Robust Estimation from Multiple Graphs

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

© Runze Tang 2017

All rights reserved
Abstract

Estimation of graph parameters based on a collection of graphs is essential for a wide range of graph inference tasks. Often, this problem is especially difficult because the sample size is relatively small as compared to the number of parameters to estimate. While using the element-wise sample mean of the adjacency matrices is a common approach, this method does not exploit any underlying structural properties of the graphs. Another challenge for this estimation problem is that, in practice, graphs are generally observed with edge contamination. We consider a weighted latent position graph model contaminated via an edge weight gross error model and propose an estimation methodology based on robust $L_q$ estimation followed by low-rank adjacency spectral decomposition. We demonstrate that, under appropriate conditions, our estimator both maintains $L_q$ robustness and wins the bias-variance trade-off by exploiting low-rank graph structure. We illustrate the improvement offered by our estimator via both simulations and a human connectome data experiment.

Primary Reader: Carey E. Priebe
ABSTRACT

Secondary Reader: Minh Tang
Acknowledgments

The past five years in graduate school has been the most wonderful and enjoyable experience in my life. I am grateful for the company of all these lovely people around me. My thanks belong first and foremost to the best advisor in the world, Carey Priebe, for his unconditional support and valuable advice at all levels. I would also like to thank Minh Tang, Joshua Vogelstein, and Daniel Sussman, who each provided varied, but consistently invaluable, guidance. I thank my co-authors and collaborators, including Avanti Athreya, Vince Lyzinski, Youngser Park, Nam Lee, Keith Levin, Greg Kiar, Shangsi Wang, Cencheng Shen, Donniell Fishkind, Bruno Jedynak and many others, who all offered their help with patience in discussions over the past few years. Special thanks go to Randal Burns, Brian Caffo, and Vladimir Braverman for their time and great suggestions in my GBO exam, which influenced this thesis for the better. I also want express my appreciation to the faculty and staff of the Applied Maths and Statistics Department at Johns Hopkins University. In particular, my thanks go to Daniel Naiman, Edward Scheinerman, Daniel Robinson, John Wierman, and John Miller for their kindness advice; and to Kristin Bechtel, Sandy Kirt,
ACKNOWLEDGMENTS

and Ann Gibbins for their help. My deepest thanks go to my fellow students and friends at Hopkins for coffee, soccer games, and all the time we have spent together. Finally, my warmest thanks go to my loving parents, Zhigui Tang and Juan Chen, who have supported me in every way and make me who I am.
Dedication

This thesis is dedicated to my parents, for their endless love.
Contents

Abstract ii

Acknowledgments iv

List of Figures xi

1 Introduction 1

2 Random Graph Models 6

2.1 Graphs and Random Graphs . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.1 Basic Concepts of Graphs . . . . . . . . . . . . . . . . . . . . . 7

2.1.2 Random Graphs . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2 Unweighted Random Graph Models . . . . . . . . . . . . . . . . . . 9

2.2.1 Independent Edge Model . . . . . . . . . . . . . . . . . . . . . 9

2.2.2 Random Dot Product Graph . . . . . . . . . . . . . . . . . . . 10

2.2.3 Stochastic Blockmodel . . . . . . . . . . . . . . . . . . . . . . 14

2.3 Weighted Random Graph Models . . . . . . . . . . . . . . . . . . . 18
CONTENTS

2.3.1 Weighted Independent Edge Model ................................................. 18
2.3.2 Weighted Random Dot Product Graph ............................................. 19
2.3.3 Weighted Stochastic Blockmodel ..................................................... 20

3 Estimation from Multiple Graphs and a Law of Large Graphs ................. 22

3.1 Model .............................................................................................. 29
3.2 Methods ........................................................................................... 30
  3.2.1 Adjacency Spectral Embedding ..................................................... 30
  3.2.2 Choosing Dimension .................................................................... 31
  3.2.3 Graph Diagonal Augmentation ...................................................... 32
3.3 Estimators ......................................................................................... 33
  3.3.1 Element-wise sample mean ............................................................ 34
  3.3.2 Low-Rank Estimator ..................................................................... 35
3.4 Theoretical Results ............................................................................ 38
3.5 Finite Sample Toy Model Simulations ................................................. 48
3.6 SWU4 Brain Graphs Experiment ....................................................... 51
  3.6.1 Dataset Description ....................................................................... 52
  3.6.2 Experiment Results ...................................................................... 53
  3.6.3 Exploration of Dimension Selection Procedures ............................ 58
  3.6.4 Interpretability of Low-rank Methods ............................................. 59
  3.6.5 Challenges of the SWU4 Dataset .................................................. 63
  3.6.6 Lobe Structure behind the Low-rank Methods ............................... 67

viii
CONTENTS

4.7 Proofs for Theory Results .......................... 114

4.7.1 Outline of the Proofs ............................. 114

4.7.2 \( \hat{\mathcal{P}}(q) \) vs. \( \hat{\mathcal{P}}(1) \) .................. 116

4.7.3 ASE Procedure of \( \hat{\mathcal{P}}(1) \) ..................... 123

4.7.4 \( \hat{\mathcal{P}}(1) \) vs. \( \tilde{\mathcal{P}}(1) \) ...................... 140

4.7.5 \( \mathcal{P}(q) \) vs. \( \mathcal{P}(q) \) ......................... 151

4.7.6 \( \mathcal{P}(q) \) vs. \( \mathcal{P}(1) \) ......................... 162

4.7.7 Other Proofs .............................. 162

5 Discussion ........................................ 165

Vita .................................................. 178
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Example illustrating the stochastic blockmodel</td>
</tr>
<tr>
<td>3.1</td>
<td>Heat maps of the population mean, the sample mean, and the low-rank estimator</td>
</tr>
<tr>
<td>3.2</td>
<td>Example illustrating different estimates under the stochastic blockmodel</td>
</tr>
<tr>
<td>3.3</td>
<td>Asymptotic scaled relative efficiency in a 2-block SBM</td>
</tr>
<tr>
<td>3.4</td>
<td>Finite sample relative efficiency based on simulations</td>
</tr>
<tr>
<td>3.5</td>
<td>Relative efficiencies of two estimators for the SWU4 dataset</td>
</tr>
<tr>
<td>3.6</td>
<td>Heat plots of absolute estimation error for both estimators</td>
</tr>
<tr>
<td>3.7</td>
<td>Top 5 regions of the brain and top 50 connections between regions with the largest differences between two estimators</td>
</tr>
<tr>
<td>3.8</td>
<td>Comparison of MSE of two estimators for three atlases at three sample sizes for the SWU4 data</td>
</tr>
<tr>
<td>3.9</td>
<td>Brain plots colored separately for each of the first 4 dimensions of embedding under the Desikan atlas</td>
</tr>
<tr>
<td>3.10</td>
<td>Relative error of the low-rank approximation of the population mean</td>
</tr>
<tr>
<td>3.11</td>
<td>Histogram of the population mean</td>
</tr>
<tr>
<td>3.12</td>
<td>Histogram of mean graph for Desikan atlas</td>
</tr>
<tr>
<td>3.13</td>
<td>Violin plot of the permutation test</td>
</tr>
<tr>
<td>3.14</td>
<td>Comparison of two estimators for synthetic data analysis</td>
</tr>
<tr>
<td>4.1</td>
<td>Roadmap among the data and four estimators</td>
</tr>
<tr>
<td>4.2</td>
<td>Relationships among our four estimators</td>
</tr>
<tr>
<td>4.3</td>
<td>MSE for different contamination ratio $\epsilon$</td>
</tr>
<tr>
<td>4.4</td>
<td>MSE for different parameter $q$</td>
</tr>
<tr>
<td>4.5</td>
<td>Screeplot and histogram of the normalized eigenvalues of the mean of 114 graphs based on ndmg1 pipeline</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Network analysis has emerged as an area of intense statistical theory and application activity. In the general parametric framework, \( G \sim f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\} \), and selecting a principled and productive estimator \( \hat{\theta} \) for the unknown graph parameter \( \theta \) 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.

When considering unweighted graphs, it is equivalent to estimating the mean in many scenarios. 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
CHAPTER 1. INTRODUCTION

space to more complex objects like images, shapes, and documents. In this work, we consider the challenges of estimating a population mean based on a sample of graphs – for example, 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

\[\text{Trunk} \ [1979].\]

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 the first part of this work, we consider a series of (unweighted) random graph models, with a particular focus on the ones which assume that the connections depend on the hidden properties of the corresponding objects. By proposing a low-rank estimator, we are able to improve the performance of an edge-wise sample mean.

When we shift our focus to weighted graphs, 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) \cite{bollobas2007} (described in Section 2.3.1 below) – 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 to be (approximately) low-rank, then by biasing towards a low-rank structure, more elaborate estimators can have greatly reduced variance and win the bias-variance tradeoff. For our connectome data (Section 4.6 Figure 4.5) we observe this approximate low-rank property. \cite{tang2016} 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
CHAPTER 1. INTRODUCTION

low-rank structure.

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; Qin and Priebe, 2013] considered in this paper.

To address these two deficiencies simultaneously, in the second part of this work, 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 trade-off by taking advantage of the low-rank structure.

We organize the work as follows:

• **Chapter 1** We introduce the estimation problem and give an overview of this work.

• **Chapter 2** A series of models for a single random graph are introduced. These important components will not only lead to our model for a collection of graphs but also motivate our estimators later.

• **Chapter 3** For multiple unweighted graphs, we propose using a low-rank method together with tools for dimension selection and diagonal augmentation to smooth the estimates and improve performance over the naive methodology.
CHAPTER 1. INTRODUCTION

for small sample sizes.

• **Chapter 4** The case of weighted graphs with contamination is investigated. We propose an estimation methodology based on robust $L_q$ estimation followed by low-rank adjacency spectral decomposition, which both maintains $L_q$ robustness and wins the bias-variance trade-off by exploiting a low-rank graph structure.

• **Chapter 5** The results and their extensions are discussed from different perspectives.
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 $a \in A$. For unweighted graphs, $A = \{0, 1\}^{n \times n}$ 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. $A = \mathbb{R}^{n \times n}_{\geq 0}$. 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.
CHAPTER 2. RANDOM GRAPH MODELS

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 $A : \Omega \mapsto A$ 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 $A_{ij} \sim \text{Bernoulli}(p)$ for $i, j \in [n]$ with $i < j$. Thus for $a \in A$ 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.
CHAPTER 2. RANDOM GRAPH MODELS

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. $A_{ij} \sim \text{ind} \text{ Bernoulli}(p_{ij})$. 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 $a \in A$ 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 \( n \ast (n - 1)/2 \) 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 \( 2^{n\ast(n-1)/2} \) 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 [Hoff et al., 2002] 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
CHAPTER 2. RANDOM GRAPH MODELS

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 $X_j$,

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

Thus for $a \in A$ 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) [Nickel, 2008, Young and Scheinerman, 2007].

**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 $X_j$,

$$A_{ij} | (X_i, X_j) \text{i.i.d.} \sim \text{Bernoulli}(X_i^\top X_j).$$
Thus for $a \in A$ 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 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 $n \ast (n - 1)/2$.

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 IEM$(P)$ with $P = XX^\top$, i.e. an edge between vertex $i$ and vertex $j$ is present with
CHAPTER 2. RANDOM GRAPH MODELS

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 $\kappa(x, y) = x^\top Ky$ 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d'} > 0 > \lambda_{d'+1} \geq \cdots \geq \lambda_d$.

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.
CHAPTER 2. RANDOM GRAPH MODELS

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 $i < j$,

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

Thus for $a \in A$ 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 $\tau$ 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 \( \tau \). 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

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

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
CHAPTER 2. RANDOM GRAPH MODELS

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 $\text{SBM}(P)$. 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 $\nu$ are given by $\nu_1^\top, \ldots, \nu_K^\top$. Then $\nu_k$ 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 $\text{SBM}(\rho, B)$ where $B = \nu^\top \nu$ with $\nu = [\nu_1, \ldots, \nu_K]^\top$. Similarly, we can define $\tau \in [K]^n$ such that $X_i = \nu_{\tau_i}$. 
CHAPTER 2. RANDOM GRAPH MODELS

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. \( A \in \mathbb{R}_{\geq 0}^{n \times n} \). 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 drawn from a Bernoulli distribution with parameter \( P_{ij} \) independent of all 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 drawn from \( f_{P_{ij}} \) independent of all other edges.

Thus IEM is a special case of WIEM, with \( \mathcal{F} \) representing the collection of Bernoulli distributions and \( \Theta = [0, 1] \).
CHAPTER 2. RANDOM GRAPH MODELS

Note that the graphs considered in this paper are undirected without self-loops, and the parameter matrix $P$ can be considered to be symmetric and hollow. That is, for convenience, we still define the parameters to be an $n$-by-$n$ matrix while only $n \times (n - 1)/2$ of them are active.

2.3.2 Weighted Random Dot Product Graph

As discussed in Section 2.2.2, vertices are selective about their adjacencies in graphs. A vertex may be frequently adjacent to one group of vertices but rarely adjacent to the other group of 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 via consideration of latent position matrix $X \in \mathbb{R}^{n \times d}$ 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 $X$, the entries of the adjacency matrix $A$ are independent and $A_{ij}$ is a random variable following distribution $f_\theta \in \mathcal{F}$ with parameter $\theta = X_i^\top X_j$ for all $i < j \in [n]$.

Under the WRDPG defined above, the parameter matrix $P = XX^\top \in \Theta^{n \times n}$ ⊂
CHAPTER 2. RANDOM GRAPH MODELS

$\mathbb{R}^{n \times n}$ is automatically symmetric because the link function is the 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 definiteness 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 follows: 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 $\tau$ 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 $B$ is assumed be positive semidefinite. Henceforth, with slight abuse of terminology, we will denote the sub-model of WSBM with positive semi-definite $B$ as simply 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

Estimation from Multiple Graphs

and 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. And many literatures focus on the low-dimensional representation, for example, Allard et al. [2012] constructs data-dependent multi-scale dictionaries that aim at efficient encoding and manipulating of the data.

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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 consider the SWU4 brain graphs available through
the Consortium for Reliability and Reproducibility (CoRR) [Zuo et al., 2014] after
processing with ndmg pipeline [Kiar et al., 2016, 2017] based on Desikan atlas. 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 3.6.1. 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
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 SWU4 dataset
example with Desikan atlases mentioned above, then we calculate the entry-wise sam-
ple 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 some node pairs being estimated very inac-
curately. Specifically, 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 bottom 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

Figure 3.1: Heat maps of the population mean $P$, the sample mean $\bar{A}$, and the estimator $\hat{P}$ 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 $\hat{P}$ for the same 5 sampled graphs with dimension $d = 11$ selected using the Zhu and Ghodsi method (right). Details about how to construct $\hat{P}$ 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 $\hat{P}$ (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 $\hat{P}$, we highlight the edges (18 edges highlighted for $\bar{A}$ and 6 for $\hat{P}$) 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, $\hat{P}$ is calculated with dimension $d = 12$. A clearly smoothing effect can be seen in the heat map of $\hat{P}$ (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 $\hat{P}$.
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 roughly a 2-block model at the highest level, representing the two hemispheres. More generally, the latent positions model (LPM) [Hoff et al., 2002] 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) [Nickel 2008, Young and Scheinerman, 2007] 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 struc-
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

ture 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 $\hat{P}$ in the top right panel. Note that compared to the sample mean $\bar{A}$ (with mean squared error equals 0.016), $\hat{P}$ (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 $\hat{P}$, 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 Figure 3.1, we plot the heat map of $\hat{P}$ based on sample size $m = 1$. From the figure, we can see that $\hat{P}$ 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, $\hat{P}$ (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 $\hat{P}$.

In this chapter, we show via theory, simulations, and real data analysis that the
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 stochastic 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.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 $E[A^{(1)}]$, where since the graphs are identically distributed $E[A^{(t)}] = E[A^{(t')}]$ 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 $\tau_i$ are firstly drawn iid from a categorical distribution with block membership probabilities 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 $A^{(1)}, \ldots, A^{(m)}$ are sampled independently from the SBM conditioned on $\tau$.

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.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

Definition 3.2.1 (Adjacency Spectral Embedding) For a symmetric $n$-by-$n$ matrix $A$, let its eigen-decomposition be $\hat{U} \hat{S} \hat{U}^\top + \tilde{U} \tilde{S} \tilde{U}^\top$, where $\hat{S}$ is a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest $d$ eigenvalues of $A$, and $\hat{U}$ has columns given by the corresponding eigenvectors. Similarly, $\tilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n - d$ eigenvalues of $A$, and $\tilde{U}$ has the columns given by the corresponding eigenvectors. Then the $d$-dimensional adjacency spectral embedding (ASE) of $A$ is defined as $\hat{X} = \hat{U} \hat{S}^{1/2} \in \mathbb{R}^{n \times d}$.

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 appropriate 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. In particular, a method proposed in Zhu and Ghodsi [2006] finds the gaps in the scree-plot by positing a Gaussian mixture model for the ordered eigenvalues. This method provides multiple choices based on different elbows. In this
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

paper, to avoid under-estimating the dimension, which is often much more harmful than over-estimating it, we choose the third elbow returned by the procedure of [Zhu and Ghodsi, 2006].

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
\( \tilde{A} \), and a low-rank estimator \( \hat{P} \). 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 $\text{Var}(\bar{A}_{ij}) = P_{ij}(1 - P_{ij})/m$. Moreover, $\bar{A}$ is the uniformly minimum-variance unbiased estimator, so it has the smallest variance among all unbiased 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, $\bar{A}$ does not have any asymptotic properties as the number of vertices $n$ becomes large for fixed $m$, while our estimator $\hat{P}$ discussed in Section 3.3.2 does.

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 $\hat{P}$

Motivated by the low-rank structure of the RDPG mean matrix, we propose the estimator $\hat{P}$ 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 $\hat{P}$ 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 $\hat{P}$

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

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

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

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

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

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

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

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

7: Set $\hat{P}$ to $\tilde{P}^{(1)}$ with values $< 0$ set to 0 and values $> 1$ set to 1.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 $\hat{U}\hat{S}\hat{U}^\top + \tilde{U}\tilde{S}\tilde{U}^\top$, where $\hat{S}$ is a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest $d$ eigenvalues of $\bar{A}$, and $\hat{U}$ has columns given by the corresponding eigenvectors. Similarly, $\tilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n - d$ eigenvalues of $\bar{A}$, and $\tilde{U}$ has the columns given by the corresponding eigenvectors. The $d$-dimensional adjacency spectral embedding (ASE) of $\bar{A}$ is given by $\hat{X} = \hat{U}\hat{S}^{1/2} \in \mathbb{R}^{n \times d}$. For an RDPG, the rows of $\hat{X}$ 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 $\hat{X}\hat{X}^\top = \hat{U}\hat{S}\hat{U}^\top$.

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

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

4: Set $\text{lowrank}_d(A)$ to $\hat{U}\hat{S}\hat{U}^\top$;
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

To compute our estimator \( \hat{P} \), 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 \( \hat{P} \) 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 \( \hat{P} \). 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 \( \hat{P} \) 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 \( \hat{P} \) 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 $\hat{P}$ motivated by RDPG. We evaluate our estimators in terms of mean squared error, either $\text{MSE}(\hat{P}_{ij}) = \mathbb{E}[(\hat{P}_{ij} - P_{ij})^2]$ 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.

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

$$\text{RE}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{MSE}(\hat{\theta}_2)}{\text{MSE}(\hat{\theta}_1)}.$$ 

In our case, this is $\text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\text{MSE}(\hat{P}_{ij})}{\text{MSE}(\bar{A}_{ij})}$, with values above 1 indicating $\bar{A}$ should be preferred while values below 1 indicate $\hat{P}$ 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 $\hat{P}$, shown in the lower left panel (see Section 3.3.2). By visual inspection, it is clear that the low-rank estimate $\hat{P}$ more closely approximates the probability matrix $P$ as compared to $\bar{A}$.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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, \( n \cdot \text{RE}(\hat{A}_{ij}, \hat{P}_{ij}) \) 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 \( \tau_i \) 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 \( \rho_k \) 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 \( A^{(1)}, \ldots, A^{(m)} \) sampled independently from the SBM conditioned on \( \tau \), we can calculate the two estimators \( \bar{A} \) and \( \hat{P} \).

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

\[
\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] \approx \frac{1}{\rho_{\tau_i}} + \frac{1}{\rho_{\tau_j}} - \frac{mn P_{ij}(1 - P_{ij})}{mn}.
\]
And
\[
\lim_{n \to \infty} n \cdot \text{Var}(\hat{P}_{ij}) = \frac{1}{m} \rho_{\tau_i} + \frac{1}{m} \rho_{\tau_j} P_{ij}(1 - P_{ij}).
\]

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 \( \hat{P} \).

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 \( \hat{P} \) 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 \( \tilde{P}_{ij} = \tilde{X}_i \top \tilde{X}_j \) is a noisy version of the dot product of \( \nu_s \top \nu_t \) and each \( \tilde{X}_i \) 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 $\hat{P}_{ij}$ converges to $(1/\rho_{\tau_i} + 1/\rho_{\tau_j}) P_{ij}(1 - P_{ij})/(n \cdot m)$ as $n \to \infty$. Since the variance of $\bar{A}_{ij}$ is $P_{ij}(1 - P_{ij})/m$, the relative efficiency between $\hat{P}_{ij}$ 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 $\hat{P}_{ij}$ 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, $\hat{P}$ 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 $\hat{P}_{ij}$ 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 $\text{rank}(B) = K = d$, then for large enough $n$, we have

$$\text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{n}. \quad (3.1)$$
\textit{And the asymptotic relative efficiency (ARE) is}

\[
\text{ARE}(\bar{A}_{ij}, \hat{P}_{ij}) = \lim_{n \to \infty} \text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = 0.
\]

\textbf{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 $n$,

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

we have for large enough $n$,

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

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

This theorem indicates that under the SBM, $\hat{P}$ is a much better estimate of the mean of the collection of graphs $P$ than $A$. Note that a relative efficiency less than 1 indicates that $\hat{P}$ should be preferred over $\bar{A}$, so under the above assumptions, as $n \to \infty$, $\hat{P}$ performs far better than $\bar{A}$. From the result, we see that the relative efficiency is of order $O(n^{-1})$ and $n \cdot \text{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j}$ when
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

$n \to \infty$. 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 $\hat{P}$ 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 of two vertices from different blocks 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 $\rho_s$ below $1/n$ correspond to graphs where typically no vertices are in that block, so the effective minimum we can consider for $\rho_s$ is $1/n$.)

To illustrate Equation 3.1 of Theorem 3.4.3 we consider a 2-block SBM with
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

Figure 3.3: Asymptotic scaled relative efficiency $n \cdot \text{RE}(\bar{A}, \hat{P})$ in a 2-block SBM. For each distinct pair of edge probabilities in a 2-block SBM specified in Equation 3.2, the scaled relative efficiency only depends on the proportion of vertices in each block. We show the scaled asymptotic relative efficiency as $\rho_1$ 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 $\rho_1$ is close to 1, while the improvements are smaller for block pairs with relatively few vertex pairs such as $B_{11}$ when $\rho_1$ is small and $B_{12}$ when $\rho_1$ 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, $\hat{P}$ performs best for large blocks while the improvements may be very minor for blocks with only a few vertices.
parameters

\[ B = \begin{bmatrix} 0.42 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \quad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \quad (3.2) \]

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 \(\hat{P}\), 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 blocks, respectively, in the 2-block SBM we specified earlier. We vary \(\rho_1\) 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 \(\hat{P}\) versus \(\bar{A}\) in the general RDPG case, using the same ideas as in Theorem 3.4.3, we can show that \(\text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = O(1/n)\).

**Proposition 3.4.4** Suppose that \(A^{(1)}, A^{(2)}, \ldots, A^{(m)}\) are independently and identi-
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

cally 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

\[
\text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = O(1/n).
\]

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, \(\text{RE}(\bar{A}, \hat{P})\) will increase as \(m\) increases for full-rank models. Indeed, for large \(m\) we could have \(\text{RE}(\bar{A}, \hat{P}) \geq 1\) since we cannot guarantee that \(\hat{P}\) 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 SWU4 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 $\hat{P}$ 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:

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

Similarly, when calculating $\hat{P}$, 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

$N \in \{30, 50, 100, 250, 500, 1000\}$ and a fixed number of graphs $m = 100$. The relative efficiency $\text{RE}(\hat{A}_{ij}, \hat{P}_{ij})$ 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

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

for $s, t \in \{1, 2\}$, where $\text{MSE}$ 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), $n \cdot \hat{\text{RE}}_{st}(\hat{A}, \hat{P})$. 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 $\hat{P}$ is performing better than $\hat{A}$.

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 $n \cdot \hat{\text{RE}}_{st}(\hat{A}, \hat{P})$ converges to scaled asymptotic
Figure 3.4: Finite sample relative efficiency based on simulations. The top panel shows the estimated relative efficiency \( \hat{\text{RE}}(\tilde{A}, \tilde{P}) \) 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 \( \tilde{P} \) is performing better than \( \tilde{A} \). 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, \( n \cdot \hat{\text{RE}}(\tilde{A}, \tilde{P}) \). 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.
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.

### 3.6 SWU4 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 $\hat{P}$ is still useful in such cases, we tested its performance on structural connectomic data. The graphs are based on diffusion tensor MR images of SWU4 dataset collected and available at the Consortium for Reliability and Reproducibility (CoRR) [Zuo et al., 2014] (see Section 3.6.1). The dataset contains
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

454 different brain scans, each of which was processed to yield an undirected, unweighted graph with no self-loops, using the pipeline described in [Kiar et al., 2016, 2017]. 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 [Oishi et al., 2010], the Desikan atlas with 70 vertices [Desikan et al., 2006], and the CPAC200 atlas with 200 vertices [Sikka et al., 2014]. 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.

\section*{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 $\hat{P}$ 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 \ref{chap3:section3.7}), 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 $\text{MSE}(\bar{A}) = \binom{n}{2}^{-1} \sum_{i<j} E[\bar{A}_{ij} - P_{ij}]$ and similarly for $\text{MSE}(\hat{P})$, 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 \( \hat{P} \) 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 relatively easy to compute. We plot the estimated relative efficiencies between \( \bar{A} \) and \( \hat{P} \) 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 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 \( \hat{P} \) 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 \( \hat{P} \). Nonetheless, when applied to subgroups inference, such as all females between the age of 21 and 25, \( \hat{P} \) 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, \( \hat{P} \) offers certain advantages, especially since low-rank
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

Figure 3.5: Relative efficiencies of $\bar{A}$ versus $\hat{P}$ for the SWU4 dataset. For each atlas, JHU, Desikan, and CPAC 200, we sampled graphs which we used to compute $\bar{A}$ and $\hat{P}$. We compared different sample sizes $m$ and different dimension selection procedures, ZG and USVT. For each of the two methods for computing $\hat{P}$, 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 $n$. 
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 \( \hat{P} \), using Zhu and Ghodsi’s 3rd elbow to select \( d = 11 \). In Figure 3.1, the estimates \( \bar{A} \) and \( \hat{P} \) 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, \( \hat{P} \) 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 \( \hat{P} \) (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 \( \hat{P} \), we highlight the edges (18 edges highlighted for \( \bar{A} \) and 6 for \( \hat{P} \)) 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 \( |\hat{P} - P| \). In addition, we include the absolute difference \( |\bar{A} - \hat{P}| \) 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 \( \hat{P} \) being highlighted in the figure,
Figure 3.6: Heat plots of absolute estimation error for $\bar{A}$ and $\hat{P}$ (lower triangle) and absolute errors above 0.4 (upper triangle). These heat plots show the absolute estimation error $|\bar{A} - P|$, $|\hat{P} - P|$ and $|\bar{A} - \hat{P}|$ for a sample of size $m = 5$ from the Desikan dataset. The embedding dimension for $\hat{P}$ 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 $\hat{P}$ are highlighted indicates that $\hat{P}$ has fewer large outliers compared to $\bar{A}$.

Further indicating the superior performance of $\hat{P}$. 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, $\hat{P}$ 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 $|\bar{A}_{ij} - P_{ij}| - |\hat{P}_{ij} - P_{ij}|$ according to the location of the corresponding regions in the brain. Red edges indicate that $\hat{P}$ overestimates $P$, while blue means that $\hat{P}$ underestimates $P$. The edge width is determined by the estimation error for $\hat{P}$, 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 $\sum_{j}(|\bar{A}_{ij} - P_{ij}| - |\hat{P}_{ij} - P_{ij}|)$. Notably, three of these top
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 $|A_{ij} - P_{ij}| - |\hat{P}_{ij} - P_{ij}|$. Red edges indicate that $\hat{P}$ overestimates $P$ while blue means that $\hat{P}$ 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 $\hat{P}$ outperforms $\bar{A}$ the most for estimating $P$.

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.

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

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

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 $\hat{\text{MSE}}$. Meanwhile, $\hat{P}$ 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, $\hat{P}$ smooths the connectomes estimate and improves the performance. Additionally, we see that there is a large range of dimensions where the performance for $\hat{P}$ 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 $\hat{P}$.

### 3.6.4 Interpretability of Low-rank Methods

Low-rank methods can often be more easily interpreted. 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

![Graphs showing MSE comparison](image)

**Figure 3.8:** Comparison of $\hat{\text{MSE}}$ of $\hat{P}$ and $\bar{A}$ for three atlases at three sample sizes for the SWU4 data. These plots show the MSE for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for three datasets (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.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

are given by the inner products of these vectors, one can analyze and visualize the
gometry of these vectors in order to interpret how each vertex is behaving in the
context of the larger graph. Now we take the SWU4 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 $\hat{X} \in \mathbb{R}^{n \times d}$, 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 separately for each of the first 4 dimensions of $\hat{X}$ as in
Figure 3.9 respectively. The color of the $i$-th brain region for the $j$-th dimension is
determined by the value of $\hat{X}_{ij}$, i.e. the $j$-th element of the estimated latent vector
for the $i$-th region. Red represents a positive value while blue represents the negative
one. And the darker the color is, the smaller the magnitude of the value is. The
1st dimension depicted in Panel (a) of the figure provides an average level of the
entire brain. Then in Panel (b) particularly, we can see a clear distinction of the left
and right hemisphere as conveyed in the 2nd dimension. For the other dimensions,
it seems that they correspond to different lobes. For example, red color represents
Frontal and Temporal in the 3rd dimension, while the light blue matches Occipital
roughly in the 4th dimension. We will explore this connection between dimensions
and lobes further in Section 3.6.6. Additionally, such a representation allows the use
of techniques from multivariate analysis to further study the estimated population
mean.
Figure 3.9: Brain plots colored separately for each of the first 4 dimensions of $\hat{X}$ under the Desikan atlas. We color the brain separately for each of the first 5 dimensions of $\hat{X}$. The color of the $i$-th brain region for the $j$-th dimension is determined by the value of $\hat{X}_{ij}$, i.e. the $j$-th element of the estimated latent vector for the $i$-th region. Red represents a positive value while blue represents the negative one. And the darker the color is, the smaller the magnitude of the value is. From the figure, we can see the embeddings have their own neuro-meaning. The first dimension depicted in Panel (a) of the figure provides an average level of the entire brain. Then in Panel (b) particularly, we can see a clear distinction of the left and right hemisphere as conveyed in the second dimension. For the other dimensions, it seems that they correspond to different lobes. And we will explore it further later in Section 3.6.6.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

3.6.5 Challenges of the SWU4 Dataset

While our estimator \( \hat{P} \) performs well when the sample size \( m \) is small and the number of vertices \( n \) is large, the SWU4 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 \( \frac{\| \text{lowrank}_d(P) - P \|_F}{\| P \|_F} \) 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 \( \hat{P}_d \) 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 where the ranks are selected by Zhu and Ghodsi’s method. In particular, the selected dimensions are 13 for JHU, 8 for Desikan, and 37 for CPAC200. As can be expected, the SWU4 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 many negative eigenvalues, which indicates that the positive semi-definiteness is also violated in this real data experiment. This makes it even harder for \( \hat{P} \) 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 \( \hat{P} \) will not capture.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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. Many negative eigenvalues indicate that $P$ is not positive semi-definite.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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 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 $\hat{P}$ 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 $\hat{P}$ will give a less accurate estimate because of the smoothing effect. So the SWU4 dataset we consider is highly preferable to $\bar{A}$ compared to $\hat{P}$. However, even in this situation, $\hat{P}$ 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, $\hat{P}$ still gives a better estimate than $\bar{A}$ for the SWU4 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 these low-rank methods. In this section, we focus on the lobe structure in particular.

Classically, the brain can be divided into lobes, originally based purely on anatomical considerations, now widely recognized to also play a functional role. While different anatomists partition brain regions differently, there is generally agreement on four cortical lobes per hemisphere: frontal, parietal, occipital, and temporal [Desikan et al., 2006, Fischl, 2012, Salat et al., 2004]. 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.

In general one might expect that properties of regions within a lobe are more similar to those across lobes, even upon conditioning an anatomical proximity. 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.

However, the anatomical geometry might contribute to the dependence between \(X\) and \(l\). So a small test statistic \(T(X, l)\) is evidence that the low-rank methods preserves the lobe structure only after we conditioning on anatomy geometry. We consider the hypothesis test as follows:

- \(H_0\): \(X\) 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, i.e. the permutations in the test is performed without any anatomical geometry constraints. 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 also maintains the constraint that regions are geometrically adjacent, 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. $S_{i_1j_1} = 1$ 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. $S_{i_2j_2} = 1$ 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 latent positions in the same lobe are more similar to each other, even after accounting for the fact that geometrically proximal regions may also have similar latent positions.
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.
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

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, $\hat{P}$ often performs well even when this assumption is false. To further illuminate this point, we performed a synthetic 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 $m = 1, 5, \text{ and } 10$ and used $\bar{A}$ and $\hat{P}$, where we varied the rank of $\hat{P}$ from 1 to 70.

Figure 3.14 shows the resulting estimated MSE for $\bar{A}$ (solid line) and $\hat{P}$ (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 $\hat{P}$ performs even better than in the real data experiments. When $m$ is small, $\hat{P}$ 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 $\hat{P}$ 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

Figure 3.14: Comparison of $\hat{P}$ and $\hat{A}$ for synthetic data analysis. As in Figure 3.4, this figure shows MSE for $\hat{A}$ (solid line) and $\hat{P}$ (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 $\hat{RE}(\hat{A}, \hat{P})$ is lower for this synthetic data analysis than for the SWU4 data.

more favorable to $\hat{P}$, 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 $\hat{P}$ in real data can at least partially be attributed to the fact that the independent edge assumption does not hold for real data. And it also suggests that more elaborate models of connectomes will be valuable for various inferential tasks.
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 $\tau_i$ 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 $B = \nu \nu^T \in [0, 1]^{K \times K}$ 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 $A^{(1)}, \ldots, A^{(m)}$ sampled independently from the SBM conditioned on $\tau$. Define $\bar{A} = \frac{1}{m} \sum_{t=1}^{m} A^{(t)}$. Let $\hat{U} \hat{S} \hat{U}^T$ be the best rank-$d$ positive semidefinite approximation of $\bar{A}$, then we define $\hat{P} = \hat{X} \hat{X}^T$, where $\hat{X} = \hat{U} \hat{S}^{1/2}$.

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 $X = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{N \times d}$ have row $i$ equal to $X_i = \nu_{\tau_i}$ (recall that $\tau_i$ are drawn from $[K]$ according to the probabilities $\rho$). 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
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

\( \tau_j = t, \)

\[ P \left\{ \sqrt{n}(W \hat{X}_i - \nu_s) \leq z, \sqrt{n}(W \hat{X}_j - \nu_t) \leq z' \right\} = \Phi(z, \Sigma(\nu_s)/m) \Phi(z', \Sigma(\nu_t)/m) + o(1) \]

\[(3.3)\]

where \( \Sigma(x) = \Delta^{-1} E[X_j X_j^\top (x^\top X_j - (x^\top X_j)^2)] \Delta^{-1} \) and \( \Delta = E[X_1 X_1^\top] \) is the second moment matrix, with all expectations taken unconditionally. The function \( \Phi \) 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 \((B, \rho)\), 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 \( R = \)
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

diag(ρ₁,⋯,ρₖ), we have

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

Thus

\[
\nu_s^\top \Sigma(\nu_t) \nu_s = \nu_s^\top \Delta^{-1} \sum_{k=1}^{K} \rho_k \nu_k \nu_k^\top (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \Delta^{-1} \nu_s \\
= \sum_{k=1}^{K} \rho_k (\nu_s^\top \Delta^{-1} \nu_k)(\nu_k^\top \Delta^{-1} \nu_s)(\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
= \sum_{k=1}^{K} \rho_k (\nu_t^\top U^\top R^{-1} U u_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
= \sum_{k=1}^{K} \rho_k (e_s^\top R^{-1} e_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
= \sum_{k=1}^{K} \rho_k \delta_{sk} \rho_s^{-2} (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\
= \frac{1}{\rho_s} \nu_t^\top \nu_s (1 - \nu_t^\top \nu_s)
\]

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

\[
\lim_{n \to \infty} n \cdot \text{Var}(\hat{P}_{ij}) = \frac{1}{\rho_{\tau_i}} + \frac{1}{\rho_{\tau_j}} \frac{m}{p_{ij}(1 - p_{ij})}.
\]
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

And for \( n \) large enough, conditioning on \( X_i = \nu_s \) and \( X_j = \nu_t \), we have

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

**Proof:** Conditioned on \( X_i = \nu_k \), we have by Theorem 3.8.1,

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

and

\[
n \cdot \text{Cov}(W\hat{X}_i, W_n\hat{X}_i) = \Sigma(\nu_k)/m.
\]

Also, Corollary 3 in Athreya et al. [2016] says \( \hat{X}_i \) and \( \hat{X}_j \) are asymptotically independent. Thus, conditioning on \( X_i = \nu_s \) and \( X_j = \nu_t \), we have

\[
\text{lim}_{n \to \infty} \mathbb{E}[\hat{X}_i^\top \hat{X}_j] = \text{lim}_{n \to \infty} \mathbb{E}[(W_n\hat{X}_i)^\top][W_n\hat{X}_j] = \nu_s^\top \nu_t = P_{ij}.
\]

Since \( \hat{P}_{ij} = \hat{X}_i^\top \hat{X}_j \) 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

\[
\mathbb{E}[\hat{X}_i^\top \hat{X}_j] = \mathbb{E}[(W_n\hat{X}_i)^\top][W_n\hat{X}_j] = \nu_s^\top \nu_t + o(1) = P_{ij} + o(1)
\]
CHAPTER 3. ESTIMATION FROM MULTIPLE GRAPHS AND A LAW OF LARGE GRAPHS

and

\[ n \cdot \text{Var}(\hat{P}_{ij}) \]

\[ = \frac{1}{m} (\nu_s^\top \Sigma(\nu_s)\nu_s + \nu_t^\top \Sigma(\nu_t)\nu_t^\top) + \frac{1}{m^2 n} (tr(\Sigma(\nu_s)\Sigma(\nu_t))) + o(1) \]

\[ = \frac{1}{m} (\nu_s^\top \Sigma(\nu_s)\nu_s + \nu_t^\top \Sigma(\nu_s)\nu_t^\top) + o(1) \]

\[ = \frac{1}{\rho_s + 1/\rho_t} P_{ij} (1 - P_{ij}) + o(1). \]

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

\[ \mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] = \text{Var}(\hat{P}_{ij}) \approx \frac{1/\rho_s + 1/\rho_t}{mn} P_{ij} (1 - P_{ij}) + o(1). \]
Chapter 4

Robust Estimation from Multiple 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 \( \hat{\theta} \) for the unknown graph parameter \( \theta \) given a sample of graphs \( \{G^{(1)}, \ldots, G^{(m)}\} \) 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
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

to estimating $\theta$ 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 dataset “Templeton114” made available through the Neurodata repository[^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 to be

[^1]: http://m2g.io/
(approximately) low-rank, then by biasing towards low-rank structure, more elaborate estimators can have greatly reduced variance and win the bias-variance trade-off, 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; Qin and Priebe, 2013] 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 trade-off 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 as adjacency matrices \( \{A^{(t)}\}(t = 1, \ldots, m) \). 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, completely accurate data is difficult to collect – there will almost 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], is considered in this work.

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

However, since all the models we consider are subsets of the
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

WIEM, it is more natural to consider the contamination 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 \(f \in F\) and \(h \in H\), where both \(F\) and \(H\) 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 \(F\). 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 \(\tau\) is the block assignments, the contamination distribution \(f_{C_{ij}}\) with \(C_{ij} = B'_{\tau'_i, \tau'_j}\) also has the block structure, where \(B'\) is the block probability matrix and \(\tau'\) is the block assignment vector. For clarity, we will consider the sampling procedure when the contamination has the same block structure, i.e. \(\tau = \tau'\). However, this simplification is not required in our theory.

To generate \(m\) graphs under this contamination model with known vertex correspondence, we first sample \(\tau\) from the categorical distribution with parameter \(\rho\) and keep \(\tau\) fixed for all \(m\) graphs as in Section 2.3.3. Then \(m\) symmetric and hollow graphs \(G^{(1)}, \ldots, G^{(m)}\) are sampled such that conditioning on \(\tau\), 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 each \(1 \leq t \leq m\), \(1 \leq i < j \leq n\), where \(P_{ij} = B_{\tau_i, \tau_j}\) and \(C_{ij} = B'_{\tau'_i, \tau'_j}\). Here \(\epsilon\) 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 contamination.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

4.2 Estimators

Under any model introduced in Section 3.1, our goal is 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 such as $\epsilon$ and $C$, our goal is still to estimate the uncontaminated parameter matrix $P$. In this section, we present four estimators as depicted in Figure 4.1, i.e. the standard entry-wise MLE $\hat{P}^{(1)}$, the low-rank approximation of the entry-wise MLE $\tilde{P}^{(1)}$, the entry-wise robust estimator ML$qE$ $\hat{P}^{(q)}$, and the low-rank approximation of the entry-wise ML$qE$ $\tilde{P}^{(q)}$. Since the observed graphs are symmetric and hollow with a symmetric parameter matrix of the model, we are not concerned with estimating the diagonal of $P$; however, the estimate itself should be at least symmetric.

4.2.1 Entry-wise Maximum Likelihood Estimator

$\hat{P}^{(1)}$

Under WIEM, the most natural estimator is the MLE, which happens to be the element-wise MLE $\hat{P}^{(1)}$ in this case. Note that this is $\bar{A}$ in Chapter 3. Moreover, when $\mathcal{F}$ is a one-parameter exponential family, such as Bernoulli, Poisson, or Exponential, the entry-wise MLE $\hat{P}^{(1)}$ is the uniformly minimum-variance unbiased estimator, i.e. it has the smallest variance among all unbiased estimators. In addition, it has desirable asymptotic properties as the number of graphs $m$ goes to infinity. However, in high
Figure 4.1: Roadmap among the data and four estimators.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

dimensional situations such as our graph setting, 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 improve as 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 WRDGP or a WSBM, then the entry-wise MLE is no longer the MLE and the performance can be improved by considering low-rank estimators.

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

Motivated by the low-rank structure of the parameter matrix $P$ in WRDGP, we consider the estimator $\tilde{P}^{(1)}$ proposed by Tang et al. [2016] based on the spectral decomposition of $\hat{P}^{(1)}$, i.e. $\tilde{P}$ in Chapter 3. Dimension selection technique discussed in Section 3.2.2 and the diagonal augmentation procedure introduced in Section 3.2.3 will also be used in this section. The construction procedure of $\tilde{P}^{(1)}$ 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 $\tilde{P}^{(1)} = \text{lowrank}_d(\hat{P}^{(1)})$ as the best rank-$d$ positive semi-definite approximation of $\hat{P}^{(1)}$. To find this best approximation, first calculate
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

the eigen-decomposition of the symmetric matrix \( \hat{P}^{(1)} = \hat{U}\hat{S}\hat{U}^\top + \hat{U}\tilde{S}\hat{U}^\top \), where \( \hat{S} \) is the diagonal matrix with the largest \( d \) eigenvalues of \( \hat{P}^{(1)} \), and \( \hat{U} \) has the corresponding eigenvectors as each column. Similarly, \( \tilde{S} \) is the diagonal matrix with non-increasing entries along the diagonal corresponding to the remaining \( n - d \) eigenvalues of \( \hat{P}^{(1)} \), and \( \tilde{U} \) has the columns given by the corresponding eigenvectors. The \( d \)-dimensional adjacency spectral embedding (ASE) of \( \hat{P}^{(1)} \) is given by \( \hat{X} = \hat{U}\tilde{S}^{1/2} \in \mathbb{R}^{n \times d} \), which follows Definition 3.2.1. The best rank-\( d \) positive semi-definite approximation \( \hat{P}^{(1)} \) is then \( \tilde{P}^{(1)} = \hat{X}\hat{X}^\top = \hat{U}\tilde{S}\hat{U}^\top \). The issue of how to select the embedding dimension \( d \) is discussed in Section 3.2.2. Based on the ASE result, we have the best rank-\( d \) positive semi-definite approximation of \( \hat{P}^{(1)} \) to be \( \tilde{P}^{(1)} = \hat{X}\hat{X}^\top = \hat{U}\tilde{S}\hat{U}^\top \). In the RDPG setting, Sussman et al. [2014] proved that each row of \( \hat{X} \) 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. In the RDPG setting, Sussman et al. [2014] proved that, provided that \( d \) is chosen appropriately, each row of \( \hat{X} \) can accurately estimate the the latent position for each vertex up to an orthogonal transformation. We will extend these results for \( \hat{X} \) to the WRDPG setting in Section 4.3.

Here, we restate the algorithm given 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 $\hat{S}$ to the $d \times d$ diagonal matrix $\text{diag}(s_1, \ldots, s_d)$

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

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

With the dimension selection procedure as described above, the detailed description for calculating our estimator $\hat{P}^{(1)}$ is then given by Algorithm 4.

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

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

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

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

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

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

4: Set $\tilde{P}^{(1)}$ with each entry $\tilde{P}^{(1)}_{ij} = \max(Q_{ij}, 0)$
4.2.3 Entry-wise Maximum $L_q$-likelihood Estimator $\hat{P}^{(q)}$

In the case of no contamination, 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 can outperform the MLE in terms of mean squared error by winning the bias-variance trade-off. Moreover, under contamination models, robust estimators can even outperform the MLE asymptotically since they are designed to be not unduly affected by outliers. We consider one such robust estimator, the maximum $L_q$-likelihood estimator (ML$q$E) proposed by Ferrari and Yang [2010].

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

$$\hat{\theta}_{MLqE} = \arg \max_{\theta \in \Theta} \sum_{i=1}^{m} L_q[f_{\theta}(X_i)],$$

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

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

$$U_{\theta}(x) = \nabla_{\theta} \log f_{\theta}(x)$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

and

\[ U^*_\theta(x; q) = U_\theta(x)f_\theta(x)^{1-q}. \]

Then the MLqE \( \hat{\theta}_{\text{MLqE}} \) can also be seen as a solution to the equation

\[ \sum_{i=1}^{m} U^*_\theta(X_i; q) = 0. \]

This form interprets \( \hat{\theta}_{\text{MLqE}} \) as a solution to a 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 are induced by \( q = 1 \) and lead to the standard MLE.

Under the WIEM, we can calculate the robust entry-wise MLqE \( \hat{P}^{(q)} \) based on the adjacency matrices \( A^{(1)}, \ldots, A^{(m)} \). Note that \( \hat{P}^{(1)} \), the entry-wise MLE, is a special case of entry-wise MLqE \( \hat{P}^{(q)} \) when \( q = 1 \). There is also a bias-variance trade-off 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 automatic selection of \( q \).
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

Intuitively, the low-rank structure of the parameter matrix $P$ in WRDPG should be preserved approximately in the entry-wise ML$q$E $\hat{P}^{(q)}$. 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 $\tilde{P}^{(1)}$, i.e. enforce a low-rank approximation on the entry-wise ML$q$E matrix $\hat{P}^{(q)}$ to get $\tilde{P}^{(q)}$. As in Algorithm 4, we apply the same dimension selection method and diagonal augmentation procedure. The only change is to substitute $\hat{P}^{(1)}$ by $\hat{P}^{(q)}$. The details of the algorithm are shown in Algorithm 5.

Intuitively, the low-rank structure of the parameter matrix $P$ in WRDPG should be preserved approximately in the entry-wise ML$q$E $\hat{P}^{(q)}$. 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 $\tilde{P}^{(1)}$, i.e. enforce a low-rank approximation on the entry-wise ML$q$E matrix $\hat{P}^{(q)}$ to get $\tilde{P}^{(q)}$. The details of the construction of $\tilde{P}^{(q)}$ are given in Algorithm 5. Algorithm 5 is almost identical to that of Algorithm 4 for constructing $\tilde{P}^{(1)}$, the main change being the use of $\hat{P}^{(q)}$ in place of $\hat{P}^{(1)}$.
Algorithm 5 Algorithm to compute $\tilde{P}(q)$

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

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

 1: Calculate the entry-wise ML$q$E $\hat{P}(q)$
 2: Select the dimension $d$ based on the eigenvalues of $\hat{P}(q)$; (see Section 3.2.2)
 3: Set $Q$ to lowrank$_d(\hat{P}(q))$; (see Algorithm 3)
 4: Set $\tilde{P}(q)$ with each entry $\tilde{P}_{ij}(q) = \max(Q_{ij}, 0)$

4.3 Theoretical Results

In this section, for illustrative purposes, we present theoretical results for the case in which the contamination model introduced in Section 4.1 is with respect to exponential distributions. That is $\mathcal{F} = \{f_\theta(x) = \exp(-x/\theta)/\theta, \theta \in [0, R] \subset \mathbb{R}\}$, where $R > 0$ is a constant. These results can be extended beyond the exponential under appropriate conditions, which will be discussed in Section 4.4.

For clarity, we restate the model settings discussed in Section 4.1. Consider the WSBM with parameters $B$ and $\rho$. First we sample the block membership $\tau$ from the categorical distribution with parameter $\rho$ and keep it fixed for all $m$ graphs. Conditioned on this $\tau$, the uncontaminated probability matrix $P$ satisfies $P_{ij} = B_{\tau_i, \tau_j}$. In this section, we assume the contamination has the same block membership $\tau$, and so the contamination matrix $C \in \mathbb{R}^{n \times n}$ has the same block structure as $P$. 

92
Denote $\epsilon$ as the probability that an edge is contaminated. Then $m$ symmetric graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on $\tau$, 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 each $1 \leq t \leq m$, $1 \leq i < j \leq n$. Note that our theoretical results do not require the contamination to have the same block structure and block membership $\tau$ as the uncontaminated probability matrix; different block structure will lead to similar results – but require higher embedding dimension – since the rank of $(1 - \epsilon)P_{ij} + \epsilon C_{ij}$ is still finite.

In the setting outlined above, 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 easily leads to the result for the total MSE for the entire matrix. We present the main results in this section. Additional results and proofs of stated results are given in Section 4.7.

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

We first compare the performance between the entry-wise MLE $\hat{P}^{(1)}$ and the entry-wise ML$q$E $\hat{P}^{(q)}$. Without using the graph 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 $0 < q < 1$ and any $P$, there exists a constant $C_0(\epsilon, q) > 0$ depending only on $\epsilon$, $q$, and $\max_{i,j} P_{ij}$ such that under the contaminated model $A_{ij}^{(t)} \sim (1 - \epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ with $C_{ij} > C_0(\epsilon, q)$ for all $i, j$ ML$qE$ has smaller entry-wise asymptotic bias compared to MLE, i.e.

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

for $1 \leq i, j \leq n$ and $i \neq j$. Moreover, for any $\epsilon \geq 0$ and any $q$,

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

And thus

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

Theorem 4.3.1 shows that the entry-wise ML$qE \hat{P}^{(q)}$ has smaller bias for estimating $P$ asymptotically compared to the entry-wise MLE $\hat{P}^{(1)}$. Although we put restrictions on the contamination matrix $C$ in the statement of the theorem, the result still holds provided that $\epsilon(C_{ij} - P_{ij}) > (1 - q)P_{ij}$ for all $i, j$. This condition can be interpreted as only requiring that the contamination of the model is large enough (either large contamination parameter matrix, or higher likelihood of encountering an outlier).

From a different perspective, by putting a condition on $q$ with respect to the amount of contamination, the above condition also corresponds to only requiring that $\hat{P}^{(q)}$ 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 $\epsilon$ or a small enough $q$.

Theorem 4.3.1 also indicates that both estimators have variances converging 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, $\hat{P}^{(q)}$ asymptotically reduces the bias while keeping the variance asymptotically the same compared to $\hat{P}^{(1)}$. Thus in terms of MSE, $\hat{P}^{(q)}$ is a better estimator than $\hat{P}^{(1)}$ when the number of graphs $m$ is large with enough contamination.

4.3.2 $\hat{P}^{(1)}$ vs. $\tilde{P}^{(1)}$

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

Before proceeding to the comparison between the two estimators, we first recall the definition of the asymptotic relative efficiency (ARE) [Serfling 2011], which is an 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 $\theta$ of a distribution $f$, and for estimators $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ such that $E[\hat{\theta}^{(1)}] = E[\hat{\theta}^{(2)}] = \theta'$, $n \cdot \text{Var}(\hat{\theta}^{(1)}) \to V_1(f)$ and $n \cdot \text{Var}(\hat{\theta}^{(2)}) \to V_2(f)$, the Asymptotic Relative Efficiency (ARE) of $\hat{\theta}^{(2)}$ to $\hat{\theta}^{(1)}$ is given by

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

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

To compare $\hat{P}^{(1)}$ and $\tilde{P}^{(1)}$, we first show that they have the same entry-wise asymptotic bias and then use the ARE criterion to compare their performance in the following theorem.

Theorem 4.3.3 Assuming that $m = O(n^b)$ for any $b > 0$, then $\tilde{P}^{(1)}$, the estimator based on the ASE of the entrywise MLE $\hat{P}$, has the same entry-wise asymptotic bias as $\tilde{P}^{(1)}$, i.e.

$$\lim_{n \to \infty} \text{Bias}(\tilde{P}^{(1)}_{ij}) = \lim_{n \to \infty} E[\tilde{P}^{(1)}_{ij}] - P_{ij} = \lim_{n \to \infty} E[\hat{P}^{(1)}_{ij}] - P_{ij} = \lim_{n \to \infty} \text{Bias}(\hat{P}^{(1)}_{ij}).$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

\[
\text{Var}(\tilde{P}^{(1)}_{ij}) = O(m^{-1}n^{-1}(\log n)^3), \text{Var}(\hat{P}^{(1)}_{ij}) = O(m^{-1}).
\]

Thus

\[
\frac{\text{Var}(\tilde{P}^{(1)}_{ij})}{\text{Var}(\hat{P}^{(1)}_{ij})} = O(n^{-1}(\log n)^3); \quad \text{ARE}(\hat{P}^{(1)}_{ij}, \tilde{P}^{(1)}_{ij}) = 0.
\]

for all \(1 \leq i < j \leq n\).

Theorem 4.3.3 says that when \(m\) is fixed or grows no faster than any polynomial with respect to \(n\), the ASE procedure applied to \(\hat{P}^{(1)}\) will not affect the asymptotic bias for estimating \(P\). Combined with the fact that the ratio of the variances of the two estimators is of order \(O(n^{-1}(\log n)^3)\), we have that the ARE of \(\tilde{P}^{(1)}_{ij}\) to that of \(\hat{P}^{(1)}_{ij}\) is 0. Thus \(\tilde{P}^{(1)}_{ij}\) is a better estimate of \(P_{ij}\) than \(\hat{P}^{(1)}_{ij}\) for 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 \(\hat{P}^{(1)}\) helps reduce the variance while keeping the bias unchanged asymptotically, leading to a better estimate \(\tilde{P}^{(1)}\) for \(P\) in terms of MSE.

### 4.3.3 \(\hat{P}^{(q)}\) vs. \(\tilde{P}^{(q)}\)

We now proceed to analyze the effect of the ASE procedure applied to the entry-wise MLqE \(\tilde{P}^{(q)}\) under the gross error contamination model in order to compare the
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

performance between \( \hat{P}^{(q)} \) and \( \tilde{P}^{(q)} \). Similarly, we first show that the two estimators have the same entry-wise asymptotic bias under appropriate conditions, and then use the ARE criterion to compare their 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 ML\( q \)E has the same entry-wise asymptotic bias as ML\( q \)E, i.e.

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

In addition, for \( 1 \leq i < j \leq n \) and \( i \neq j \),

\[
\text{Var}(\tilde{P}^{(q)}_{ij}) = O(n^{-1}(\log n)^3), \quad \text{Var}(\hat{P}^{(q)}_{ij}) = O(m^{-1}).
\]

Thus

\[
\frac{\text{Var}(\tilde{P}^{(q)}_{ij})}{\text{Var}(\hat{P}^{(q)}_{ij})} = O(mn^{-1}(\log n)^3).
\]

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

\[
\text{ARE}(\tilde{P}^{(q)}_{ij}, \hat{P}^{(q)}_{ij}) = 0.
\]

The proof for Theorem 4.3.4 is almost the same as the proof for Theorem 4.3.3. But unlike the results for comparing \( \tilde{P}^{(1)} \), we are missing the term \( m^{-1} \) in the variance bound for \( \text{Var}(\tilde{P}^{(q)}) = O(n^{-1}(\log n)^3) \) for arbitrary \( q \) due to the structure of maximum
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Lq likelihood equation. As a result, while the ASE procedure to obtain \( \widehat{P}^{(q)} \) still does not affect the asymptotic bias compared to \( \widehat{P}^{(q)} \), the ratio of variances for \( \widehat{P}^{(q)} \) and \( \widehat{P}^{(q)} \) has an extra term \( m \) compared to the ratio of variances for \( \widehat{P}^{(1)} \) and \( \widehat{P}^{(1)} \). This leads to some difference in the conclusion of Theorem 4.3.4 compared to that of Theorem 4.3.3. Specifically, when \( m \) is fixed, the order of the ratio of the variances of \( \widehat{P}^{(q)} \) and \( \widehat{P}^{(q)} \) is \( O(n^{-1}(\log n)^3) \), which converges to 0 as \( n \to \infty \). If \( m \) also increases as \( n \) increases, then provided that \( m \) grows on the order of \( o(n(\log n)^{-3}) \), the ARE of \( \widehat{P}^{(q)} \) and \( \widehat{P}^{(q)} \) still converges to 0.

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

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

Finally, we compare the performance between \( \widehat{P}^{(1)} \) and \( \widehat{P}^{(q)} \) by combining the previous results.

**Theorem 4.3.5** For sufficiently large values of \( \{C_{ij}\} \) and any \( 1 \leq i, j \leq n \), if \( m \to \infty \) at order \( m = O(n^b) \) for any \( b > 0 \), then the estimator based on ASE of MLqE has smaller entry-wise asymptotic bias compared to the estimator based on ASE of MLE, i.e.

\[
\lim_{m,n \to \infty} \text{Bias}(\widehat{P}_{ij}^{(1)}) > \lim_{m,n \to \infty} \text{Bias}(\widehat{P}_{ij}^{(q)})
\]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Moreover, if \( m = O(n^b) \) for any \( b > 0 \), then

\[
\lim_{n \to \infty} \text{Var}(\tilde{P}_{ij}^{(1)}) = \lim_{n \to \infty} \text{Var}(\tilde{P}_{ij}^{(q)}) = 0.
\]

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 \( \tilde{P}^{(q)} \) inherits the robustness from the entry-wise MLqE \( \hat{P}^{(q)} \) and has a smaller asymptotic bias compared to \( \tilde{P}^{(1)} \) while both estimates have variance going to 0 as \( n \to \infty \). Thus \( \tilde{P}^{(q)} \) is a better estimator than both \( \tilde{P}^{(1)} \) and \( \hat{P}^{(1)} \). Finally, if \( m = o(n(\log n)^{-3}) \), then by Theorem 4.3.4, \( \tilde{P}^{(q)} \) is also better than \( \hat{P}^{(q)} \) and hence is the best estimator among all four estimators.

4.3.5 Summary of our Four Estimators

We summarize all four estimators and their relationships in Figure 4.2. From top to bottom in 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 contamination, whenever the number of graphs \( m \) is large enough, the bias term which dominates the MSE will be improved. Thus in Figure 4.2 we have quantified the qualitative roadmap introduced in Figure 4.1. We will evaluate these four estimators on simulated and real data experiment in Section 4.5 and Section 4.6.
4.4 Extensions

The results in Section 4.3 are presented in the specific setting of the exponential distribution and the robust MLqE estimator. These results, however, can also be generalized to a broader class of distribution families and/or different entry-wise robust estimators (denoted as $\hat{P}(R)$), provided that the following conditions are satisfied:

1. Letting $A_{ij} \overset{\text{ind}}{\sim} (1 - \epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$, then we require $f_\theta$ to satisfy $E[(A_{ij} - E[\hat{P}^{(1)}_{ij}])^k] \leq \text{const}^k \cdot k!$, where $\hat{P}^{(1)}$ 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_{ij} >
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\[ C_0(P_{ij}, \epsilon), \]

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

3. \( \hat{P}_ij^{(R)} \leq \text{const} \cdot \hat{P}_ij^{(1)}; \)

4. \( \text{Var}(\hat{P}_ij^{(R)}) = O(m^{-1}), \) where \( m \) is the number of graph observations.

Condition 1 is to ensure that the observations \( A_{ij} \) does not deviate too far from their expectation so that Bernstein-like concentration inequalities can be applied. Condition 2 is similar to that used in Section 4.3.1 in particular, it assumes that the contamination of the model is large enough (a restriction on the distribution) and/or \( \hat{P}^{(R)} \) is sufficiently robust with respect to the contamination (a condition on the estimator). By taking advantage of Condition 1 which controls \( \hat{P}^{(1)} \), Condition 3 allows one to derive Bernstein-like concentration inequalities for \( \hat{P}^{(R)} \). Condition 4 ensures that the variance of \( \hat{P}_ij^{(R)} \) is comparable to the variance of the entry-wise MLE \( \hat{P}_ij^{(1)} \), which is of order \( O(m^{-1}) \). (Note that in the absence of Condition 4, similar but weaker results can still be derived.)

As an example to clarify the above four conditions, we sketch the argument that the results of Section 4.3 also holds for the Poisson distribution when the entry-wise MLqE estimator is used. The Poisson distribution is a commonly used distribution for nonnegative graphs with integer weights. Lemma 4.7.34 verifies Condition 1; intuitively, since the exponential distribution has a fatter tail compared to the Poisson, we should have the bound for the central moment of the Poisson directly from the
results for the exponential distribution. Condition 2 is satisfied when we use the MLqE with the Poisson distribution. More specifically, since the gross error contamination is to the right, i.e. \( C_{ij} > P_{ij} \), the weights in the MLqE equation will be smaller when the observed values are larger observations, in contrast to the equal weighting in the MLE equation. Thus under the gross error model, for sufficiently large \( m \), \( \hat{P}_{ij}^{(R)} \) will be less biased than \( \hat{P}_{ij}^{(1)} \). For condition 3, \( \hat{P}_{ij}^{(R)}/\hat{P}_{ij}^{(1)} \) is maximized when there are \( m \) data points \( x_1, \ldots, x_m \) 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 MLqE larger than MLE \( \bar{x} \), we need the weights of the first \( m \) data points to be smaller than the weights of the remaining \( m - k \) points. Thus \( \exp(-\bar{x}) < \bar{x}^m \exp(-\bar{x})/x_m! \). But then \( x_m! < \bar{x}^m \). By the lower bound in Stirling’s formula, we have \( x_m < e \cdot \bar{x} \) when \( x_m > 0 \). Note that if \( x_m = 0 \) then MLE equals MLqE since all data points equal zero. Thus MLqE is bounded by \( e \cdot \bar{x} \). As a result, \( \hat{P}_{ij} \leq e \cdot \hat{P}_{ij}^{(1)} \) and Condition 3 is satisfied. Finally, Condition 4 follows directly from the theory of minimum contrast estimators.

In summary, all theorems in Section 4.3 hold for the Poisson distribution together with the MLqE. The four conditions presented in this section provide a general framework for extending the theory to more general models and robust estimators.
4.5 Simulations

4.5.1 Simulation Setting

Here we consider a 2-block WSBM with respect to the exponential distribution parameterized by

\[
B = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}, \quad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.
\]

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

\[
B' = \begin{bmatrix} 9 & 6 \\ 6 & 13 \end{bmatrix}, \quad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.
\]

With these parameters specified, we sample graphs according to Section 4.1. For ease of presentation, in the simulation we assume the true dimension \( d = \text{rank}(B) = 2 \) is known, and thus we eliminate the dimension selection step in Algorithm 4 and Algorithm 5.

As suggested in [Tang et al., 2016], in this work we combine both ideas by first using Marchette’s row-averaging method and then one step of Scheinerman’s iterative method.
4.5.2 Simulation Results

To see how the performance of the four estimators vary with respect to contamination, we first run 1000 Monte Carlo replicates based on the contaminated WSBM 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 \( \epsilon \) from 0 to 0.4. Given each sample, four estimators can be computed following Algorithm 4 and Algorithm 5. Since we are not focusing on how to select the parameter \( q \) in the MLqE estimator, we shall use a fixed \( q = 0.9 \) unless specified otherwise. The MSE of each estimator can be estimated since the probability matrix \( P \) is known in this simulation.

The results are presented in Figure 4.3. Different curves represent the simulated MSE associated with the four different estimators. Firstly, we see that MLE \( \hat{P}^{(1)} \) is a better estimator compared to MLqE \( \hat{P}^{(q)} \) when there is little or no contamination (i.e. \( \epsilon \leq 0.01 \) in the figure); however this estimator degrades dramatically as the contamination probability increases. On the other hand, the MLqE \( \tilde{P}^{(q)} \) is slightly less efficient than the MLE \( \tilde{P}^{(1)} \) when the contamination probability is small, but is much more robust under a large contamination probability compared to the MLE.

Secondly, we see that even with a relatively small number of vertices \( n = 100 \), the ASE procedure which takes advantage of the low-rank structure already helps improve the performance of \( \hat{P}^{(1)} \) and lets \( \tilde{P}^{(1)} \) win the bias-variance trade-off. Since the MLqE \( \hat{P}^{(q)} \) approximately preserves the low-rank structure of the original graph, the ASE procedure also helps and makes \( \tilde{P}^{(q)} \) a better estimate. Although both
\( \tilde{P}(q) \) and \( \tilde{P}(1) \) take advantage of the low-rank structure and have reduced variances, \( \tilde{P}(q) \) constructed based on \( \text{ML}_{q}E \) inherits the robustness from \( \text{ML}_{q}E \), so when the contamination probability is large enough, \( \tilde{P}(q) \) outperforms \( \tilde{P}(1) \) and degrades more slowly.

Figure 4.4 shows additional simulation results by varying the parameter \( q \) in \( \text{ML}_{q}E \) with fixed \( n = 100, m = 20 \) and \( \epsilon = 0.1 \) based on 1000 Monte Carlo replicates. 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, for a wide range of \( q \), the \( \text{ML}_{q}E \) wins the bias-variance trade-off and exhibits the robustness property compared to the MLE. And as \( q \) goes to 1, \( \text{ML}_{q}E \) goes to the MLE as expected.

### 4.6 Brain Graphs Experiment

We now compare the four estimators on a structural connectomic dataset. 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 ndmg pipeline [Kiar et al., 2016, 2017]. There are different versions of this pipeline. In particular, the estimators are calculated based on graphs generated through version “ndmg-v0.0.1” (ndmg1). Note that we will consider another version “ndmg-v0.0.33” (ndmg2) to assess the performance of
**Figure 4.3:** Mean squared error in average by varying contamination ratio $\epsilon$ with fixed $n = 100$ and $m = 20$ based on 1000 Monte Carlo replicates, using $q = 0.9$ when applying ML$q$E. Different curves represent the simulated MSE associated with four different estimators. 1. MLE $\hat{P}^{(1)}$ vs. ML$q$E $\tilde{P}^{(q)}$ (Relationship (a) in Figure 4.2): MLE outperforms by a small amount when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination probability increases; 2. MLE $\hat{P}^{(1)}$ vs. ASE $\circ$ MLE $\tilde{P}^{(1)}$ (Relationship (b) in Figure 4.2): ASE procedure takes the low rank structure into account and $\tilde{P}^{(1)}$ wins the bias-variance trade-off; 3. ML$q$E $\tilde{P}^{(q)}$ vs. ASE $\circ$ ML$q$E $\tilde{P}^{(q)}$ (Relationship (c) in Figure 4.2): ML$q$E approximately preserves the low rank structure of the original graph, so ASE procedure still helps and $\tilde{P}^{(q)}$ wins the bias-variance trade-off; 4. ASE $\circ$ ML$q$E $\tilde{P}^{(q)}$ vs. ASE $\circ$ MLE $\tilde{P}^{(1)}$ (Relationship (d) in Figure 4.2): When contamination probability is large enough, $\tilde{P}^{(q)}$ based on ML$q$E is better, since it inherits the robustness from ML$q$E.
Figure 4.4: Mean squared error in average by varying the parameter $q$ in ML$q$E with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different curves represent the simulated MSE associated with the four different estimators. 1. MLE $\hat{P}^{(1)}$ vs. ML$q$E $\hat{P}^{(q)}$ (Relationship (a) in Figure 4.2): Within an appropriate range of $q$, $\hat{P}^{(q)}$ wins the bias-variance trade-off and exhibits robustness compared to $\hat{P}^{(1)}$. Also as $q$ goes to 1, $\hat{P}^{(q)}$ goes to $\hat{P}^{(1)}$ as expected; 2. MLE $\hat{P}^{(1)}$ vs. ASE $\tilde{P}^{(1)}$ (Relationship (b) in Figure 4.2): Both estimators are not affected by different choices of $q$. $\tilde{P}^{(1)}$ outperforms $\hat{P}^{(1)}$ as shown in Figure 4.3; 3. ML$q$E $\hat{P}^{(q)}$ vs. ASE $\circ$ ML$q$E $\tilde{P}^{(q)}$ (Relationship (c) in Figure 4.2): ASE procedure takes advantage of the graph structure and $\tilde{P}^{(q)}$ improves the performance of $\hat{P}^{(q)}$ independent of the selection of $q$; 4. ASE $\circ$ ML$q$E $\tilde{P}^{(q)}$ vs. ASE $\circ$ MLE $\tilde{P}^{(1)}$ (Relationship (d) in Figure 4.2): Within an appropriate range of $q$, $\tilde{P}^{(q)}$ inherits robustness from $\hat{P}^{(q)}$ and outperforms $\tilde{P}^{(1)}$. 

$n = 100, m = 20, \epsilon = 0.1$
our estimators later in this section. 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 experiment. The weight of an edge between two vertices represents the number of white-matter tracts connecting the corresponding two regions of the brain.

Generally, we do not expect the graphs to perfectly follow an RDPG model, or even an IEM. Before proceeding with our analysis, we will perform some exploratory data analysis to check whether the data can reasonably be assumed to have approximate low-rank structure. Indeed, without at least some approximately low-rank structure, we will not expect the ASE procedure to improve the bias-variance trade-off 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 ndmg1 pipeline. The eigenvalues first decrease dramatically and then stay around 0 for a large range of dimensions. In addition, we also plot the histogram in the right panel of Figure 4.5. From the figure, we see that many eigenvalues are concentrated around zero. This exploration suggests that the information is mostly contained in the first few dimensions. Such approximate low-rank property provides an opportunity to win the bias-variance trade-off by applying the 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
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

![Screeplot and histogram of normalized eigenvalues](image)

**Figure 4.5:** Screeplot and the histogram of the normalized eigenvalues (divided by the largest eigenvalue 230267.8) of the mean of 114 graphs based on ndmg1 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 normalized eigenvalues (divided by the largest eigenvalue 230267.8) of the mean graph of all 114 graphs with diagonal augmentation. Many eigenvalues are around zero, which lead to an approximate low-rank structure.

parameter matrix $P$. However, unlike the simulation experiment in Section 4.5.1, $P$ is definitely not available in practice since the 114 graphs themselves are also a sample from the population. We address this issue by finding a surrogate estimate for $P$ and using it to calculate the MSE. Recently, Kiar et al. [2016, 2017] updated the ndmg pipeline to a newer version “ndmg-v0.0.33” (ndmg2), which generates graphs of better quality compared to the previous version “ndmg-v0.0.1” (ndmg1). So the MLE derived from the 114 graphs in ndmg2 should be a relatively more accurate estimate of the actual probability matrix $P$ for the population. We use this as our surrogate for $P$ when calculating the MSE. However, such a $P$ generally has full rank, which breaks the low-rank assumptions and thus makes it harder for $\tilde{P}^{(1)}$ and
\( \tilde{P}(q) \) to improve over \( \hat{P}^{(1)} \) and \( \tilde{P}^{(q)} \). Thus any improvement arising from low-rank approximation is likely to be conservative. Moreover, it is still possible that the 114 graphs from ndmg2 contain outliers. Thus by using the MLE of the ndmg2 data as \( P \), the performance of MLqE-related estimators \( \hat{P}^{(q)} \) and \( \tilde{P}^{(q)} \) are also underestimated.

In summary, our approach to constructing a workable surrogate for \( P \) relies on the availability of a better pipeline ndmg2, but is biased against both ASE-based and MLqE-based estimators; still, as we shall see, ASE \( \circ \) MLqE yields the best estimate of our surrogate \( P \).

In this experiment, we build the four estimates based on the sample of size \( m \) from the ndmg1 pipeline, while using the MLE of all 114 graphs from the ndmg2 dataset as the surrogate probability matrix \( P \). Note that diagonal augmentation procedure introduced in Section 3.2.3 is also applied here to compensate for the bias introduced by the zero diagonals of the adjacency matrices. 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 ndmg1 pipeline and compute the four estimates based on the \( m \) sampled graphs. Once again for simplicity, we set \( q \) to be 0.9 without further exploration. However, the results are qualitatively similar for many choices of \( q \). We then compare these estimates to the MLE of all 114 graphs in the ndmg2 dataset. For the two low-rank estimators \( \tilde{P}^{(1)} \) and \( \tilde{P}^{(q)} \), we apply ASE into all possible dimensions, i.e. \( d \) ranges from 1 to \( n \). The MSE results are shown in Figure 4.6.
Figure 4.6: Comparison of MSE of the four estimators using the Desikan atlas ndmg graphs at three sample sizes. Three panels represent different sample sizes $m = 2, 5, 10$. The more samples we have, the better each estimator performs. The horizontal-axis represents the number of embedded dimensions while the vertical-axis characterizes the MSE of each estimator. And four estimators perform similarly according to different dimensions under different sample size $m$. 1. MLE $\hat{P}^{(1)}$ vs. MLqE $\tilde{P}^{(q)}$: $\tilde{P}^{(q)}$ outperforms $\hat{P}^{(1)}$ since in practice observations are always contaminated and robust estimators are preferred; 2. MLE $\hat{P}^{(1)}$ vs. ASE $\circ$ MLE $\tilde{P}^{(1)}$: $\tilde{P}^{(1)}$ wins the bias-variance trade-off when being embedded into a proper dimension; 3. MLqE $\hat{P}^{(q)}$ vs. ASE $\circ$ MLqE $\tilde{P}^{(q)}$: $\tilde{P}^{(q)}$ wins the bias-variance trade-off when being embedded into a proper dimension; 4. ASE $\circ$ MLqE $\tilde{P}^{(q)}$ vs. ASE $\circ$ MLE $\tilde{P}^{(1)}$: $\tilde{P}^{(q)}$ is better, since it inherits the robustness from $\tilde{P}^{(q)}$. The squares and circles represent the dimensions selected by the Zhu and Ghodsi method, which we see are reasonable choices. And more importantly, a wide range of dimensions lead to an improvement.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

When \( d \) is small, the ASE procedures underestimate the dimension and fail to capture 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 ZG dimension of \( \hat{P}^{(1)} \) by square and denote the ZG dimensions of \( \hat{P}^{(q)} \) by circle in the figure. We see that the Zhu and Ghodsi’s algorithm performs adequately for selecting a dimension in which to embed. More importantly, there is a wide range of dimensions for which applying ASE leads to improved performance. Although the \( P \) we are estimating is arguably not low-rank, ASE procedures still win the bias-variance trade-off and improve performance even in this unfavorable setting.

We observed here that the robust estimator \( \tilde{P}^{(q)} \) also performs relatively better than \( \hat{P}^{(1)} \), even though \( P \) (presumably) still contains outliers. This strongly indicates that there are many outliers in the original graphs from the ndmg1 pipeline, and \( \tilde{P}^{(q)} \) successfully inherits the robustness from MLqE and outperforms \( \hat{P}^{(1)} \).

For all three sample sizes \((m = 2, 5, 10)\), \( \tilde{P}^{(q)} \) estimates the surrogate for \( P \) most accurately even though the surrogate for \( P \) is itself an edge-wise MLE, and thus is biased in favor of the other three estimators due to its high rank and non-robustness. As such, we expect \( \tilde{P}^{(q)} \) to provide an even better estimate for the true but unknown \( P \).
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

4.7 Proofs for Theory Results

4.7.1 Outline of the Proofs

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

In Section 4.7.3, we analyze the properties of the ASE procedure. We first prove Theorem 4.7.9 which provides an upper bound for the spectral norm of the difference between the estimator $\hat{P}^{(1)}$ and its expectation $H^{(1)}_{ij} = E[\hat{P}^{(1)}_{ij}]$. Lemma 4.7.11 shows that $U^\top \hat{U}$ can be approximated by an orthogonal matrix $W* = W_1 W_2^\top$, where $U$ and $\hat{U}$ are the eigenspaces with respect to the largest $d$ eigenvalues of $H^{(1)}_{ij}$ and $\hat{P}^{(1)}$ 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 $\|\hat{Z} - ZW\|_F$, which controls the error of the $\hat{Z}$ for estimating the true latent positions $Z$ up to orthogonal transformation. Using Lemma 4.7.13, we can then give a bound for the $2 \to \infty$-norm of $\hat{Z} - ZW$, i.e. we bound $\max_i \|\hat{Z}_i - WZ_i\|_2$ in Theorem 4.7.14.

In Section 4.7.4, we give a bound of the estimation error $\|\hat{Z}^\top_i \hat{Z}_j - Z^\top_i Z_j\|$ in Lemma 4.7.15 based on the results in Section 4.7.3. In order to bound the vari-
of our estimator $\tilde{P}^{(1)}$, all results in this section will be based on a truncated version of $\tilde{P}^{(1)}$ defined in Definition 4.7.16. This is purely for technical reasons and will not affect the estimation procedure in practice, which is discussed in details in Remark 4.7.17. We then bound the expectation (Lemma 4.7.18) and variance (Theorem 4.7.19) of $\tilde{P}^{(1)}$ by carefully choosing a truncation point $a$ and applying the above truncation argument. As a direct result, we obtain the bound for the relative efficiency between $\hat{P}^{(1)}_{ij}$ and $\tilde{P}^{(1)}_{ij}$ in Theorem 4.7.20.

In Section 4.7.5, we compare the performance between $\tilde{P}^{(q)}$ and $\hat{P}^{(q)}$. The results in this section are proved in a similar manner to those in Section 4.7.3 and Section 4.7.4. However, since the ML$q$E estimator for a mixture distribution model does not have a closed form expression, we explore a relationship between MLE and ML$q$E to bound $\tilde{P}^{(q)}$ and $\hat{P}^{(q)}$; this technique could be of independent interest. Finally, in Section 4.7.6, we compare the performance between $\tilde{P}^{(q)}$ and $\tilde{P}^{(1)}$.

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

Before presenting the proofs, we first define the following notion of “with high probability” that is used throughout this work.

**Definition 4.7.1** Let $(E_n)$ for $n \geq 1$ be a sequence of events. We say that the events hold with high probability if, for any constant $c > 0$ there exits a constant $n_0(c)$ such that for all $n \geq n_0$, the event $E_n$ holds with probability greater than $1 - n^{-c}$.
4.7.2 \( \hat{P}(q) \) vs. \( \hat{P}(1) \)

Lemma 4.7.2 Let \( X_1, \cdots, X_m \overset{\text{iid}}{\sim} \text{Exp}(p) \) with \( m \geq 2 \) and \( E[X_1] = p \). Then with probability 1,

- 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 \( \hat{p}^{(q)} \), the root closest to the MLE, is well defined.

Proof: Let \( x_1, \cdots, x_m \) be the observed values of \( X_1, X_2, \ldots, X_m \). Then with probability 1, the \( x_i \) are unique and \( x^{(1)} = \min_i x_i > 0 \). The MLE is

\[
\hat{p}^{(1)}(x) = \bar{x}.
\]

Let \( g(\theta, x) = \sum_{i=1}^{m} e^{-\frac{(1-q)x_i}{\theta}} (x_i - \theta) \). Then the MLq equation is \( g(\theta, x) = 0 \). Now let \( l \) be the smallest index such that \( x^{(1)} \leq \cdots \leq x^{(l)} \leq \bar{x} \leq x^{(l+1)} \leq \cdots \). 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

\[
g(\theta, x) = \sum_{i=1}^{m} e^{-\frac{(1-q)x(i)}{\sigma}}(x(i) - \theta) = \sum_{i=1}^{m} e^{-\frac{(1-q)x(i)}{\sigma}}(x(i) - \bar{x} + \bar{x} - \theta)
\]

\[
= - \sum_{i=1}^{l} e^{-\frac{(1-q)x(i)}{\sigma}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i + \sum_{i=1}^{m} e^{-\frac{(1-q)x(i)}{\sigma}}(\bar{x} - \theta)
\]

\[
\leq - \sum_{i=1}^{l} e^{-\frac{(1-q)x(i)}{\sigma}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i
\]

\[
\leq - e^{-\frac{(1-q)x(l+1)}{\sigma}} \sum_{i=1}^{l} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i
\]

\[
\leq - e^{-\frac{(1-q)x(l+1)}{\sigma}} \sum_{i=1}^{m-l} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i
\]

\[
\leq - \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x(i+1)}{\sigma}} t_i
\]

\[
= 0,
\]

and equality holds if and only if all \( x_i \)'s are the same, which occurs with probability 0. Thus with probability 1, \( g(\theta, x) < 0 \) for all \( \theta \geq \bar{x} \).

Denote any solution to the MLqE equation as \( \hat{p}^{(q)}(x) \); we then have that

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

Thus there exists at least one solution to the MLqE equation. And since all solutions to the MLqE equation are in the interval \( (x_{(1)}, \bar{x}) \), we have \( \hat{p}^{(q)}(x_1, \ldots, x_m) \leq \)
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\( \hat{p}^{(1)}(x_1, \ldots, x_m) \).

**Lemma 4.7.3** Given observed data points \( x_1, \ldots, x_m \), recall that the ML\( q \)E equation under the exponential distribution based on \( m \) data points \( x_1, \ldots, x_m \) is

\[
\sum_{i=1}^{m} \exp(-(1-q)x_i/\theta)(x_i-\theta) = 0
\]

while the MLE equation under the exponential distribution based on the same data is

\[
\sum_{i=1}^{m} (x_i - \theta) = 0.
\]

Now let \( X_1, \ldots, X_m \overset{iid}{\sim} (1-\epsilon)\text{Exp}(p) + \epsilon\text{Exp}(c) \) be \( m \) data points sampled from a mixture of two exponential distribution. Denoting this mixture as \( F \), there exists exactly one real solution \( \theta(F) \) of

\[
E_F[\exp(-(1-q)X/\theta(F))(X-\theta(F))] = 0.
\]

Note that \( E_F[\exp(-(1-q)X/\theta(F))(X-\theta(F))] = 0 \) is the population version of ML\( q \)E equation under the exponential distribution. Moreover, the ML\( q \)E solution is less than the MLE solution under the exponential distribution, i.e. \( \theta(F) < E_F[\bar{X}] = (1-\epsilon)p + \epsilon c \).

**Proof:** For the MLE, i.e. \( \bar{X} \), we have \( E[\bar{X}] = (1-\epsilon)p + \epsilon c \). According to Equation (3.2) in [Ferrari and Yang 2010], \( \theta(F) \) satisfies

\[
\frac{ec}{(c(1-q)+\theta)^2} - \frac{\epsilon}{c(1-q)+\theta} + \frac{(1-\epsilon)p}{(p(1-q)+\theta)^2} - \frac{(1-\epsilon)}{p(1-q)+\theta} = 0,
\]

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...
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\( h(E[\bar{X}]) = 0, \) we have three roots and one of them is \( q = 1 \) obviously. The other two roots are

\[
\frac{(p + c) ((p - c)^2 \epsilon (1 - \epsilon) + 2pc)}{2pc(p\epsilon + c(1 - \epsilon))} \pm \sqrt{\frac{\epsilon(1 - \epsilon)(c - p)^2 (\epsilon(1 - \epsilon)(c - p)^4 - 4p^2\epsilon^2)}{4p^2c^2(p\epsilon + c(1 - \epsilon))^2}}.
\]

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

\[
\frac{(p + c) ((p - c)^2 \epsilon (1 - \epsilon) + 2pc)}{2pc(p\epsilon + c(1 - \epsilon))} - \sqrt{\frac{\epsilon(1 - \epsilon)(c - p)^2 (\epsilon(1 - \epsilon)(c - p)^4 - 4p^2\epsilon^2)}{4p^2c^2(p\epsilon + c(1 - \epsilon))^2}} > 1.
\]

For the first part,

\[
\frac{(p + c) ((p - c)^2 \epsilon (1 - \epsilon) + 2pc)}{2pc(p\epsilon + c(1 - \epsilon))} > 1 + \frac{(p - c)^2 \epsilon (1 - \epsilon)(p + c)}{2pc(p\epsilon + c(1 - \epsilon))}.
\]

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 \( 0 < q < 1, \) \( h(E[\bar{X}]) < 0. \)
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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 \( C_{ij} > C_0(P_{ij}, \epsilon, q) \),

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

for \( 1 \leq i, j \leq n \) and \( i \neq j \).

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

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

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 Assume similar setting as Lemma 4.7.3, i.e. let m data points sample from the contamination model $X, X_1, \ldots, X_m \overset{iid}{\sim} (1 - \epsilon)\text{Exp}(p) + \epsilon\text{Exp}(c)$. The MLqE under the exponential distribution $\text{Exp}(p)$ 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. $E_p[\exp(-(1 - q)X/\theta(F))(X - \theta(F))] = 0$. Let $r(\theta(F)) = p$. Then we can define $\rho(x; \theta) = f(x; r(\theta))^{1-q}/1-q$, where $q \in (0, 1)$ is a constant. By reparameterizing $\rho(x; \theta)$ to $\tilde{\rho}(x; r)$ such that $\tilde{\rho}(x; r(\theta)) = \rho(x; \theta)$, we can use the proof of Lemma 4.7.3.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

directly to prove that $D(\theta_0, \theta) = E_{\theta_0}[\rho(X, \theta)]$ is uniquely minimized at $\theta_0$. Thus the ML$qE$ is a minimum contrast estimator.

Lemma 4.7.6 Uniform convergence of the ML$q$ equation, i.e.

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

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 $\theta$ for all $x > 0$ and measurable function of $x$ at each $\theta$, we have the uniform convergence by Lemma 2.4 in [Newey and McFadden, 1994].

Lemma 4.7.7 $\hat{F}_{ij}^{(q)} \xrightarrow{P} \theta(F_{ij})$ 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 ML$q$ is consistent based on Theorem 5.2.3 in [Bickel and Doksum, 2007].
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

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

And thus

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

Proof: Both MLE and ML$q$E are minimum contrast 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]. \hfill \blacksquare

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

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

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

for $1 \leq t \leq m$. Let $\hat{P}^{(1)}$ be the element-wise MLE based on exponential distribution with $m$ observations. Define $H_{ij}^{(1)} = E[\hat{P}_{ij}^{(1)}] = (1 - \epsilon)P_{ij} + \epsilon C_{ij}$, then for any constant $c > 0$, there exists another constant $n_0(c)$, independent of $n$, $P$, $C$ and $\epsilon$, such that if...
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

$n > n_0$, then for all $\eta$ satisfying $n^{-c} \leq \eta \leq 1/2$,

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

**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 $G_{ij}$:

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

Thus

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

and

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

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

for $1 \leq t \leq m$ and $1 \leq i < j \leq n$. 

---

124
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

$$E[(A_{ij}(t) - H_{ij}^{(1)})^k] \leq (1 - \epsilon) \cdot \exp(-H_{ij}/P_{ij})P_{ij}^k \Gamma(1 + k, -H_{ij}/P_{ij})$$
$$+ \epsilon \cdot \exp(-H_{ij}/C_{ij})C_{ij}^k \Gamma(1 + k, -H_{ij}/C_{ij})$$
$$\leq (1 - \epsilon) \cdot \exp(-H_{ij}/P_{ij})P_{ij}^k + \epsilon \cdot \exp(-H_{ij}/C_{ij})C_{ij}^k) \ k!$$
$$\leq ((1 - \epsilon) \cdot P_{ij}^k + \epsilon \cdot C_{ij}^k) \ k!$$
$$\leq R^k k!, \quad (4.1)$$

Combined with

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

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,

$$E[(X_{ij}^{(t)})^k] = E[(A_{ij}(t) - H_{ij}^{(1)})^k] G_{ij} \preceq k!R^k G_{ij}^2.$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

So

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

Let

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

Notice that random matrices \( X_{ij}^{(t)} \) are independent, self-adjoint and have mean zero, apply Theorem 6.2 in [Tropp 2012] we have

\[
P \left( \lambda_{\max}(\hat{P}^{(1)} - H^{(1)}) \geq t \right) = P \left( \lambda_{\max} \left( \frac{1}{m} \sum_{1 \leq t \leq m, 1 \leq i < j \leq n} X_{ij}^{(t)} \right) \geq t \right)
\]

\[
= P \left( \lambda_{\max} \left( \sum_{1 \leq t \leq m, 1 \leq i < j \leq n} X_{ij}^{(t)} \right) \geq mt \right)
\]

\[
\leq n \exp \left( -\frac{(mt)^2/2}{\sigma^2 + Rmt} \right)
\]

\[
\leq n \exp \left( -\frac{mt^2/2}{2R^2 n + Rt} \right).
\]

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

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

126
1. When \( k \) is even,

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

2. When \( k \) is odd,

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

Thus by similar arguments,

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

Therefore we have

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

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 \( \epsilon \) 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

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

Define $H^{(1)} = E[\hat{P}^{(1)}] = (1 - \epsilon)P + \epsilon C$, where $P = XX^\top$, $X \in \mathbb{R}^{n \times d}$, $C = YY^\top$, $Y \in \mathbb{R}^{n \times d'}$. Let $d^{(1)} = \text{rank}(H^{(1)})$ be the dimension in which we are going to embed $\hat{P}^{(1)}$. Then we can define $H^{(1)} = ZZ^\top$ where $Z \in \mathbb{R}^{n \times d^{(1)}}$. Since $H^{(1)} = [\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y][\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y]^\top$, we have $d^{(1)} \leq d + d'$.

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

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

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

**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 $\|\hat{S}\|_2 = \Theta(n)$ with high probability.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

Proof: Note that $\lambda_i(H) = \lambda_i(ZZ^\top) = \lambda_i(Z^\top Z)$ when $i \leq k$. Since each entry of $Z^\top Z$ is a sum of $n$ independent random variables each in $[0, R]$, i.e. $(Z^\top Z)_{ij} = \sum_{i=1}^n Z_i Z_{ij}$. By Hoeffding’s inequality,

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

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

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

we have

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

By the union bound, we have

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

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

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

with probability at least $1 - k^2 \eta$. Thus $\lambda_i(H) = S_{ii} = \Theta(n)$ with probability at least
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

1 - \frac{2k^2}{n^2}$ when $i \leq k$. Moreover,

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

Combined with Theorem [4.7.9] with high probability we have $\|\hat{S}\|_2 = \Theta(n)$.

**Lemma 4.7.11** Let $W_1 \Sigma W_2^\top$ be the singular value decomposition of $U^\top \hat{U}$. Then for sufficiently large $n$,

$$\|U^\top \hat{U} - W_1 W_2^\top\|_F = O(m^{-1} n^{-1} \log n)$$

with high probability.

**Proof:** Let $\sigma_1, \cdots, \sigma_k$ denote the singular values of $U^\top \hat{U}$. Then $\sigma_i = \cos(\theta_i)$ where the $\theta_i$ are the principal angles between the subspaces spanned by $\hat{U}$ 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]

$$\|\hat{U} \hat{U}^\top - UU^\top\|_2 = \max_i |\sin(\theta_i)|$$

$$\leq \frac{\|\hat{P} - H\|_2}{\lambda_k(H)} \leq C \frac{\sqrt{n \log n / m}}{n}$$  

$$= O(m^{-1/2} n^{-1/2} \sqrt{\log n})$$  

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

$$
\|U^\top \hat{U} - W_1 W_2^\top \|_F = \|\Sigma - I\|_F = \sqrt{\sum_{i=1}^{k} (1 - \sigma_i)^2}
\leq \sum_{i=1}^{k} (1 - \sigma_i) \leq \sum_{i=1}^{k} (1 - \sigma_i^2)
= \sum_{i=1}^{k} \sin^2(\theta_i) \leq k \|\hat{U}^\top \hat{U} - UU^\top\|_2^2
= O(m^{-1} n^{-1} \log n).
$$

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

**Lemma 4.7.12** For sufficiently large $n$,

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

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

and

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

with high probability.

**Proof:** By Proposition 2.1 in [Rohe et al., 2011] and Equation (4.2), we have for
some orthogonal matrix $W$,

$$
\| \hat{U} - UW \|_F^2 \leq \frac{2\| \hat{U}\hat{U}^\top - UU^\top \|_F^2}{\delta^2} \leq \frac{8k^2\| \hat{U}\hat{U}^\top - UU^\top \|_2^2}{\delta^2} = O(m^{-1}n^{-1}\log n),
$$

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

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

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

$$
W^* \hat{S} = (W^* - U^\top \hat{U}) \hat{S} + U^\top \hat{U} \hat{S} = (W^* - U^\top \hat{U}) \hat{S} + U^\top \hat{P} \hat{U}
$$

$$
= (W^* - U^\top \hat{U}) \hat{S} + U^\top (\hat{P} - H) \hat{U} + U^\top H \hat{U}
$$

$$
= (W^* - U^\top \hat{U}) \hat{S} + U^\top (\hat{P} - H) Q + U^\top (\hat{P} - H) UU^\top \hat{U} + U^\top H \hat{U}
$$

$$
= (W^* - U^\top \hat{U}) \hat{S} + U^\top (\hat{P} - H) Q + U^\top (\hat{P} - H) UU^\top \hat{U} + SU^\top \hat{U}.
$$
Combined with Theorem [4.7.9], Lemma [4.7.10], Lemma [4.7.11], we have

\[ \|W^*\hat{S} - SW^*\|_F \]

\[ = \|(W^* - U^T\hat{U})\hat{S} + U^T(\hat{P} - H)Q + U^T(\hat{P} - H)UU^T\hat{U} + S(U^T\hat{U} - W^*)\|_F \]

\[ \leq \|W^* - U^T\hat{U}\|_F(\|\hat{S}\|_2 + \|S\|_2) + \|U^T\|_F\|\hat{P} - H\|_2\|Q\|_F + \|U^T(\hat{P} - H)U\|_F \]

\[ \leq O(m^{-1} \log n) + O(m^{-1/2} \log n) + \|U^T(\hat{P} - H)U\|_F \]

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

\[ u_i^T(\hat{P} - H)u_j = \sum_{s=1}^{n} \sum_{t=1}^{n} (\hat{P}_{st} - H_{st})u_isu_jt = 2 \sum_{s<t} (\hat{P}_{st} - H_{st})u_isu_jt \]

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

By Equation [4.1], we have

\[ E\left[\left((A^{(t')}_{st} - H_{st})u_isu_jt\right)^k\right] \leq k!R^k u_is^k u_jt^k \leq \frac{k!}{2} R^{k-2}(\sqrt{2}u_isu_jtR)^2. \]

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

\[
P \left( \left| \sum_{s<t} (\hat{P}_{st} - H_{st})u_isu_jt \right| \geq t \right) \leq \exp \left( \frac{-mt^2/8}{R^2 + Rt/2} \right).
\]

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

\[
P \left( \left| \sum_{s<t} (\hat{P}_{st} - H_{st})u_isu_jt \right| \geq Cm^{-1/2} \log n \right) \leq n^{-c}.
\]

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

\[
\|U^T (\hat{P} - H) U\|_F = O(m^{-1/2} \log n)
\]

(4.4)

with high probability. Hence

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

with high probability. Also, since

\[
W_{ij}^* (\lambda_j^{1/2}(\hat{P}) - \lambda_i^{1/2}(H)) = W_{ij}^* \frac{\lambda_j(\hat{P}) - \lambda_i(H)}{\lambda_j^{1/2}(\hat{P}) + \lambda_i^{1/2}(H)}
\]

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

\[
\|W^* \hat{S}^{1/2} - S^{1/2} W^*\|_F = O(m^{-1/2} n^{-1/2} \log n)
\]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

with high probability. Similarly, since

\[ W_{ij}^*(\lambda_j^{-1/2}(\hat{P}) - \lambda_i^{-1/2}(H)) = W_{ij}^* \frac{\lambda_i(H) - \lambda_j(\hat{P})}{(\lambda_j^{-1/2}(\hat{P}) + \lambda_i^{-1/2}(H))\lambda_j(\hat{P})\lambda_i(H)} \]

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

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

\[ \square \]

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

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

with high probability.

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

135
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\[ \hat{U} - UU^\top \hat{U} + Q_1 = Q + Q_1. \]  
Then since \( UU^\top H = H \) and \( \hat{U}S^{1/2} = \hat{P}\hat{U}S^{-1/2} \),

\[ \hat{Z} - US^{1/2}W^* = \hat{U}S^{1/2} - UW^*S^{1/2} + U(W^*S^{1/2} - S^{1/2}W^*) \]
\[ = (\hat{U} - UU^\top \hat{U})S^{1/2} + Q_1S^{1/2} + UQ_2 \]
\[ = (\hat{P} - H)\hat{U}S^{-1/2} - UU^\top (\hat{P} - H)\hat{U}S^{-1/2} + Q_1S^{1/2} + UQ_2 \]
\[ = (\hat{P} - H)UU^\top S^{-1/2} - UU^\top (\hat{P} - H)UW^*S^{-1/2} \]
\[ + (I - UU^\top) (\hat{P} - H)Q_3S^{-1/2} + Q_1S^{1/2} + UQ_2. \]

By Lemma 4.7.11, with high probability,

\[ \|Q_1\|_F \leq \|U\|_F \|U^\top \hat{U} - W^*\|_F = O(m^{-1}n^{-1}\log n). \]

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,

\[ \|Q_3\|_F \leq \|Q\|_F + \|Q_1\|_F = O(m^{-1/2}n^{-1/2}(\log n)^{1/2}). \]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

By Equation (4.4), with high probability,

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

By Lemma 4.7.12 with high probability,

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

Therefore, with high probability,

$$\|\hat{Z} - US^{1/2} W^*\|_F$$

$$=\|(\hat{P} - H) U W^* \hat{S}^{-1/2}\|_F + O(m^{-1/2} n^{-1/2} \log n) + \|I - UU^\top\|_2 \|\hat{P} - H\|_2 O(m^{-1/2} n^{-1/2} (\log n)^{1/2})$$

$$+ O(m^{-1/2} n^{-1/2} \log n) + O(m^{-1/2} n^{-1/2} \log n)$$

$$=\|(\hat{P} - H) U W^* \hat{S}^{-1/2}\|_F + O(m^{-1/2} n^{-1/2} \log n)$$

$$\leq\|(\hat{P} - H) U S^{-1/2} W^*\|_F + \|(\hat{P} - H) U (W^* \hat{S}^{-1/2} - S^{-1/2} W^*)\|_F + O(m^{-1/2} n^{-1/2} \log n)$$

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

$$=\|(\hat{P} - H) U S^{-1/2}\|_F + O(m^{-1/2} n^{-1/2} (\log n)^{3/2}).$$

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

**Theorem 4.7.14** There exists a rotation matrix $W$ such that for sufficiently large
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\[ n, \quad \max_i \| \hat{Z}_i - WZ_i \|_2 = O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \]

with high probability.

**Proof:** By Lemma 4.7.13 we have

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

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

\[ \max_i \| \hat{Z}_i - WZ_i \|_2 \leq \frac{1}{\lambda_k^{1/2}(H)} \max_i \| (\hat{P} - H)U \|_2 + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \]

\[ \leq \frac{k^{1/2}}{\lambda_k^{1/2}(H)} \max_j \| (\hat{P} - H)u_j \|_\infty + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \]

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

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

Thus, conditioned on \(H\), the \(i\)-th element of the vector \((\hat{P} - H)u_j\) is a sum of inde-
pendent mean 0 random variables. By Equation (4.1), we have

\[ E \left[ \left( (A_{is}^{(t)} - H_{is})u_{js} \right)^k \right] \leq k! R^k u^k_{js} \leq \frac{k!}{2} R^{-2} (\sqrt{2} Ru_{js})^2. \]

Also we have

\[ \sigma^2 := | \sum_{t,s \neq i} 2R^2 u^2_{js} | \leq 2R^2 m, \]

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

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

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

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

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,

\[ \max_i \| \hat{Z}_i - WZ_i \|_2 \leq \frac{Ck^{1/2}}{\lambda_k^{1/2} (H)} m^{-1/2} (\log n)^{3/2} + O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) \]

\[ = O(m^{-1/2} n^{-1/2} (\log n)^{3/2}). \]
4.7.4 $\widetilde{P}^{(1)}$ vs. $\widehat{P}^{(1)}$

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

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

$$\left| \hat{Z}^\top_i \hat{Z}_j - Z^\top_i Z_j \right| = \left| \hat{Z}^\top_i \hat{Z}_j - \hat{Z}^\top_i W Z_j + \hat{Z}^\top_i W Z_j - (W Z_i)^\top W Z_j \right|$$

$$\leq \left| \hat{Z}^\top_i (\hat{Z}_j - W Z_j) + (\hat{Z}^\top_i - (W Z_i)^\top) W Z_j \right|$$

$$\leq \| \hat{Z}_i \|_2 \| \hat{Z}_j - W Z_j \|_2 + \| Z_j \|_2 \| \hat{Z}^\top_i - (W Z_i)^\top \|_2.$$

Since $\|Z_i\|_2^2 = Z^\top_i Z_i = H^{(1)}_{ii} = E[\hat{P}^{(1)}_{ii}] = (1 - \epsilon)P_{ij} + \epsilon C_{ij} \leq R$, we have $\|Z_i\|_2 = O(1)$.

Combined with Theorem 4.7.14

$$\left| \hat{Z}^\top_i \hat{Z}_j - Z^\top_i Z_j \right| = (\| \hat{Z}_i \|_2 + \| Z_j \|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

$$\leq (\| \hat{Z}_i - W Z_i \|_2 + \| W Z_i \|_2 + \| Z_j \|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

$$= O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability.

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

140
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.

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

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

$$
E[|\hat{Z}_i^\top \hat{Z}_j|_{\text{tr}} - Z_i^\top Z_j|]
= E[|\hat{Z}_i^\top \hat{Z}_j|_{\text{tr}} - Z_i^\top Z_j\mathbb{I}\{\hat{P}_{ij} \leq a\}] + E[|\hat{Z}_i^\top \hat{Z}_j|_{\text{tr}} - Z_i^\top Z_j\mathbb{I}\{\hat{P}_{ij} > a\}].
$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

For the first term, we have

\[
E[|\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j| \text{tr}| \{\hat{P}_{ij} \leq a\}] \\
\leq E[|\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j| \text{tr}| \{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 holds}\}] \\
+ E[|\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j| \text{tr}| \{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\leq E[|\mathbf{Z}_i^\top \mathbf{Z}_j| \text{tr}| \{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 holds}\}] \\
+ n^{-c} E[|\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j| \text{tr}| \{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\leq O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\
+ n^{-c} E[|\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j| \text{tr}| \{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
+ n^{-c} E[|\hat{\mathbf{P}}_{ij} - \mathbf{Z}_i^\top \mathbf{Z}_j| \mathbb{I}\{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\leq O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + n^{-c} E[|\hat{\mathbf{P}}_{ij} - \mathbf{Z}_i^\top \mathbf{Z}_j| \mathbb{I}\{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
+ n^{-c} E[|\hat{\mathbf{P}}_{ij} - \mathbf{Z}_i^\top \mathbf{Z}_j| \mathbb{I}\{\hat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.15 does not hold}\}] \\
\leq O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + an^{-c} + (a + R)n^{-c} \\
\leq O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a + R).
Notice that

\[
E[\hat{P}_{ij}\mathbb{I}\{\hat{P}_{ij} > a\}] = E\left[\left(\frac{1}{m} \sum_{1 \leq t \leq m} A_{ij}^{(t)}\right) \mathbb{I}\{\hat{P}_{ij} > a\}\right]
\]

\[
= \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)} \mathbb{I}\{\hat{P}_{ij} > a\}\right] \leq \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)} \mathbb{I}\{\max_{1 \leq s \leq m} A_{ij}^{(s)} > a\}\right]
\]

\[
\leq \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)} \left( \sum_{1 \leq s \leq m} \mathbb{I}\{A_{ij}^{(s)} > a\}\right)\right] = E[A_{ij}^{(1)} \left( \sum_{1 \leq s \leq m} \mathbb{I}\{A_{ij}^{(s)} > a\}\right)]
\]

\[
= E[A_{ij}^{(1)}] + (m - 1) E[A_{ij}^{(1)}] P(A_{ij}^{(2)} > a)]
\]

\[
= E[A_{ij}^{(1)}] + (m - 1) E[A_{ij}^{(1)}] P(A_{ij}^{(1)} > a),
\]

and similarly

\[
E[(\hat{P}_{ij} + R)\mathbb{I}\{\hat{P}_{ij} > a\}] = E[\hat{P}_{ij}\mathbb{I}\{\hat{P}_{ij} > a\}] + R \cdot P(\hat{P}_{ij} > a)
\]

\[
\leq E[A_{ij}^{(1)}] + (m - 1) E[A_{ij}^{(1)}] P(A_{ij}^{(1)} > a) + R \cdot m \cdot P(A_{ij}^{(1)} > a).
\]
Thus for the second term,
\[
E[(\hat{Z}_i^\top \hat{Z}_j)_{ij} - Z_i^\top Z_j][\{\hat{P}_{ij} > a\}]
\]
\[
\leq E[(\hat{Z}_i^\top \hat{Z}_j)_{ij} - \hat{P}_{ij}[\{\hat{P}_{ij} > a\}] + E[|\hat{P}_{ij} - Z_i^\top Z_j|][\{\hat{P}_{ij} > a\}]
\]
\[
\leq E[\hat{P}_{ij}[\{\hat{P}_{ij} > a\}] + E[(\hat{P}_{ij} + R)[\{\hat{P}_{ij} > a\}]
\]
\[
\leq 2E[A_{ij}^{(1)}[\{A_{ij}^{(1)} > a\}]] + 2(m - 1)E[A_{ij}^{(1)}P(A_{ij}^{(1)} > a)
\]
\[
+ R \cdot m \cdot P(A_{ij}^{(1)} > a)
\]
\[
\leq 2e^{-a/R}(a + 2mR).
\]

Thus
\[
E[(\hat{Z}_i^\top \hat{Z}_j)_{ij} - Z_i^\top Z_j]
\]
\[
\leq O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a + R) + 2e^{-a/R}(a + 2mR).
\]

Let \(a = m^{-1}n^{2b}\) for any \(b > 0\), and \(c = 2b + 3\), combined with the assumption
m = O(n^b), we have

\[
E[(\hat{Z}_i^\top \hat{Z}_j)_{ij} - Z_i^\top Z_j] = O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b} \cdot O(e^{-m^{-1} n^{2b}}))
\]
\[
= O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b} \cdot O(e^{-n^b}))
\]
\[
= O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3}) + O(m^{-1} n^{2b}) \cdot O(n^{-2b-3})
\]
\[
= O(m^{-1/2} n^{-1/2} (\log n)^{3/2}) + O(m^{-1} n^{-3})
\]
\[
= O(m^{-1/2} n^{-1/2} (\log n)^{3/2}).
\]

\[\blacksquare\]

**Theorem 4.7.19** Assuming that \( m = O(n^b) \) for any \( b > 0 \), then \( \text{Var}((\hat{Z}_i^\top \hat{Z}_j)_{ij}) = O(m^{-1} n^{-1} (\log n)^3) \).
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Proof: By Lemma 4.7.15

\[
\text{Var}((\hat{Z}_i^\top \hat{Z}_j)_{tr}) = E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])^2]
\]
\[
= E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j + Z_i^\top Z_j - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])^2]
\]
\[
= E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2] + E[(Z_i^\top Z_j - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])^2]
\]
\[
+ 2E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)(Z_i^\top Z_j - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])]
\]
\[
\leq E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2] + E[(Z_i^\top Z_j - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])^2]
\]
\[
+ 2\sqrt{E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2]E[(Z_i^\top Z_j - E[(\hat{Z}_i^\top \hat{Z}_j)_{tr}])^2]}
\]
\[
\leq 4E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2].
\]

Fix some \(a > 0\), we have

\[
E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2]
\]
\[
= E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2I\{\hat{P}_{ij} \leq a\}] + E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2I\{\hat{P}_{ij} > a\}).
\]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

For the first term, we have

\[ E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\}] \]
\[ \leq E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\} I\{\text{Lemma 4.7.15 holds}\}] \]
\[ + E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\} I\{\text{Lemma 4.7.15 does not hold}\}] \]
\[ \leq E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 holds}] \]
\[ + n^{-c} E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 does not hold}] \]
\[ \leq O(m^{-1} n^{-1} (\log n)^3) + 2n^{-c} E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - \hat{P}_{ij})^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 does not hold}] \]
\[ + 2n^{-c} E[(\hat{P}_{ij} - Z_i^\top Z_j)^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 does not hold}] \]
\[ \leq O(m^{-1} n^{-1} (\log n)^3) + 2n^{-c} E[\hat{P}_{ij}^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 does not hold}] \]
\[ + 2n^{-c} E[\hat{P}_{ij}^2 + R)^2 I\{\hat{P}_{ij} \leq a\} | \text{Lemma 4.7.15 does not hold}] \]
\[ \leq O(m^{-1} n^{-1} (\log n)^3) + 2a^2 n^{-c} + 2(a + R)^2 n^{-c} \]
\[ \leq O(m^{-1} n^{-1} (\log n)^3) + 4n^{-c}(a + R)^2. \]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Notice that

\[ E[\hat{P}_{ij}^2 \{ \hat{P}_{ij} > a \}] = E\left[\left( \frac{1}{m} \sum_{1 \leq t \leq m} A_{ij}^{(t)} \right)^2 \{ \hat{P}_{ij} > a \} \right] \]

\[ \leq \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)^2} \{ \hat{P}_{ij} > a \} \right] \leq \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)^2} \{ \max_{1 \leq s \leq m} A_{ij}^{(s)} > a \} \right] \]

\[ \leq \frac{1}{m} E\left[ \sum_{1 \leq t \leq m} A_{ij}^{(t)^2} \left( \sum_{1 \leq s \leq m} \mathbb{I}\{ A_{ij}^{(s)} > a \} \right) \right] = E[A_{ij}^{(1)^2} \left( \sum_{1 \leq s \leq m} \mathbb{I}\{ A_{ij}^{(s)} > a \} \right) \]

\[ = E[A_{ij}^{(1)^2} \{ A_{ij}^{(1)} > a \}] + (m - 1) E[A_{ij}^{(1)^2} \{ A_{ij}^{(2)} > a \}] \]

\[ = E[A_{ij}^{(1)^2} \{ A_{ij}^{(1)} > a \}] + (m - 1) E[A_{ij}^{(1)^2}] P(A_{ij}^{(1)} > a), \]

and similarly

\[ E[\hat{P}_{ij} + R]^2 \{ \hat{P}_{ij} > a \}] \]

\[ = E[\hat{P}_{ij}^2] \{ \hat{P}_{ij} > a \} + 2R \cdot E[\hat{P}_{ij}] \{ \hat{P}_{ij} > a \} + R^2 P(\hat{P}_{ij} > a) \]

\[ \leq E[A_{ij}^{(1)^2} \{ A_{ij}^{(1)} > a \}] + (m - 1) E[A_{ij}^{(1)^2}] P(A_{ij}^{(1)} > a) \]

\[ + 2R \left( E[A_{ij}^{(1)} \{ A_{ij}^{(1)} > a \}] + (m - 1) E[A_{ij}^{(1)}] P(A_{ij}^{(1)} > a) \right) \]

\[ + R^2 \cdot m \cdot P(A_{ij}^{(1)} > a). \]
Thus for the second term,

\[
E[(\hat{Z}_i^\top \hat{Z}_j - Z_i^\top Z_j)^2 \mathbb{I}\{\hat{P}_{ij} > a\}]
\leq 2E[((\hat{Z}_i^\top \hat{Z}_j - \hat{P}_{ij})^2 \mathbb{I}\{\hat{P}_{ij} > a\} + 2E[(\hat{P}_{ij} - Z_i^\top Z_j)^2 \mathbb{I}\{\hat{P}_{ij} > a\}]
\leq 2E[\hat{P}_{ij}^2 \mathbb{I}\{\hat{P}_{ij} > a\} + 2E[\hat{P}_{ij} + R)^2 \mathbb{I}\{\hat{P}_{ij} > a\}]
\leq 4E[A_{ij}^{(1)} \mathbb{I}\{A_{ij}^{(1)} > a\}] + 4(m - 1)E[A_{ij}^{(1)^2}]P(A_{ij}^{(1)} > a)
+ 4R \cdot E[A_{ij}^{(1)} \mathbb{I}\{A_{ij}^{(1)} > a\}] + 2R(m - 1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)} > a)
+ 2R^2 \cdot m \cdot P(A_{ij}^{(1)} > a)
\leq 4e^{-a/R} (\alpha^2 + 3Ra + 3(m + 1)R^2)
\leq 4e^{-a/R} (\alpha + 2m^{1/2}R)^2.
\]

Thus,

\[
\text{Var}((\hat{Z}_i^\top \hat{Z}_j)_{tr}) \leq O(m^{-1}n^{-1}(\log n)^3) + 16(a + R)^2n^{-c} + 16(a + 2m^{1/2}R)^2e^{-a/R}.
\]

Let \( a = m^{-1/2}n^b \) for any \( b > 0 \), and \( c = 2b + 3 \), combined with the assumption
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

\[ m = O(n^b), \text{ we have} \]

\[ \text{Var}(\hat{Z}_i^\top \hat{Z}_j) = O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(e^{-m^{-1/2}n^b}) \]

\[ = O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(e^{-n^{b/2}}) \]

\[ = O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b}) \cdot O(n^{-2b-3}) \]

\[ = O(m^{-1}n^{-1}(\log n)^3) + O(m^{-1}n^{-3}) \]

\[ = O(m^{-1}n^{-1}(\log n)^3). \]

\[ \medspace \]

**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 \( i \neq j \),

\[ \frac{\text{Var}(\tilde{P}_{ij}^{(1)})}{\text{Var}(\hat{P}_{ij}^{(1)})} = O(n^{-1}(\log n)^3). \]

And thus

\[ \text{ARE}(\tilde{P}_{ij}^{(1)}, \hat{P}_{ij}^{(1)}) = 0. \]

**Proof:** The results are direct from Theorem 4.7.19 and Theorem 4.3.1. \[ \medspace \]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

4.7.5 $\tilde{P}^{(q)}$ vs. $\hat{P}^{(q)}$

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

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

Proof: Similar to the proof of Theorem 4.7.9

By Lemma 4.7.2 we have

$$\left|\hat{P}^{(q)}_{ij} - H^{(q)}_{ij}\right| = \left|\hat{P}^{(q)}_{ij} - \hat{P}^{(1)}_{ij} + \hat{P}^{(1)}_{ij} - H^{(1)}_{ij} + H^{(1)}_{ij} - H^{(q)}_{ij}\right|$$

$$\leq \left|\hat{P}^{(q)}_{ij} - \hat{P}^{(1)}_{ij}\right| + \left|\hat{P}^{(1)}_{ij} - H^{(1)}_{ij}\right| + \left|H^{(1)}_{ij} - H^{(q)}_{ij}\right|$$

$$\leq \hat{P}^{(1)}_{ij} + \left|\hat{P}^{(1)}_{ij} - H^{(1)}_{ij}\right| + H^{(1)}_{ij}$$

$$\leq 2 \left(\left|\hat{P}^{(1)}_{ij} - H^{(1)}_{ij}\right| + H^{(1)}_{ij}\right).$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Also,

\[
E[(\hat{P}_{ij}^{(q)} - H_{ij}^{(q)})^k] \leq E \left[ \left| \hat{P}_{ij}^{(q)} - H_{ij}^{(q)} \right|^k \right]
\]

\[
\leq 2^k E \left[ \left( \left| \hat{P}_{ij}^{(1)} - H_{ij}^{(1)} \right| + H_{ij}^{(1)} \right)^k \right]
\]

\[
\leq 2^k \sum_{s=0}^{k} \binom{k}{s} E \left[ \left| \hat{P}_{ij}^{(1)} - H_{ij}^{(1)} \right|^s \right] (H_{ij}^{(1)})^{k-s}
\]

\[
\leq 2^k \sum_{s=0}^{k} \binom{k}{s} R^s s! (H_{ij}^{(1)})^{k-s}
\]

\[
\leq 2^k k! \sum_{s=0}^{k} \binom{k}{s} R^s (H_{ij}^{(1)})^{k-s}
\]

\[
= 2^k k! \left( R + H_{ij}^{(1)} \right)^k
\]

\[
\leq 2^{2k} R^k.
\] (4.5)

Therefore we have

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

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 \( \epsilon \) 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

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

As we define $H(q) = E[\hat{P}(q)]$, let $d(q) = \text{rank}(H(q))$ be the dimension in which we are going to embed $\hat{P}(q)$. Notice that it is less than or equal to $K \times K'$ based on the SBM assumption. Then we can define $H(q) = ZZ^\top$ where $Z \in \mathbb{R}^{n \times d(q)}$.

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

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

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

**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 $\|\hat{S}\|_2 = \Theta(n)$ with high probability.

**Proof:** Exactly the same as proof for Lemma 4.7.10.
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

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

with high probability.

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

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

**Lemma 4.7.24** For sufficiently large $n$,

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

$$\|W^* \hat{S}_1^{1/2} - S^{1/2}W^*\|_F = O(n^{-1/2} \log n)$$

and

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

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 $n$,

$$\|\hat{Z} - ZW\|_F = \|(\hat{P} - H)US^{-1/2}\|_F + O(n^{-1/2} (\log n)^{3/2})$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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 $n$,

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

with high probability.

**Proof:** Similar to the proof of Theorem 4.7.14.

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

**Proof:** Similar to the proof of Lemma 4.7.15.

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

**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 ML$qE$ has the same entry-wise asymptotic bias as ML$qE$, i.e.

$$
\lim_{n \to \infty} \text{Bias}(\hat{P}_{ij}^{(q)}) = \lim_{n \to \infty} E[\hat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n \to \infty} E[\hat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n \to \infty} \text{Bias}(\hat{P}_{ij}^{(q)}).
$$
Proof: Fix some $a > 0$, we have

$$E[|((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)|]
= E[|((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)I\{\hat{P}_{ij}^{(1)} \leq a\}] + E[|((\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)I\{\hat{P}_{ij}^{(1)} > a\}].$$

Note that we are thresholding according to $\hat{P}^{(1)}$ instead of $\hat{P}^{(q)}$. By Lemma 4.7.2, we
know $\hat{P}(q) < \hat{P}(1)$ given any data. For the first term, we have

$$E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a}\]$$

$$\leq E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a}\{\text{Lemma 4.7.27 holds}\}$$

$$+ E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$\leq E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 holds}\}$$

$$+ n^{-c} E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$\leq O(n^{-1/2} (\log n)^{3/2})$$

$$+ n^{-c} E[|(\hat{Z}_i^\top \hat{Z}_j)_\text{tr} - \hat{P}_{ij}^{(1)}|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$+ n^{-c} E[|\hat{P}_{ij}^{(1)} - Z_i^\top Z_j|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$\leq O(n^{-1/2} (\log n)^{3/2})$$

$$+ n^{-c} E[|\hat{P}_{ij}^{(1)}|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$+ n^{-c} E[|\hat{P}_{ij}^{(1)} + R|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$\leq O(n^{-1/2} (\log n)^{3/2})$$

$$+ n^{-c} E[|\hat{P}_{ij}^{(1)}|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$+ n^{-c} E[|\hat{P}_{ij}^{(1)} + R|\{\hat{P}_{ij}^{(1)} \leq a\}\{\text{Lemma 4.7.27 does not hold}\}$$

$$\leq O(n^{-1/2} (\log n)^{3/2}) + an^{-c} + (a + R)n^{-c}$$

$$\leq O(n^{-1/2} (\log n)^{3/2}) + 2n^{-c}(a + R).$$
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

For the second term, we have

\[
E[(\hat{Z}_i^\top \hat{Z}_j)_{\text{tr}} - Z_i^\top Z_j|\{\hat{P}^{(1)}_{ij} > a\}]
\]

\[
\leq E[(\hat{Z}_i^\top \hat{Z}_j)_{\text{tr}} - \hat{P}^{(q)}_{ij} |\{\hat{P}^{(1)}_{ij} > a\}] + E[|\hat{P}^{(q)}_{ij} - Z_i^\top Z_j|\{\hat{P}^{(1)}_{ij} > a\}]
\]

\[
\leq E[|\hat{P}^{(q)}_{ij} |\{\hat{P}^{(1)}_{ij} > a\}] + E[|\hat{P}^{(q)}_{ij} - Z_i^\top Z_j|\{\hat{P}^{(1)}_{ij} > a\}]
\]

\[
\leq E[|\hat{P}^{(1)}_{ij} |\{\hat{P}^{(1)}_{ij} > a\}] + E[|\hat{P}^{(1)}_{ij} + R|\{\hat{P}^{(1)}_{ij} > a\}]
\]

\[
\leq 2e^{-a/R}(a + 2mR).
\]

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

\[
E||(\hat{Z}_i^\top \hat{Z}_j)_{\text{tr}} - Z_i^\top Z_j|| = O(n^{-1/2}(\log n)^{3/2}).
\]

\[\blacksquare\]

**Theorem 4.7.30** Assuming that \( m = O(n^b) \) for any \( b > 0 \), then

\[
\text{Var}((\hat{Z}_i^\top \hat{Z}_j)_{\text{tr}}) = O(n^{-1}(\log n)^3).
\]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

Proof: By Lemma 4.7.27

\[
\text{Var}((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}) = E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])^2] \\
= E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j + \mathbf{Z}_i^\top \mathbf{Z}_j - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])^2] \\
= E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2] + E[(\mathbf{Z}_i^\top \mathbf{Z}_j - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])^2] \\
+ 2E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)(\mathbf{Z}_i^\top \mathbf{Z}_j - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])] \\
\leq E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2] + E[(\mathbf{Z}_i^\top \mathbf{Z}_j - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])^2] \\
+ 2\sqrt{E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2]E[(\mathbf{Z}_i^\top \mathbf{Z}_j - E[(\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr}])^2]} \\
\leq 4E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2].
\]

Fix some \( a > 0 \), we have

\[
E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2] \\
= E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2 \mathbb{I}\{\hat{P}_{ij}^{(1)} \leq a\}] + E[((\hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j)_{tr} - \mathbf{Z}_i^\top \mathbf{Z}_j)^2 \mathbb{I}\{\hat{P}_{ij}^{(1)} > a\}].
\]

Note that we are thresholding according to \( \hat{P}^{(1)} \) instead of \( \hat{P}^{(q)} \). By Lemma 4.7.2 we
know $\hat{P}^{(q)} < \hat{P}^{(1)}$ given any data. For the first term, we have

$$E[((\hat{Z}_i^T \hat{Z}_j)_{tr} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\}]$$

$$\leq E[((\hat{Z}_i^T \hat{Z}_j)_{tr} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 holds}\}]$$

$$+ E[((\hat{Z}_i^T \hat{Z}_j)_{tr} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$\leq E[((\hat{Z}_i^T \hat{Z}_j)_{tr} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 holds}\}]$$

$$+ n^{-c} E[((\hat{Z}_i^T \hat{Z}_j)_{tr} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$\leq O(n^{-1}(\log n)^3)$$

$$+ 2n^{-c} E[(\hat{Z}_i^T \hat{Z}_j)_{tr} - \hat{P}^{(q)}_{ij})^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$+ 2n^{-c} E[(\hat{P}^{(q)}_{ij} - Z_i^T Z_j)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$\leq O(n^{-1}(\log n)^3)$$

$$+ 2n^{-c} E[\hat{P}^{(q)}_{ij}^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$+ 2n^{-c} E[(\hat{P}^{(q)}_{ij} + R)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$\leq O(n^{-1}(\log n)^3) + 2n^{-c} E[\hat{P}^{(1)}_{ij}^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$+ 2n^{-c} E[(\hat{P}^{(1)}_{ij} + R)^2 \mathbb{I}\{\hat{P}^{(1)}_{ij} \leq a\} \mathbb{I}\{\text{Lemma 4.7.27 does not hold}\}]$$

$$\leq O(n^{-1}(\log n)^3) + 2a^2 n^{-c} + 2(a + R)^2 n^{-c}$$

$$\leq O(n^{-1}(\log n)^3) + 4n^{-c}(a + R)^2.$$
For the second term, we have

\[
E[(\hat{Z}_i^\top \hat{Z}_j)_{tr} - Z_i^\top Z_j)^2 I\{\hat{P}^{(1)}_{ij} > a\}] \\
\leq 2E[((\hat{Z}_i^\top \hat{Z}_j)_{tr} - \hat{P}^{(q)}_{ij})^2 I\{\hat{P}^{(1)}_{ij} > a\}] + 2E[(\hat{P}^{(1)}_{ij} - Z_i^\top Z_j)^2 I\{\hat{P}^{(1)}_{ij} > a\}] \\
\leq 2E[\hat{P}^{(q)}_{ij}^2 I\{\hat{P}^{(1)}_{ij} > a\}] + 2E[(\hat{P}^{(q)}_{ij} + R)^2 I\{\hat{P}^{(1)}_{ij} > a\}] \\
\leq 2E[\hat{P}^{(1)}_{ij}^2 I\{\hat{P}^{(1)}_{ij} > a\}] + 2E[(\hat{P}^{(1)}_{ij} + R)^2 I\{\hat{P}^{(1)}_{ij} > a\}] \\
\leq 4e^{-a/R}(a + 2m^{1/2}R^2).
\]

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

\[
\text{Var}((\hat{Z}_i^\top \hat{Z}_j)_{tr}) = O(n^{-1}(\log n)^3).
\]

\[\Box\]

**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 \( i \neq j \),

\[
\frac{\text{Var}(\hat{P}^{(q)}_{ij})}{\text{Var}(\hat{P}^{(q)}_{ij})} = O(mn^{-1}(\log n)^3).
\]

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

\[
\text{ARE}(\hat{P}^{(q)}_{ij}, \tilde{P}^{(q)}_{ij}) = 0.
\]
CHAPTER 4. ROBUST ESTIMATION FROM MULTIPLE GRAPHS

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

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

Theorem 4.7.32 For sufficiently large values of \( \{C_{ij}\} \) and any \( 1 \leq i, j \leq n \), if \( m \to \infty \) at order \( m = O(n^b) \) for any \( b > 0 \), then the estimator based on ASE of MLE has smaller entry-wise asymptotic bias compared to the estimator based on ASE of MLE, i.e.

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

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

Theorem 4.7.33 For any \( 1 \leq i, j \leq n \), if \( m = O(n^b) \) for any \( b > 0 \), then

\[
\lim_{n \to \infty} \text{Var}(\tilde{P}^{(1)}_{ij}) = \lim_{n \to \infty} \text{Var}(\tilde{P}^{(q)}_{ij}) = 0.
\]

Proof: Direct result from Theorem 4.3.3 and Theorem 4.3.4.

4.7.7 Other Proofs

Lemma 4.7.34 Let \( A_{ij} \sim (1 - \epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}} \) with \( f \) to be Poisson, then \( E[(A_{ij} - E[\hat{P}^{(1)}_{ij}])^k] \leq \text{const} \cdot k! \), where \( \hat{P}^{(1)} \) 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 $\theta$ is bounded by

$$E[(X - \theta)^k] \leq k!(E[e^{X-\theta}] + E[e^{\theta-X}])$$

$$= k!(e^{-\theta}E[e^X] + e^\theta E[e^{-X}])$$

$$= k! \left( e^{\theta(e-2)} + e^{\theta e^{-1}} \right).$$

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

\[ E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \]

\[ = (1 - \epsilon) E[(X_1 - P_{ij} + P_{ij} - E[\hat{P}_{ij}^{(1)}])] + \epsilon E[(X_2 - C_{ij} + C_{ij} - E[\hat{P}_{ij}^{(1)}])] \]

\[ = (1 - \epsilon) \sum_{j=0}^{k} \binom{k}{j} (P_{ij} - E[\hat{P}_{ij}^{(1)}])^{k-j} E[(X_1 - P_{ij})^j] \]

\[ + \epsilon \sum_{j=0}^{k} \binom{k}{j} (C_{ij} - E[\hat{P}_{ij}^{(1)}])^{k-j} E[(X_2 - C_{ij})^j] \]

\[ \leq (1 - \epsilon) \sum_{j=0}^{k} \binom{k}{j} (P_{ij} - E[\hat{P}_{ij}^{(1)}])^{k-j} \cdot j! \cdot \text{const} \]

\[ + \epsilon \sum_{j=0}^{k} \binom{k}{j} (C_{ij} - E[\hat{P}_{ij}^{(1)}])^{k-j} \cdot j! \cdot \text{const} \]

\[ \leq (1 - \epsilon) k! \cdot \text{const}^k + \epsilon k! \cdot \text{const}^k \]

\[ \leq \text{const}^k \cdot k!. \]
Chapter 5

Discussion

When estimating the mean of a collection of unweighted graphs, motivated by the RDPG model, our methodology takes advantage of the low-rank structure of the graphs by applying low-rank approximation to the entry-wise MLE. We give a closed form for the asymptotic relative efficiency between the entry-wise MLE $\bar{A}$ and our estimator $\hat{P}$ in the case of a stochastic blockmodel, demonstrating that when the number of vertices $n$ is sufficiently large, low-rank methods provide a substantial improvement. In particular, we show that for a stochastic blockmodel with fixed number of blocks $K$, block size proportion $\rho$, and number of graphs $m$, the low-rank estimator $\hat{P}$ has MSE which is on the order of $n$ times lower than the MSE for $\bar{A}$.

Moreover, our estimator outperforms the entry-wise MLE in a cross validation analysis of the SWU4 brain graphs and in low- and full-rank simulation settings. These results illustrate that $\hat{P}$ performs well even when the low-rank assumption is
violated and that $\hat{P}$ is robust and can be applied in practice.

One of the key observations from our real data analysis was that the largest improvements using the low-rank method occurred when the number of graphs $m$ was small, and that it provided only minor improvements or even degraded performance slightly when $m$ was large. However, even in large scale studies the low-rank methods will be useful for estimating graph means for subpopulations, e.g. the population of females over 60 with some college education. Using the element-wise sample mean for such small strata, which may have fewer than ten subjects, will frequently result in a degradation of performance. Similarly, Durante et al. [2014] used low-rank deviations from a full rank population mean to model collections of graphs and our methods could be easily adapted to those ideas.

While the low-rank methods considered in this work will often offer substantial improvements, further refinements of these methods which account for the particular traits of connectomics data would be useful to improve estimation further. For example, an issue that arose in our analysis of the connectome dataset was the presence of structural ones in the mean graph for the population. These structural ones appear since edges between certain regions of the brain are present in all or nearly all members of the healthy population. The low-rank methods tend to miss these always-present edges while the sample mean will always capture them. Detecting and incorporating structural ones and zeros could yield methods that share the best elements of both methods considered here.
CHAPTER 5. DISCUSSION

For the SWU4 dataset, we used a cross-validation framework where we compared the estimates based on a subsample to the mean for the held-out set. Another option would be to compare the estimates $\hat{A}$ and $\hat{P}$ to the mean for the entire population including the subsample. Both of these analyses lead to very similar results in the cases presented above, but for various reasons one may prefer one analysis over another. The cross-validation method is most reasonable from a prediction perspective where prediction about new samples is of interest. If instead one is interested in learning directly about the mean of a population, especially a finite population, the sub-sampling approach may be the most logical choice.

In practice there is noise in the observed graphs and one may seek to account for this noise with more robust methods. Thus we consider an edge weight gross error model and propose an estimator based on the robust estimation followed by a low-rank decomposition. Under appropriate conditions, theoretical results show that our estimator not only inherits the robust property from robust estimators, but also wins the bias-variance tradeoff by exploiting the low-rank graph structure under appropriate conditions.

In Section 4.3, we present theory based on the exponential distribution with MLqE for clarity. Section 4.4 indicates that these results can be extended to other distributions and robust estimators. Note that the most important condition is Condition 1, which requires that the MLE under the corresponding edge weight and contamination distribution is concentrated so that we obtain the required matrix bounds. This
generalization makes the theory more flexible and powerful.

In this work, our theoretical analysis is performed mostly in the weighted stochastic blockmodel setting. Note that the results can be extended to the weighted random dot product graph, i.e. our estimator does not require the block structure. Indeed, the WSBM assumption is just to ensure \( \text{rank}(E[\hat{P}(q)]) \) has an upper bound under the contamination model that is invariant in the number of vertices. With this assumption on the rank, all the theory still holds in the WRDPG setting. In practice, graphs are not exactly low rank. However, as shown in Figure 4.5 and Figure 4.6, our estimator still provides large improvement with approximate low-rank structure. Thus our method can be applied to a much more general setting instead of being restricted to WSBM.

Selecting a good distortion parameter \( q \) based on real data in MLqE is an important but difficult task. Qin and Priebe [2017] presents a thorough, successful, and decidedly non-trivial example of such a selection methodology in the context of hypothesis testing in the one-sample univariate location problem. While we use a fixed \( q = 0.9 \) in our real data experiments without presenting a formal automatic selection methodology, Figure 4.4 demonstrates that a poor choice of \( q \) can significantly degrade performance. We suggest that a program to develop an adaptive approach in our setting, perhaps along the lines of Qin and Priebe [2017], is a promising avenue for widening the applicability of our estimator.

The models considered in this work assume that the vertex correspondence across
graphs is known. In some applications this may not be the case. Our methods may still be applicable after applying graph matching algorithms such as [Lyzinski et al., 2014, 2015, 2016, Vogelstein et al., 2015].

In general, improvement in estimation performance is important not only for the estimation itself but also is important for subsequent statistical inference procedures such as clustering, vertex classification, etc. For example, [Priebe et al., 2015] and [Chen et al., 2016] both discuss vertex classification based on a single unweighted graph with contamination. Additional investigation into subsequent inference tasks based on multiple contaminated weighted graphs should lead to important refinements and extensions of our robust estimation method.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


Vince Lyzinski, Sancar Adali, Joshua T Vogelstein, Youngser Park, and Carey E
BIBLIOGRAPHY


Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Lapla-
BIBLIOGRAPHY


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.