Outline for Robust LLG

November 20, 2016

1 Introduction

- Network analysis is becoming more and more widely used recently. And estimating the mean of a population based on the sample is one of the most important tasks. 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 expected value of the edge weight between any pair of vertices when graphs are sampled i.i.d. from some random graph distribution.

- 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.

- 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.

- Using the estimates of the latent positions in an RDPG setting based on a truncated eigen-decomposition of the adjacency matrix, LLG consider a new estimator for the mean of the collection of graphs which captures the low-rank structure. And they prove that when the graphs follow a Bernoulli distribution, the estimator is better than element-wise MLE.
- In LLG, it is assumed that the adjacency matrix is observed without contamination, however in practice there will be noise in the observed graph.
- We first proved that the estimator based on ASE proposed in LLG is still better than entry-wise MLE when the observations are contaminated under proper conditions.
- Also, with contaminations, it is natural to use more robust methods, like MLqE considered in this paper. And we proved that robust estimators improve the performance with relative large noise.
- Similarly, in order to take advantage of the low rank structure, we enforce a low rank approximation on the entry-wise MLqE. And we prove that, under proper assumptions, the new estimator inherits the robust property from MLqE and wins the bias-variance tradeoff by considering the low rank structure.
- The estimation of the mean graph is considered in this paper but not restricted to, since the method can be applied to estimate the parameters of any one-parameter exponential family.

2 Models

- For this work, we are in the scenario that m graphs are given in the adjacency matrices form.
- In this section, we present three nested models, the weighted independent edge model, the weighted random dot product model, and the stochastic blockmodel. Moreover, we introduce a contaminated model based on them.

2.1 Weighted Independent Edge Model

- We first consider the weighted independent edge model (WIEM).

2.2 Weighted Random Dot Product Graph

- The latent positions model proposed by Hoff et. al. (2002) captures the unobserved properties of each vertex. And here we generalize the idea and define the weighted latent positions model as following.
- A specific instance of this model is the weighted random dot product graph model (WRDPG) in which the link function is the dot product.

2.3 Stochastic Block Model as a Weighted 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 weighted random dot product graph, all vertices in the same block would have identical latent positions.

2.4 Stochastic Block Model as a Weighted Random Dot Product Graph with Contaminations

- In practice, we always get noisy data. Thus to better describe the real world, a contaminated model is needed.
- Here we define the edge contaminated model as following.

3 Estimators

3.1 Entry-wise Maximum Likelihood Estimator $\hat{P}^{(1)}$

- Under the WIEM, the most natural estimator is the element-wise MLE $\hat{P}^{(1)}$ based on the adjacency matrices $A^{(1)}, \cdots, A^{(m)}$.
- In many cases, the entry-wise MLE happens to be the mean graph $\bar{A}$, which is the UMVUE under WIEG with no constraints. But, it doesn't exploit any graph structure.

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

- In order to take advantage of the underlying low rank structure of the WRDPG, we use the adjacency spectral embedding (ASE) studied by Sussman et. al. to enforce a low rank approximation on the entry-wise MLE matrix $\hat{P}^{(1)}$, which will decrease the variance without losing much in bias if we embed it into the right dimension.
- There are varies ways dealing with dimension selection. In this paper, we consider Zhu and Ghodsi's elbow selection method.
- Detailed description of the algorithm for our estimator $\widetilde{P}^{(1)}$.

3.3 Entry-wise Maximum Lq Likelihood Estimator $\hat{P}^{(q)}$

- The MLE is asymptotically efficient, i.e. when sample size is large enough, the MLE is at least as accurate as any other estimator. However, when the sample size is moderate, robust estimators always outperforms MLE in terms of mean squared error by winning the bias-variance tradeoff.
- Moreover, in practice, the observations are generally contaminated. In this case, robust estimators can even beat MLE asymptotically.
- The class of parametric estimators based on the q-entropy function, the MLqE, is considered in this paper.

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

- Intuitively, the low rank structure of the WRDPG should be preserved more or less in the entry-wise MLqE.
- Similarly, in order to take advantage of such low rank structure, we apply the same idea for building $\widetilde{P}^{(1)}$, i.e. enforce a low rank approximation on the entry-wise MLqE matrix $\hat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$.
- Detailed description of the algorithm for our estimator $\widetilde{P}^{(q)}$.

4 Theoretical Results

In this section, for illustrative purpose, we are going to present theoretical results when the contaminated model is based on exponential distributions, i.e. $\mathcal{F} = \{f_\theta(x) = \frac{1}{\theta}e^{-x/\theta}, \theta \in [0, R] \subset \mathbb{R}\}$, where $R > 0$ is a constant. The results can be extended to a general situation with proper assumptions, which will be discussed in Section 5.

4.1 $\hat{P}^{(q)}$ vs $\hat{P}^{(1)}$

Lemma 4.1 *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

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

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

Lemma 4.2

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

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

4.2 $\widetilde{P}^{(1)}$ vs $\hat{P}^{(1)}$

Corollary 4.3 *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLE has the same entry-wise asymptotic bias as MLE, i.e.*

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

Theorem 4.4 *Assuming that $m = O(n^b)$ for any $b > 0$, then $\text{Var}((\hat{Z}_i^T \hat{Z}_j)_{\text{tr}}) = O(m^{-1}n^{-2}(\log n)^6)$.*

Theorem 4.5 *For fixed m, $1 \le i, j \le n$ and $i \ne j$,*

$$\frac{\text{Var}(\widetilde{P}_{ij}^{(1)})}{\text{Var}(\hat{P}_{ij}^{(1)})} = O(n^{-2}(\log n)^6).$$

Thus

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

Furthermore, as long as m goes to infinity of order $O(n^b)$ for any $b > 0$,

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

4.3 $\widetilde{P}^{(q)}$ vs $\hat{P}^{(q)}$

Corollary 4.6 *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has the same entry-wise asymptotic bias as MLqE, i.e.*

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

Theorem 4.7 *Assuming that $m = O(n^b)$ for any $b > 0$, then $\text{Var}((\hat{Z}_i^T \hat{Z}_j)_{\text{tr}}) = O(n^{-2}(\log n)^6)$.*

Theorem 4.8 *For fixed m, $1 \le i, j \le n$,*

$$\frac{\text{Var}(\widetilde{P}_{ij}^{(q)})}{\text{Var}(\hat{P}_{ij}^{(q)})} = O(mn^{-2}(\log n)^6).$$

Thus

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

Furthermore, as long as m goes to infinity of order $o(n^2(\log n)^{-6})$,

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

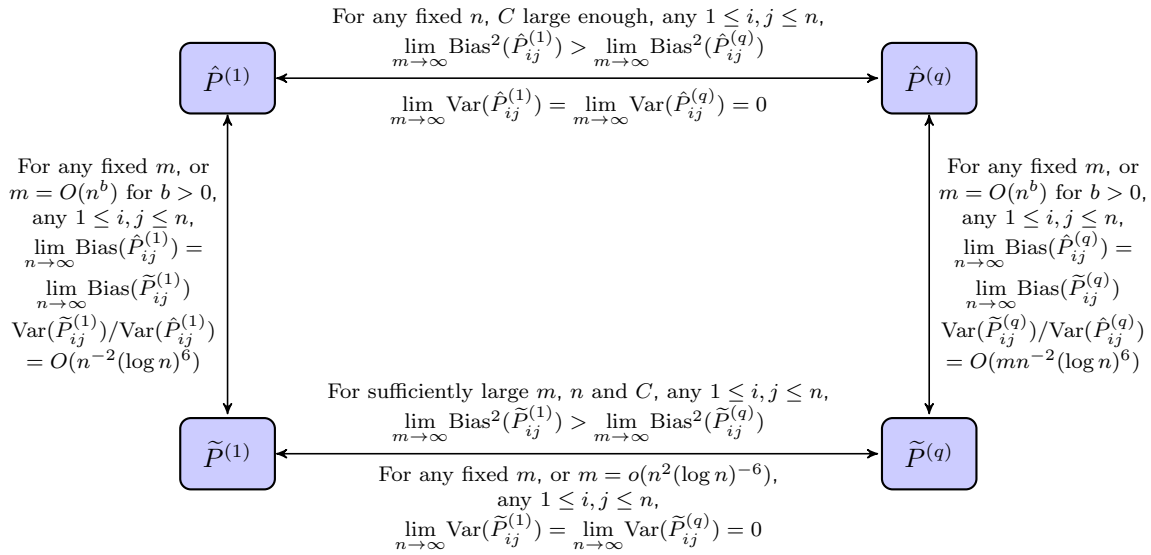


Figure 1: Relationship between four estimators.

4.4 $\widetilde{P}^{(q)}$ vs $\widetilde{P}^{(1)}$

Theorem 4.9 *For sufficiently large n and C, any $1 \le i,j \le n$,*

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

Theorem 4.10 *For any fixed m, any $1 \le i,j \le n$,*

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

Furthermore, as long as m goes to infinity of order $o(n^2(\log n)^{-6})$, any $1 \le i,j \le n$,

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

4.5 Summary

In summary, we plot the relationship among four estimators in Figure 1.

5 Extensions

We can generalize both the exponential distribution and the entry-wise MLqE to an arbitrary F and any entry-wise estimator $\hat{P}$ with the following assumptions:

- Let $A \overset{iid}{\sim} (1-\epsilon)F(P) + \epsilon F(C)$ and $H_{ij}^{(1)} = E[\hat{P}_{ij}^{(1)}] = (1-\epsilon)E_F(P_{ij}) + \epsilon E_F(C_{ij})$, then $E[(A_{ij} - H_{ij}^{(1)})^k] \le \text{const} \cdot k!$.

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

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

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

- $\hat{P}_{ij} \le \text{const} \cdot \hat{P}_{ij}^{(1)}$. This might be generalized to with high probability later.

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

6 Empirical Results

6.1 Simulation Results

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

6.1.1 Simulation Setting

- Here we consider the 2-block SBM parameterized by

$$B = \begin{bmatrix} 4.2 & 2 \\ 2 & 7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

- The contamination is also a 2-block SBM parameterized by

$$B = \begin{bmatrix} 20 & 18 \\ 18 & 25 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

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

6.1.2 Simulation Results

- Figure 2 plots the mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 10$ based on 1000 Monte Carlo replicates. And we use $q = 0.8$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. **1. MLE $\hat{P}^{(1)}$ vs MLqE $\hat{P}^{(q)}$:** MLE outperforms a little bit when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination increases; **2. MLE $\hat{P}^{(1)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** ASE procedure takes the low rank structure into account and $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff; **3. MLqE $\hat{P}^{(q)}$ vs ASE o MLqE $\widetilde{P}^{(q)}$:** MLqE preserves the low rank structure of the original graph more or less, so ASE procedure still helps and $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff; **4. ASE o MLqE $\widetilde{P}^{(q)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

- Figure 3 show the mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 10$ and $\epsilon = 0.2$ based on 1000 Monte Carlo replicates. Different colors represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators; 2. MLqE shows the robustness property compare to the MLE. And as q goes to 1, MLqE goes to the MLE as expected.

- By comparing the performance of the four estimators based on different setting, we demonstrate the theoretical results in Section 4.

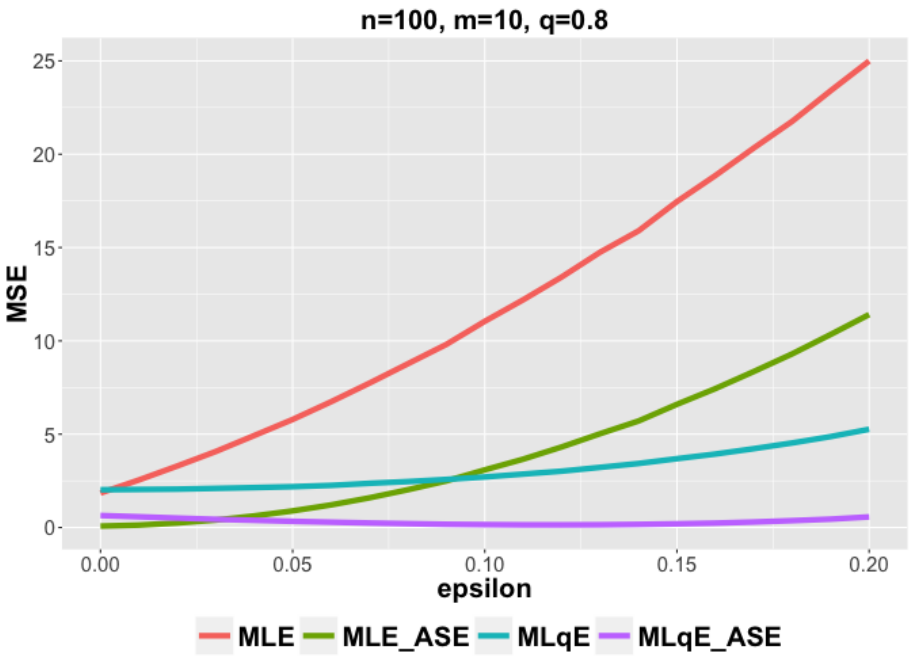


Figure 2: Mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 10$ based on 1000 Monte Carlo replicates. And we use $q = 0.8$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. **1. MLE $\hat{P}^{(1)}$ vs MLqE $\hat{P}^{(q)}$:** MLE outperforms a little bit when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination increases; **2. MLE $\hat{P}^{(1)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** ASE procedure takes the low rank structure into account and $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff; **3. MLqE $\hat{P}^{(q)}$ vs ASE o MLqE $\widetilde{P}^{(q)}$:** MLqE preserves the low rank structure of the original graph more or less, so ASE procedure still helps and $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff; **4. ASE o MLqE $\widetilde{P}^{(q)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

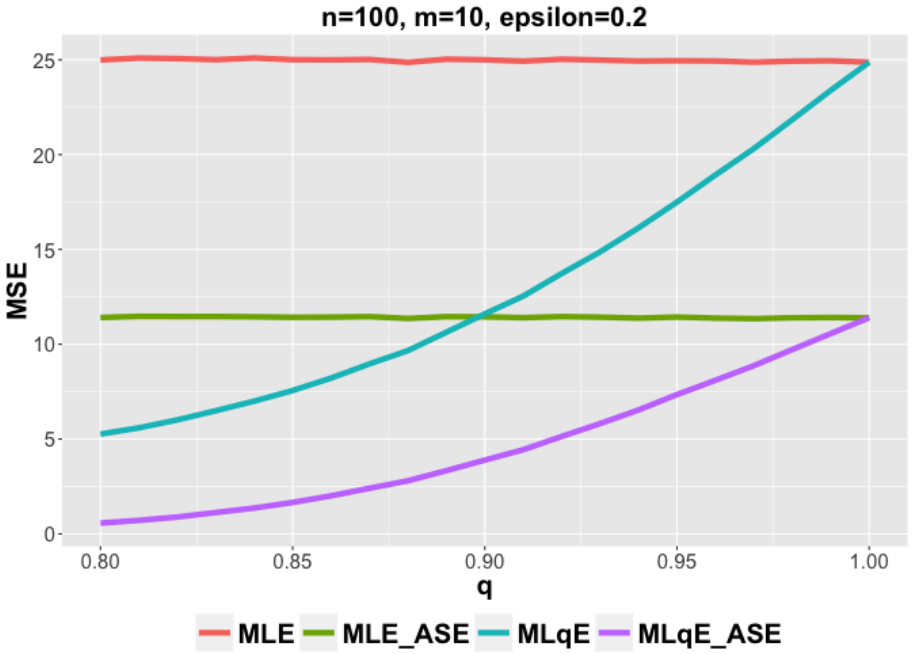
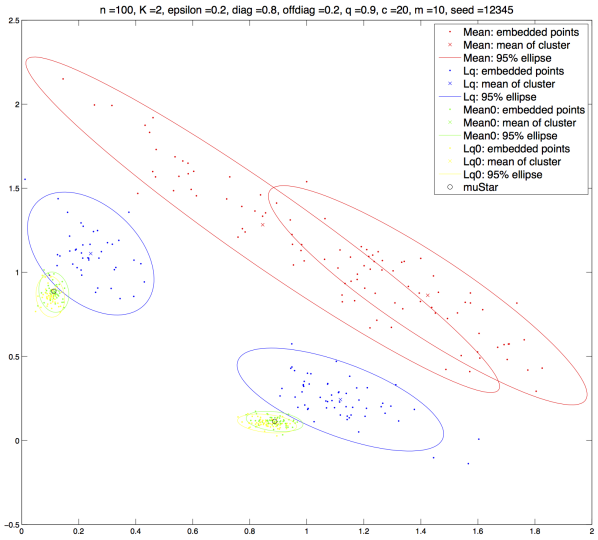


Figure 3: Mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 10$ and $\epsilon = 0.2$ based on 1000 Monte Carlo replicates. Different colors represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators; 2. MLqE shows the robustness property compare to the MLE. And as q goes to 1, MLqE goes to the MLE as expected.

6.2 CoRR Graphs

- In practice, the graphs may not perfectly follow an RDPG, or even not IEM. But we are still interested in the discussed approach. To demonstrate that the estimates are still valid in such cases, we examine the datasets, CPAC200, which is a set of 454 brain connectomes with different number of nodes generated from fMRI scans available at the Consortium for Reliability and Reproducibility (CoRR).
- The dataset has 454 different brain scans in the form of weighted, undirected graph with no self loop, based on the pipeline described in [2] and [1].
- To compare the four estimators, we perform a cross-validation study on 454 graphs.
- We run 1000 simulations on the dataset for each sample size $m = 1$, $m = 2$, $m = 5$. And we apply ASE for all possible dimensions, i.e. d ranges from 1 to n. The result is shown in Figure 4.
- Since it is real data, MLqE outperforms MLE because of the robustness



property. Moreover, as suggested in the previous theorems, such property is kept after the ASE procedure.

- When d is small, ASE procedure underestimates the dimension and fail to get important information, which leads to poor performance. In practice, we use algorithms like Zhu and Ghodsi's method to select the dimension d. We can see Zhu and Ghodsi's algorithm does a pretty good job for selecting the dimension to embed.
- When m is small, MLE and MLqE have large variances which lead to large MSE. Meanwhile, the ASE procedure reduces the variance by taking advantages of the graph structure.

7 Discussion

8 Appendix

All proofs here.

References

[1] Neurodata's mri to graphs pipeline. `http://m2g.io`. Accessed: 2016-05-23.

[2] Gregory Kiar. Gremlin: Graph estimation from mr images leading to inference in neuroscience. 2016.

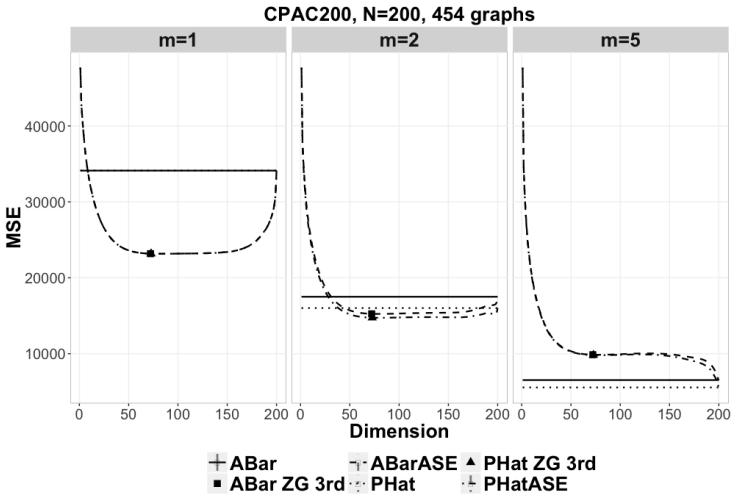


Figure 4: Comparison of mean squared error in average among four estimators while embedding the graphs into different dimensions with different size m of the subsamples. The dimensions chosen by the 3d elbow of Zhu and Ghodsi are denoted in triangle and square. When m is small, both robust estimation and ASE procedure help improving the performance, making $\widetilde{P}^{(q)}$ the best among four estimators.