Outline for LLG

May 16, 2016

1 Introduction

- Mean is one of the most important and basic concept in statistics. Motivated by the law of large numbers, sample mean is always considered to be the best estimate of the population mean. As a result, nowadays we take averages almost everywhere, from the fundamental elements in Euclidean space to more general objects, like shapes, documents and graphs.

  However, contradict to the general intuition, arithmetic average should not be our first choice all the time. In 1955, Stein's paradox shows the inadmissibility of the sample mean when there are more than three normal distributions. The fact that James-Stein estimator dominates the sample mean makes it less preferable to take the average in that situation. 27 years later, Gutmann proved that this cannot occur when the sample spaces are finite. But even when sample mean is admissible, it doesn't close the door of other estimators to be better in some cases. So in a specific situation, for instance in this paper a collection of graphs is considered, there is always a chance to have a better estimator compared to the sample mean.

- The mean of a collection of graphs can be defined in various ways. One natural definition is the proportion of the existence of an edge between any pair of vertices. Estimating the mean of a population of graphs based on a sample is becoming more and more important both in statistical inference and in various applications like connectomics, social networks, etc.

- Element-wise maximum likelihood estimate, which happens to be the sample mean in many situations, is a reasonable estimator if we only consider the independent edge graph model without taking any graph structure into account. However, it does not perform very well especially when we have a few observations, which is likely the case in real world.

- (Challenges) Generally, we don't have any information about the structure of the graphs. So it is hard to take advantage of the unknown graph structure.

- One of the most important structures is the community structure in which vertices are clustered into different communities such that vertices of the same community behave similarly. The stochastic blockmodel (SBM) captures such structural property and is widely used in modeling networks.

- Meanwhile, the latent positions model (LPM), a much more general model compared to SBM, proposes a way to parameterize the graph structure by latent positions associated with each vertex. However, the random dot product graph (RDPG) which is a special case of LPM stays in between and motivates our estimator. In this paper, we analyze our estimator in terms of RDPG specifically.
- Using the estimates of the latent positions in an RDPG setting based on a truncated eigen-decomposition of the adjacency matrix, we consider a new estimator for the mean of the collection of graphs which captures the low-rank structure. And we prove via both theory and simulations/real data analysis that it is better than element-wise MLE.
- (Future work) Robust estimation, dimension selection, diagonal augmentation, etc.

2 Models and Estimators

2.1 Independent Edge Model

- Under the independent edge model (IEM), each edge A_{ij} independently follows the Bernoulli distribution with parameter P_{ij}.

2.2 Estimator $\bar{A}$

- Under the IEM, element-wise mean of the adjacency matrices $\bar{A}$ is the MLE as well as the least squared estimate.
- $\bar{A}$ is unbiased for IEM and has entry-wise variance $\text{Var}(\bar{A}_{ij}) = P_{ij}(1 - P_{ij})/M$. And it is the UMVUE under IEG with no constraints. But, it doesn't exploit any structure.

2.3 Random Dot Product Graph

- Hoff et. al. (2002) proposed a model for random graphs called Latent Positions Graph Model.
- A specific instance of this model that we will examine is the random dot product graph model (RDPG) in which the link function is the dot product, i.e. the probability of an edge being present between two nodes is the dot product of their latent vectors.

2.4 Estimator $\hat{P}$ Based on Adjacency Spectral Embedding

- In order to take advantage of the underlying low rank structure of the RDPG, we use the adjacency spectral embedding (ASE) studied by Sussman et. al. to enforce a low rank approximation on the entry-wise mean matrix $\bar{A}$, which will decrease the variance without losing much in bias if we embed it into the right dimension.

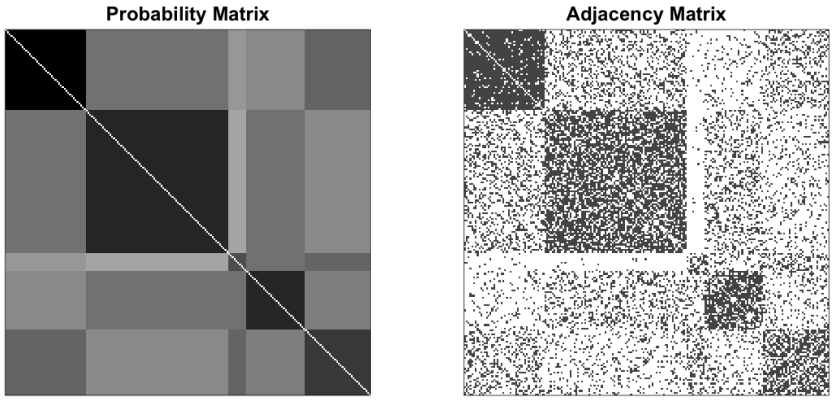

Figure 1: Example illustrating the SBM. The figure on the left is the probability matrix P that follows a SBM with $K = 5$ blocks; The other figure shows the adjacency matrix A for 200 vertices generated from the SBM with probability matrix P.

- There are varies ways dealing with dimension selection. In this paper, we consider Zhu and Ghodsi's elbow selection method and the universal singular value thresholding (USVT) method. Details are discussed in Section 5.1.
- Detailed description of the algorithm for our estimator $\hat{P}$.

2.5 Stochastic Block Model as a Random Dot Product Graph

- One of the most important structures is the community structure in which vertices are clustered into different communities such that vertices of the same community behave similarly. Such structural property is captured by the SBM, 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.
- Formally, the SBM is determined by the number of blocks K (generally way less than the number of vertices N), block proportion vector ρ, and block probability matrix B.
- Now if we consider the SBM as a random dot product graph, all vertices in the same block would have identical latent positions.

3 Results

3.1 Theoretical Results

- To compare the performance between $\hat{P}$ and $\bar{A}$, we examine the relative efficiency (RE), in mean squared error (MSE), among the two defined as: $RE_{ij} = \frac{MSE(\hat{P}_{ij})}{MSE(\bar{A}_{ij})}$.

- **Theorem 3.1** *For any i and j, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$, we have*

$$\mathrm{ARE}(\bar{A}_{ij}, \hat{P}_{ij}) = 0.$$

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

$$\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{N}.$$

- **Lemma 3.2** *In the same setting as above, for any i, j, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$, we have*

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

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

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

3.2 Validation with Simulations

- We demonstrate the theoretical results in Section 3.1, the relative efficiency of $\hat{P}$, via various Monte Carlo simulation experiments.

3.2.1 Simulation Setting

- Here we consider the 2-block SBM parameterized by

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

- And we embed the graphs into the dimension $d = \mathrm{rank}(B) = 2$.

3.2.2 Simulation Results

- Figure 2 plots the scaled RE in average with different N and fixed M based on 1000 Monte Carlo replicates. Different types of dashed lines denote the simulated scaled RE associated with the edges we are averaging over. Solid line represents the theoretical value for scaled RE. From the figure, we see that $N \cdot \mathrm{RE}_{st}(\bar{A}, \hat{P})$ converges to $1/\rho_s + 1/\rho_t$ represented as the black solid line, as suggested in Lemma 3.1. Notice that this means $\mathrm{RE}_{st}(\bar{A}, \hat{P})$ is decreasing at rate $1/N$.

- To verify Theorem 3.1 holds with different ρ, Figure 3 shows the scaled RE with $N = 500$ and $M = 100$ while changing ρ_1 from 0.1 to 0.9. The simulated values agree with the theoretical values perfectly.

- By checking the RE of the two estimates $\hat{P}$ and $\bar{A}$ over 1000 Monte Carlo replicates, we demonstrate that the theoretical results in Section 3.1.

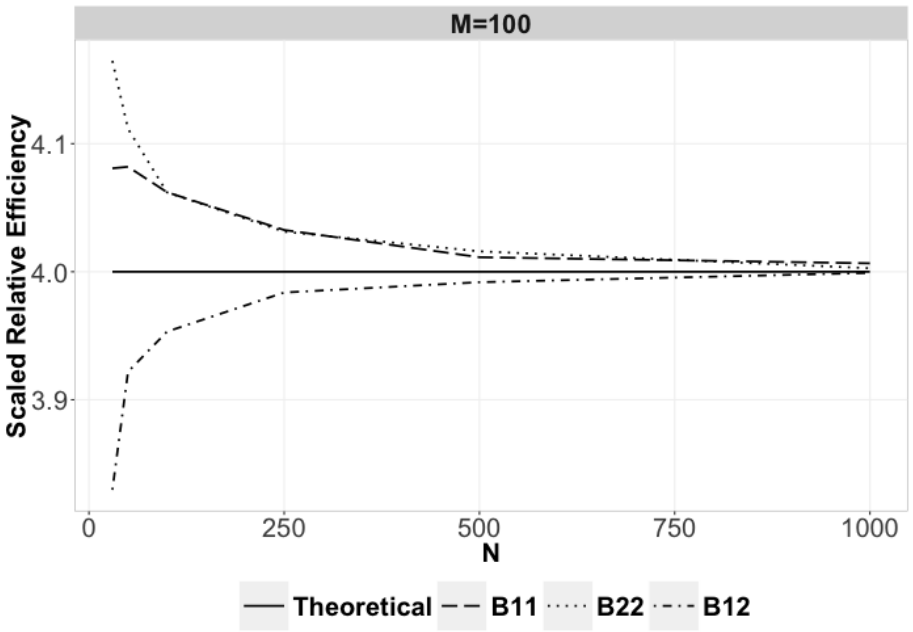

Figure 2: Scaled relative efficiency in average with different N and fixed M based on 1000 Monte Carlo replicates. Different types of dashed lines denote the simulated scaled RE associated with the edges we are averaging over. Solid line represents the theoretical value for scaled RE. Observe that $N \cdot \mathrm{RE}_{st}(\bar{A}, \hat{P})$ converges to $1/\rho_s + 1/\rho_t$ as expected.

3.3 CoRR Brain Graphs: Cross-Validation

- In practice, the graphs may not perfectly follow an RDPG, or even not IEM. But we are still interested in the mean graph. To demonstrate that the $\hat{P}$ estimate is still valid in such cases, we examine three datasets, JHU, desikan and CPAC200, which are sets of 454 brain connectomes with different number of nodes generated from fMRI scans available at the Consortium for Reliability and Reproducibility (CoRR).
- To compare $\bar{A}$ and $\hat{P}$ we perform a cross-validation study to examine the impact of the number of available graphs M.
- Figure 4 demonstrates that our algorithm gives a better estimate $\hat{P}$ according to all three datasets.
- When M is small, $\bar{A}$ has large variance which leads to large MSE. Meanwhile, $\hat{P}$ reduces the variance by taking advantages of the graph structure and outperforms $\bar{A}$ dramatically.
- Moreover, Zhu and Ghodsi's algorithm and USVT algorithm both do a good job for selecting the dimension to embed.
- Simulation with P being the mean graph of the real data shows that $\hat{P}$ still does a good job when the low rank assumption is violated.

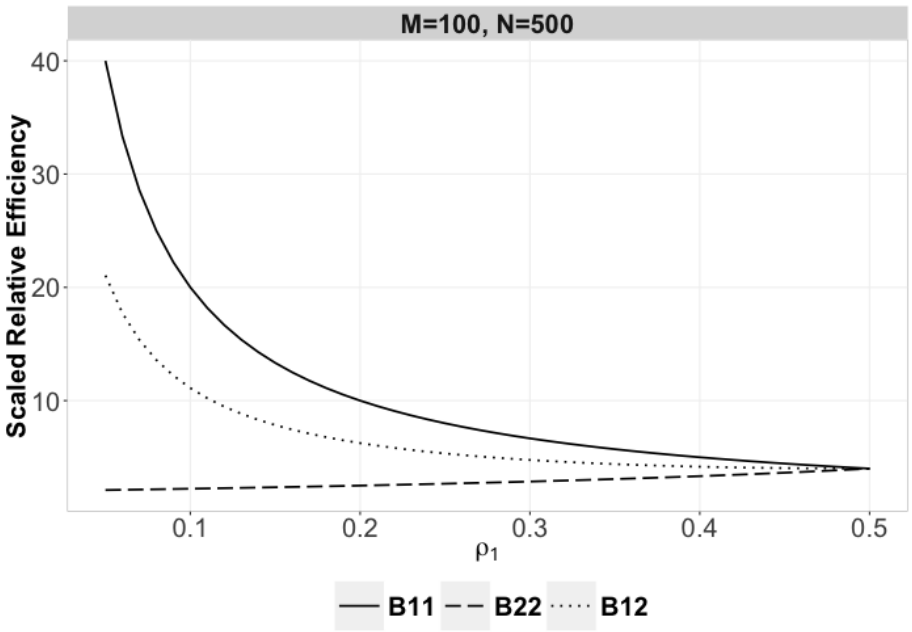

Figure 3: Simulated results for scaled RE, i.e. $N \cdot \mathrm{RE}_{st}(\bar{A}, \hat{P})$ with $N = 500$ and $M = 100$ of 1000 Monte Carlo replicates while changing ρ_1 from 0.1 to 0.9. Scaled relative efficiency in average with different N and fixed M based on 1000 Monte Carlo replicates. Different types of lines denote the simulated values associated with the edges we are averaging over. Notice that when $\rho_1 = 0.5$, the scaled RE has value 4.0, which agrees with the result in Figure 2 as expected.

3.4 Simulation under the Full Rank Independent Edge Model

- While the theory we have is based on the low rank assumption, $\hat{P}$ sometimes wins the bias-variance tradeoff even when the graphs have full rank structure. To illustrate this point, instead of the low rank SBM, we run simulations under the full rank independent edge model with the probability matrix P to be the sample mean of the 454 graphs in the desikan dataset.

- Figure 8 compare the MSE between $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data based on the full rank probability matrix P as the sample mean in desikan dataset while embedding the graphs into different dimensions with different size M of the subsamples. Vertical intervals represent the 95% confidence interval. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including what Zhu and Ghodsi selects.

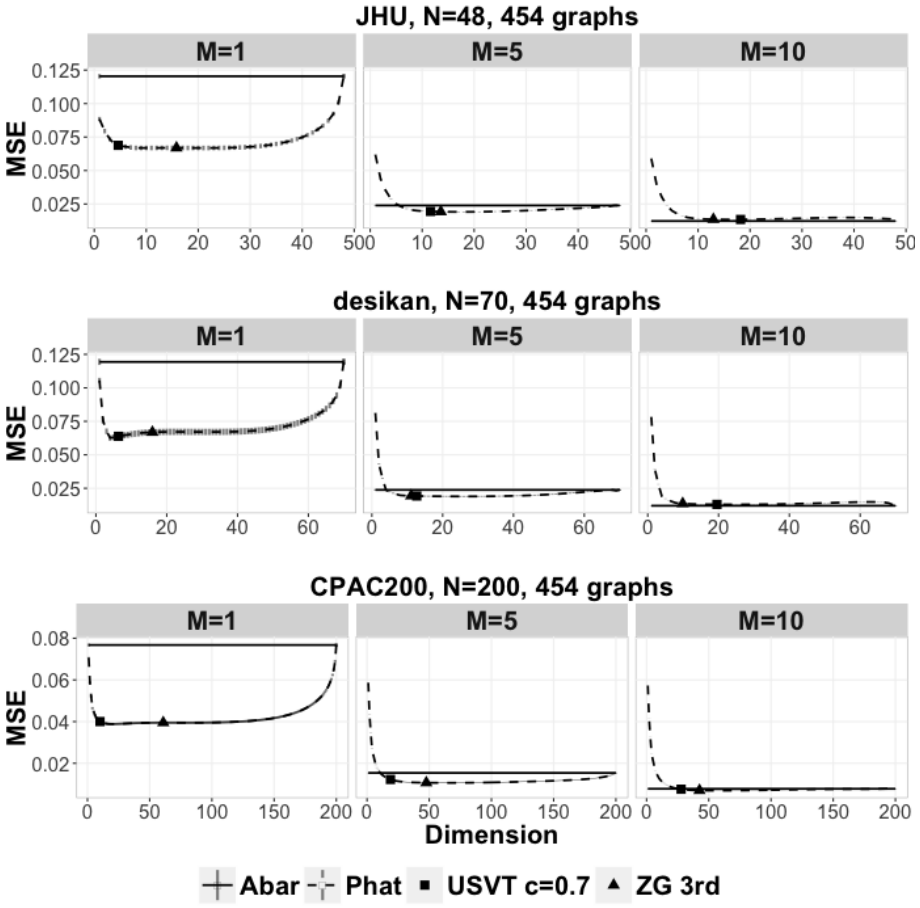

Figure 4: Comparison of MSE between $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for three dataset (JHU, desikan and CPAC200) while embedding the graphs into different dimensions with different size M of the subsamples. The dimension chosen by the 3rd elbow of Zhu and Ghodsi is denoted in triangle, and chosen by USVT with threshold equals 0.7 is denoted in square. Vertical intervals represent the 95% confidence interval. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including what Zhu and Ghodsi selects.

4 Discussion

4.1 Summary

- In this paper, we propose a better way to estimate the mean of a collection of graphs by taking advantage of the low rank structure of the graphs.

4.2 Future Work

- Generally the observations we have are always contaminated in practice. In this case, improved robust estimator based on the low rank structure

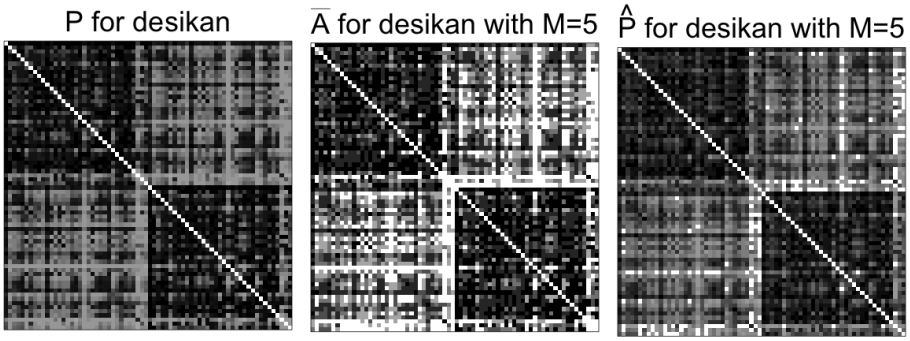

Figure 5: Comparison between the mean of 454 graphs P and two estimates $\bar{A}$ and $\hat{P}$ derived from a sample of size $M = 5$ from desikan dataset while embedding the graphs into dimension $d = 11$ selected by the 3rd elbow of ZG method. From the figure, we can see that $\hat{P}$ is a better estimation of P than $\bar{A}$.

of the graphs is preferred.

- Estimating the rank of the graph structure accurately will certainly help improve the results.
- In this paper, we are using Scheinerman's method with 1 iteration for diagonal augmentation.

5 Methods

5.1 Choosing Dimension

- Often in dimensionality reduction techniques, the choice for dimension, d, relies on visually analyzing a plot of the ordered eigenvalues, looking for a "gap" or "elbow" in the scree-plot.
- USVT is a simple estimation procedure that works for any matrix that has "a little bit of structure".

5.2 Graph Diagonal Augmentation

- The graphs examined in this work are hollow, in that there are no self-loops and thus the diagonal entries of the adjacency matrix are 0. This leads to a bias in the calculation of the eigenvectors.
- We minimize this bias by using an iterative method developed by Scheinerman and Tucker.

5.3 Source Code

5.4 Dataset Description

- The original dataset is from the Emotion and Creativity One Year Retest Dataset provided by Qiu, Zhang and Wei from Southwest University. It is

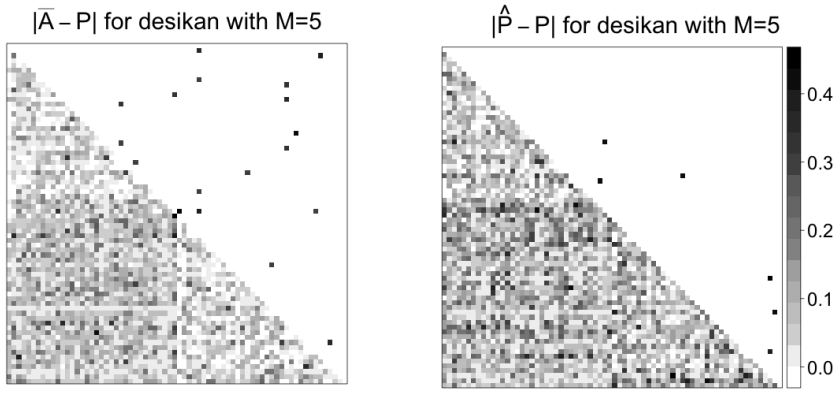

Figure 6: Heat plot of the absolute estimation error $|\bar{A} - P|$ and $|\hat{P} - P|$ for a sample of size $M = 5$ from desikan dataset while embedding the graphs into dimension $d = 11$ selected by the 3rd elbow of ZG method. The lower triangular matrix shows the actual abstolute difference, while the upper triangular matrix only highlights the edges with absolute differences larger than 0.4. The fact that 18 edges from $\bar{A}$ and 6 edges from $\hat{P}$ being highlighted shows the better performance of $\hat{P}$.

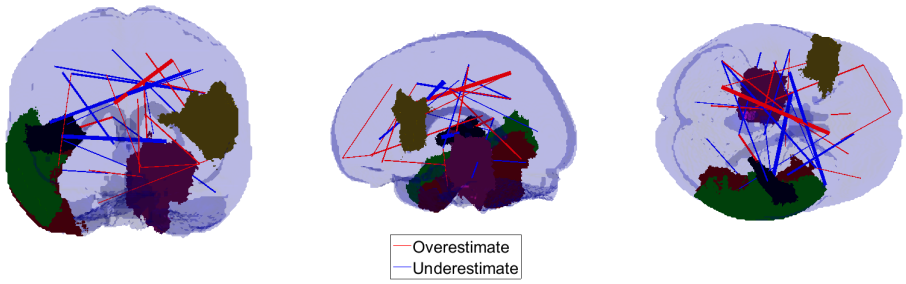

Figure 7: Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with largest difference $|\bar{A} - P| - |\hat{P} - P|$. Red indicates that $\hat{P}$ overestimate P while blue means that $\hat{P}$ underestimate P. The edge width is determined by the 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 $\bar{A}$ mostly for estimate P.

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 the incomplete data, the true number of scans is 454.

- When deriving MR connectomes, the NeuroData team parcellate the brain into groups of nodes as defined by anatomical atlases. The atlases are defined either physiologically or structurally by neuroanatomists (Desikan and JHU), or are generated using a segmentation algorithm looking for certain features or groupings (CPAC200).

- The graphs we are using are processed by NeuroData team from DTI data of the original dataset generated with different atlases (desikan, JHU and

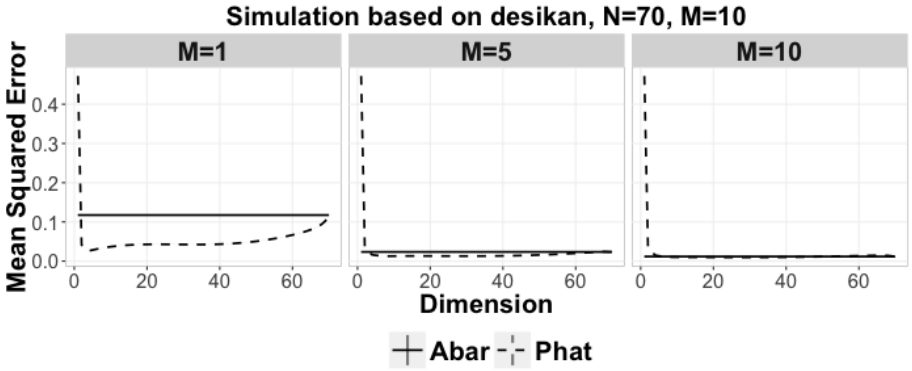

Figure 8: Comparison of MSE between $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data based on the full rank probability matrix P as the sample mean in desikan dataset while embedding the graphs into different dimensions with different size M of the subsamples. Vertical intervals represent the 95% confidence interval. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including what Zhu and Ghodsi selects.

CPAC200), each containing different region/node definitions.

- The graphs are undirected, unweighted and with no self-loops. An edge exists between two regions when there is at least one white-matter tract connecting the corresponding two parts of the brain.

5.5 Outline for the Proof of the Theorems