbb{I}\{A_{ij}^{(1)} > a\})] + 2(m-1) E[A_{ij}^{(1)}] P(A_{ij}^{(1)} > a) \\
&+ R \cdot m \cdot P(A_{ij}^{(1)} > a) \\
\leq &2e^{-a/R}(a + 2mR).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] \\
\leq &O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + 2n^{-c}(a + R) + 2e^{-a/R}(a + 2mR).
\end{aligned}
$$

Let $a = m^{-1} n^{2b}$ for any $b > 0$, and $c = 2b + 3$, combined with the assumption

$m = O(n^b)$, we have

$$
\begin{aligned}
& E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] \\
=& O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b}) \cdot O(e^{-m^{-1} n^{2b}}) \\
=& O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b}) \cdot O(e^{-n^b}) \\
=& O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b}) \cdot O(n^{-2b-3}) \\
=& O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) \\
=& O(m^{-1/2} n^{-1/2} (\log n)^{3/2}).
\end{aligned}
$$

■

Theorem 4.7.19 *Assuming that* $m = O(n^b)$ *for any* $b > 0$, *then* $\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) = O(m^{-1} n^{-1} (\log n)^3)$.

Proof: By Lemma 4.7.15,

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) =& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j + Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])] \\
\leq& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2\sqrt{E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2]E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2]} \\
\leq& 4E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2].
\end{aligned}$$

Fix some $a > 0$, we have

$$\begin{aligned}
& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij} \leq a\}] + E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij} > a\}].
\end{aligned}$$

For the first term, we have

$$
\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}] \\
\le& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}\mathbb{I}\{\text{Lemma 4.7.15 holds}\}] \\
&+ E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}\mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\le& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 holds}] \\
&+ n^{-c} E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le& O(m^{-1}n^{-1}(\log n)^3) \\
&+ 2n^{-c} E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij})^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le& O(m^{-1}n^{-1}(\log n)^3) + 2n^{-c} E[\widehat{P}_{ij}^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij} + R)^2 \mathbb{I}\{\widehat{P}_{ij} \le a\}|\text{Lemma 4.7.15 does not hold}] \\
\le& O(m^{-1}n^{-1}(\log n)^3) + 2a^2 n^{-c} + 2(a+R)^2 n^{-c} \\
\le& O(m^{-1}n^{-1}(\log n)^3) + 4n^{-c}(a+R)^2.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&E[\widehat{P}_{ij}^2\mathbb{I}\{\widehat{P}_{ij}>a\}] = E[(\frac{1}{m}\sum_{1\le t\le m} A_{ij}^{(t)})^2\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
\le&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)2}\mathbb{I}\{\widehat{P}_{ij}>a\}] \le \frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)2}\mathbb{I}\{\max_{1\le s\le m} A_{ij}^{(s)}>a\}] \\
\le&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)2}(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)}>a\})] = E[A_{ij}^{(1)2}(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)}>a\})] \\
=&E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(2)}>a\})] \\
=&E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)2}]P(A_{ij}^{(1)}>a),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&E[(\widehat{P}_{ij}+R)^2\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
=&E[\widehat{P}_{ij}^2\mathbb{I}\{\widehat{P}_{ij}>a\}] + 2R\cdot E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij}>a\}] + R^2P(\widehat{P}_{ij}>a) \\
\le&E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)2}]P(A_{ij}^{(1)}>a) \\
&+2R\left(E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a)\right) \\
&+R^2\cdot m\cdot P(A_{ij}^{(1)}>a).
\end{aligned}
$$

Thus for the second term,

$$
\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq & 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij})^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] + 2E[(\widehat{P}_{ij} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq & 2E[\widehat{P}_{ij}^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] + 2E[(\widehat{P}_{ij} + R)^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq & 4E[A_{ij}^{(1)2} \mathbb{I}\{A_{ij}^{(1)} > a\})] + 4(m-1)E[A_{ij}^{(1)2}] P(A_{ij}^{(1)} > a) \\
& + 4R \cdot E[A_{ij}^{(1)} \mathbb{I}\{A_{ij}^{(1)} > a\})] + 2R(m-1)E[A_{ij}^{(1)}] P(A_{ij}^{(1)} > a) \\
& + 2R^2 \cdot m \cdot P(A_{ij}^{(1)} > a) \\
\leq & 4e^{-a/R} \left(a^2 + 3Ra + 3(m+1)R^2\right) \\
\leq & 4e^{-a/R}(a + 2m^{1/2}R)^2.
\end{aligned}
$$

Thus,

$$
\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) \leq O(m^{-1}n^{-1}(\log n)^3) + 16(a+R)^2 n^{-c} + 16(a + 2m^{1/2}R)^2 e^{-a/R}.
$$

Let $a = m^{-1/2}n^b$ for any $b > 0$, and $c = 2b + 3$, combined with the assumption

$m = O(n^b)$, we have

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) =& O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(e^{-m^{-1/2}n^b}) \\
=& O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(e^{-n^{b/2}}) \\
=& O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(n^{-2b-3}) \\
=& O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) \\
=& O(m^{-1}n^{-1}(\log n)^3).
\end{aligned}$$

■

Theorem 4.7.20 (Theorem 4.3.3 Part 2) *Assuming that $m = O(n^b)$ for any $b > 0$, then for $1 \leq i, j \leq n$ and $i \neq j$,*

$$\frac{\mathrm{Var}(\widetilde{P}_{ij}^{(1)})}{\mathrm{Var}(\widehat{P}_{ij}^{(1)})} = O(n^{-1}(\log n)^3).$$

And thus

$$\mathrm{ARE}(\widehat{P}_{ij}^{(1)}, \widetilde{P}_{ij}^{(1)}) = 0.$$

Proof: The results are direct from Theorem 4.7.19 and Theorem 4.3.1. ■

4.7.5 $\widetilde{P}^{(q)}$ vs. $\widehat{P}^{(q)}$

Theorem 4.7.21 *Let P and C be two n-by-n symmetric and hollow matrices satisfying element-wise conditions $0 < P_{ij} \leq C_{ij} \leq R$ for some constant $R > 0$. For $0 < \epsilon < 1$, we define m symmetric and hollow matrices as*

$$A^{(t)} \overset{iid}{\sim} (1-\epsilon)\text{Exp}(P) + \epsilon\text{Exp}(C)$$

for $1 \leq t \leq m$. Let $\widehat{P}^{(q)}$ be the entry-wise MLqE based on exponential distribution with m observations. Define $H^{(q)} = E[\widehat{P}^{(q)}]$, then for any constant $c > 0$ there exists another constant $n_0(c)$, independent of n, P, C and ϵ, such that if $n > n_0$, then for all η satisfying $n^{-c} \leq \eta \leq 1/2$,

$$P\left(\|\widehat{P}^{(q)} - H^{(q)}\|_2 \leq 8R\sqrt{2n\ln(n/\eta)}\right) \geq 1-\eta.$$

Proof: Similar to the proof of Theorem 4.7.9.

By Lemma 4.7.2 we have

$$\begin{aligned}
\left|\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)}\right| =& \left|\widehat{P}_{ij}^{(q)} - \widehat{P}_{ij}^{(1)} + \widehat{P}_{ij}^{(1)} - H_{ij}^{(1)} + H_{ij}^{(1)} - H_{ij}^{(q)}\right| \\
\leq& \left|\widehat{P}_{ij}^{(q)} - \widehat{P}_{ij}^{(1)}\right| + \left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + \left|H_{ij}^{(1)} - H_{ij}^{(q)}\right| \\
\leq& \widehat{P}_{ij}^{(1)} + \left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)} \\
\leq& 2\left(\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)}\right).
\end{aligned}$$

Also,

$$\begin{aligned}
E[(\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)})^k] \leq & E\left[\left|\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)}\right|^k\right] \\
\leq & 2^k E\left[\left(\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)}\right)^k\right] \\
\leq & 2^k \sum_{s=0}^{k} \binom{k}{s} E\left[\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right|^s\right] \left(H_{ij}^{(1)}\right)^{k-s} \\
\leq & 2^k \sum_{s=0}^{k} \binom{k}{s} R^s s! \left(H_{ij}^{(1)}\right)^{k-s} \\
\leq & 2^k k! \sum_{s=0}^{k} \binom{k}{s} R^s \left(H_{ij}^{(1)}\right)^{k-s} \\
= & 2^k k! \left(R + H_{ij}^{(1)}\right)^k \\
\leq & 2^{2k} k! R^k. \qquad (4.5)
\end{aligned}$$

Therefore we have

$$P\left(\|\widehat{P}^{(q)} - H^{(q)}\| \geq t\right) \leq n \exp\left(-\frac{t^2/2}{32R^2 n + Rt}\right).$$

Now let $c > 0$ be given and assume $n^{-c} \leq \eta \leq 1/2$. Then there exists a $n_0(c)$ independent of n, P, C and ϵ such that whenever $n > n_0(c)$,

$$t = 8R\sqrt{2n \ln(n/\eta)} \leq 32Rn.$$

Plugging this t into the equation above, we get

$$P(\|\widehat{P}^{(q)} - H^{(q)}\| \geq 8R\sqrt{2n\ln(n/\eta)}) \leq n\exp\left(-\frac{t^2}{64R^2n}\right) = \eta.$$

■

As we define $H^{(q)} = E[\widehat{P}^{(q)}]$, let $d^{(q)} = \text{rank}(H^{(q)})$ be the dimension in which we are going to embed $\widehat{P}^{(q)}$. Notice that it is less than or equal to $K \times K'$ based on the SBM assumption. Then we can define $H^{(q)} = ZZ^T$ where $Z \in \mathbb{R}^{n\times d^{(q)}}$.

For simplicity, from now on, we will use $\widehat{P}$ to represent $\widehat{P}^{(q)}$, use H to represent $H^{(q)}$ and use k to represent the dimension $d^{(q)}$ we are going to embed. Assume $H = USU^T = ZZ^T$, where $Z = [Z_1, \cdots, Z_n]^T$ is a n-by-k matrix. Then our estimate for Z up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^T$ is the rank-d spectral decomposition of $|\widehat{P}| = (\widehat{P}^T\widehat{P})^{1/2}$.

Furthermore, we assume that the second moment matrix $E[Z_1Z_1^T]$ is rank k and has distinct eigenvalues $\lambda_i(E[Z_1Z_1^T])$. In particular, we assume that there exists $\delta > 0$ such that

$$\delta < \lambda_k(E[Z_1Z_1^T])$$

Lemma 4.7.22 *Under the above assumptions, $\lambda_i(H) = \Theta(n)$ with high probability when $i \leq k$, i.e. the largest k eigenvalues of H is of order n. Moreover, we have $\|S\|_2 = \Theta(n)$ and $\|\widehat{S}\|_2 = \Theta(n)$ with high probability.*

Proof: Exactly the same as proof for Lemma 4.7.10. ■

Lemma 4.7.23 *Let $W_1 \Sigma W_2^T$ be the singular value decomposition of $U^T \widehat{U}$. Then for sufficiently large n,*

$$\|U^T \widehat{U} - W_1 W_2^T\|_F = O(n^{-1} \log n)$$

with high probability.

Proof: Exactly the same as proof for Lemma 4.7.11. ∎

We will denote the orthogonal matrix $W_1 W_2^T$ by W^*.

Lemma 4.7.24 *For sufficiently large n,*

$$\|W^* \widehat{S} - S W^*\|_F = O(\log n),$$

$$\|W^* \widehat{S}^{1/2} - S^{1/2} W^*\|_F = O(n^{-1/2} \log n)$$

and

$$\|W^* \widehat{S}^{-1/2} - S^{-1/2} W^*\|_F = O(n^{-3/2} \log n)$$

with high probability.

Proof: Similar to the proof of Lemma 4.7.12. ∎

Lemma 4.7.25 *There exists a rotation matrix W such that for sufficiently large n,*

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P} - H) U S^{-1/2}\|_F + O(n^{-1/2} (\log n)^{3/2})$$

with high probability.

Proof: Exactly the same as proof for Lemma 4.7.13. ■

Theorem 4.7.26 *There exists a rotation matrix W such that for sufficiently large n,*

$$\max_i \|\widehat{Z}_i - W Z_i\|_2 = O(n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: Similar to the proof of Theorem 4.7.14. ■

Lemma 4.7.27 $\left|\widehat{Z}_i^T \widehat{Z}_j - Z_i^T Z_j\right| = O(n^{-1/2}(\log n)^{3/2})$ *with high probability.*

Proof: Similar to the proof of Lemma 4.7.15. ■

Definition 4.7.28 *Define $\widetilde{P}_{ij}^{(q)} = (\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}}$, our estimator for P_{ij}, to be a projection of $\widehat{Z}_i^T \widehat{Z}_j$ onto $[0, \min(\widehat{P}_{ij}^{(q)}, R)]$.*

Lemma 4.7.29 (Theorem 4.3.4 Part 1) *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has the same entry-wise asymptotic bias as MLqE, i.e.*

$$\lim_{n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} \text{Bias}(\widehat{P}_{ij}^{(q)}).$$

Proof: Fix some $a > 0$, we have

$$\begin{aligned} & E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|] \\ =& E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] + E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}]. \end{aligned}$$

Note that we are thresholding according to $\widehat{P}^{(1)}$ instead of $\widehat{P}^{(q)}$. By Lemma 4.7.2, we

know $\widehat{P}^{(q)} < \widehat{P}^{(1)}$ given any data. For the first term, we have

$$
\begin{aligned}
& E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] \\
\leq & E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 holds}\}] \\
& + E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}] \\
\leq & E[(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \{\text{Lemma 4.7.27 holds}] \\
& + n^{-c} E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq & O(n^{-1/2} (\log n)^{3/2}) \\
& + n^{-c} E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - \widehat{P}_{ij}^{(q)}| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
& + n^{-c} E[|\widehat{P}_{ij}^{(q)} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq & O(n^{-1/2} (\log n)^{3/2}) \\
& + n^{-c} E[\widehat{P}_{ij}^{(q)} \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
& + n^{-c} E[(\widehat{P}_{ij}^{(q)} + R) \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq & O(n^{-1/2} (\log n)^{3/2}) \\
& + n^{-c} E[\widehat{P}_{ij}^{(1)} \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
& + n^{-c} E[(\widehat{P}_{ij}^{(1)} + R) \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq & O(n^{-1/2} (\log n)^{3/2}) + a n^{-c} + (a + R) n^{-c} \\
\leq & O(n^{-1/2} (\log n)^{3/2}) + 2 n^{-c} (a + R).
\end{aligned}
$$

For the second term, we have

$$\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq &E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}^{(q)}|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[|\widehat{P}_{ij}^{(q)} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq &E[\widehat{P}_{ij}^{(q)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[(\widehat{P}_{ij}^{(q)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq &E[\widehat{P}_{ij}^{(1)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[(\widehat{P}_{ij}^{(1)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq &2e^{-a/R}(a + 2mR).
\end{aligned}$$

Similarly, assuming $m = O(n^b)$ for any $b > 0$, we have

$$E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] = O(n^{-1/2}(\log n)^{3/2}).$$

■

Theorem 4.7.30 *Assuming that $m = O(n^b)$ for any $b > 0$, then* $\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) = O(n^{-1}(\log n)^3)$.

Proof: By Lemma 4.7.27,

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) =& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j + Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])] \\
\leq& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2\sqrt{E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2]E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2]} \\
\leq& 4E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2].
\end{aligned}$$

Fix some $a > 0$, we have

$$\begin{aligned}
& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] + E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}].
\end{aligned}$$

Note that we are thresholding according to $\widehat{P}^{(1)}$ instead of $\widehat{P}^{(q)}$. By Lemma 4.7.2, we

know $\widehat{P}^{(q)} < \widehat{P}^{(1)}$ given any data. For the first term, we have

$$
\begin{aligned}
&E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] \\
\leq &E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 holds}\}] \\
&+ E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}] \\
\leq &E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \{\text{Lemma 4.7.27 holds}] \\
&+ n^{-c} E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq &O(n^{-1} (\log n)^3) \\
&+ 2n^{-c} E[((\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - \widehat{P}_{ij}^{(q)})^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij}^{(q)} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq &O(n^{-1} (\log n)^3) \\
&+ 2n^{-c} E[\widehat{P}_{ij}^{(q)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij}^{(q)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq &O(n^{-1} (\log n)^3) + 2n^{-c} E[\widehat{P}_{ij}^{(1)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij}^{(1)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\} | \text{Lemma 4.7.27 does not hold}] \\
\leq &O(n^{-1} (\log n)^3) + 2a^2 n^{-c} + 2(a + R)^2 n^{-c} \\
\leq &O(n^{-1} (\log n)^3) + 4n^{-c} (a + R)^2.
\end{aligned}
$$

For the second term, we have

$$\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}^{(q)})^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(q)} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[\widehat{P}_{ij}^{(q)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(q)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[\widehat{P}_{ij}^{(1)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(1)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 4e^{-a/R}(a + 2m^{1/2}R)^2.
\end{aligned}$$

Similarly, assuming $m = O(n^b)$ for any $b > 0$, we have

$$\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) = O(n^{-1}(\log n)^3).$$

■

Theorem 4.7.31 (Theorem 4.3.4 Part 2) *Assuming that $m = O(n^b)$ for any $b > 0$, then for $1 \leq i, j \leq n$ and $i \neq j$,*

$$\frac{\mathrm{Var}(\widetilde{P}_{ij}^{(q)})}{\mathrm{Var}(\widehat{P}_{ij}^{(q)})} = O(mn^{-1}(\log n)^3).$$

Moreover, if $m = o(n(\log n)^{-3})$, then

$$\mathrm{ARE}(\widehat{P}_{ij}^{(q)}, \widetilde{P}_{ij}^{(q)}) = 0.$$

Proof: The results are direct from Theorem 4.7.30 and Theorem 4.3.1. ■

4.7.6 $\widetilde{P}^{(q)}$ vs. $\widetilde{P}^{(1)}$

Theorem 4.7.32 *For sufficiently large C and any $1 \leq i, j \leq n$, if $m = O(n^b)$ for any $b > 0$, then*

$$\lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(1)}) > \lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)})$$

Proof: Direct result from Theorem 4.3.1, Theorem 4.3.3 and Theorem 4.3.4. ■

Theorem 4.7.33 *For sufficiently large C and any $1 \leq i, j \leq n$, if $m = O(n(\log n)^{-3})$, then*

$$\lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(q)}) = 0.$$

Proof: Direct result from Theorem 4.3.3 and Theorem 4.3.4. ■

4.7.7 Other Proofs

Lemma 4.7.34 *Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ with f to be Poisson, then $E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\widehat{P}^{(1)}$ is the entry-wise MLE as defined before.*

Proof: First we prove $(x-\theta)^k \leq k!(e^{x-\theta} + e^{\theta-x})$.

1. k is even. Then by Taylor expansion, $e^{x-\theta} + e^{\theta-x} \geq \frac{(x-\theta)^k}{k!}$
2. k is odd. When $x \geq \theta$, still by Taylor expansion, $(x-\theta)^k \leq k!e^{x-\theta}$. When $x < \theta$, $(x-\theta)^k < 0 \leq k!e^{x-\theta}$.

Thus $(x-\theta)^k \leq k!(e^{x-\theta}+e^{\theta-x})$. So the k-th central moment of Poisson distribution with parameter θ is bounded by

$$\begin{aligned} E[(X-\theta)^k] &\leq k!\left(E[e^{X-\theta}]+E[e^{\theta-X}]\right) \\ &= k!\left(e^{-\theta}E[e^X]+e^{\theta}E[e^{-X}]\right) \\ &= k!\left(e^{\theta(e-2)}+e^{\theta e^{-1}}\right). \end{aligned}$$

Let $X_1 \sim \text{Poisson}(P_{ij})$ and $X_2 \sim \text{Poisson}(C_{ij})$. Then if A_{ij} is distributed from a mixture model as in the statement, we have

$$\begin{aligned} &E[(A_{ij}-E[\widehat{P}_{ij}^{(1)}])^k] \\ =&(1-\epsilon)E[(X_1-P_{ij}+P_{ij}-E[\widehat{P}_{ij}^{(1)}])]+\epsilon E[(X_2-C_{ij}+C_{ij}-E[\widehat{P}_{ij}^{(1)}])] \\ =&(1-\epsilon)\sum_{j=0}^{k}\binom{k}{j}(P_{ij}-E[\widehat{P}_{ij}^{(1)}])^{k-j}E[(X_1-P_{ij})^j] \\ &+\epsilon\sum_{j=0}^{k}\binom{k}{j}(C_{ij}-E[\widehat{P}_{ij}^{(1)}])^{k-j}E[(X_2-C_{ij})^j] \\ \leq&(1-\epsilon)\sum_{j=0}^{k}\binom{k}{j}(P_{ij}-E[\widehat{P}_{ij}^{(1)}])^{k-j}\cdot j!\cdot \text{const} \\ &+\epsilon\sum_{j=0}^{k}\binom{k}{j}(C_{ij}-E[\widehat{P}_{ij}^{(1)}])^{k-j}\cdot j!\cdot \text{const} \\ \leq&(1-\epsilon)k!\cdot \text{const}^k+\epsilon k!\cdot \text{const}^k \\ \leq&\text{const}^k\cdot k!. \end{aligned}$$

■

Chapter 5

Discussion

Bibliography

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9 (Sep):1981–2014, 2008.

Avanti Athreya, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, 2016.

Peter J Bickel and Kjell A Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume i. Pearson Prentice Hall, second edition, 2007.

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

Gerald G Brown and Herbert C Rutemiller. Means and variances of stochastic vector products with applications to random linear models. *Management Science*, 24(2): 210–216, 1977.

BIBLIOGRAPHY

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

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Davide Ferrari and Yuhong Yang. Maximum Lq-likelihood estimation. *The Annals of Statistics*, 38(2):753–783, 2010.

Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959.

Cedric E Ginestet, Prakash Balanchandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *arXiv preprint arXiv:1407.5525*, 2014.

Krzysztof J Gorgolewski, Natacha Mendes, Domenica Wilfling, Elisabeth Wladimirow, Claudine J Gauthier, Tyler Bonnen, Florence JM Ruby, Robert Trampel, Pierre-Louis Bazin, Roberto Cozatl, et al. A high resolution 7-tesla resting-state fmri test-retest dataset with cognitive and physiological measures. *Scientific data*, 2, 2015.

Sam Gutmann. Stein's paradox is impossible in problems with finite sample space. *The Annals of Statistics*, pages 1017–1020, 1982.

BIBLIOGRAPHY

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

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif., 1961.

Xiaoyi Jiang, Andreas Müunger, and Horst Bunke. On median graphs: Properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1144–1151, 2001.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Gregory Kiar, William Gray Roncal, Disa Mhembere, Eric Bridgeford, Randal Burns, and Joshua T. Vogelstein. ndmg: NeuroData's MRI graphs pipeline, 2016. URL `https://doi.org/10.5281/zenodo.60206`.

R. S. H. Mah and A. C. Tamhane. Detection of gross errors in process data. *AIChE Journal*, 28(5):828–830, 1982. ISSN 1547-5905. doi: 10.1002/aic.690280519. URL `http://dx.doi.org/10.1002/aic.690280519`.

BIBLIOGRAPHY

David Marchette, Carey Priebe, and Glen Coppersmith. Vertex nomination via attributed random dot product graphs. In *Proceedings of the 57th ISI World Statistics Congress*, volume 6, page 16, 2011.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

Yichen Qin and Carey E Priebe. Robust hypothesis testing via Lq-likelihood. *Statistica Sinica preprint SS-2015-0441R2*, 2017.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

William Gray Roncal, Zachary H Koterba, Disa Mhembere, Dean M Kleissas, Joshua T Vogelstein, Randal Burns, Anita R Bowles, Dimitrios K Donavos, Sephira Ryman, Rex E Jung, et al. MIGRAINE: MRI graph reliability analysis and inference for connectomics. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 313–316. IEEE, 2013.

BIBLIOGRAPHY

Edward R Scheinerman and Kimberly Tucker. Modeling graphs using dot product representations. *Computational Statistics*, 25(1):1–16, 2010.

Robert Serfling. Asymptotic relative efficiency in estimation. In *International encyclopedia of statistical science*, pages 68–72. Springer, 2011.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. I*, pages 197–206. University of California Press, Berkeley and Los Angeles, 1956.

Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2014.

Runze Tang, Michael Ketcha, Joshua T Vogelstein, Carey E Priebe, and Daniel L Sussman. Law of large graphs. *arXiv preprint arXiv:1609.01672*, 2016.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

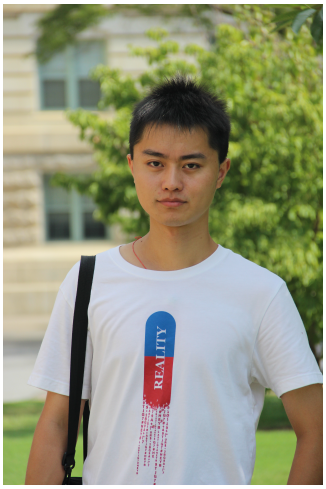
G. V. Trunk. A problem of dimensionality: A simple example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (3):306–307, 1979.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot

via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2): 918–930, 2006.

Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1, 2014.

Vita



Runze Tang received the B.S. degree in mathematics and applied mathematics from the University of Science and Technology of China in 2012, the M.S.E. degree in computer science from the Johns Hopkins University in 2015 and enrolled in the Applied Mathematics and Statistics Ph.D. program at Johns Hopkins University in 2012. Runze has received the Teaching Fellow Recognition Award.