Law of Large Graphs

August 24, 2015

1 Introduction

OCP/HCP/FCP connectome setting

Examples of $\bar{A}$ being used when vertex correspondence known

Looking for better estimators of “mean” graph by making assumption that many vertices will behave similarly
[1] [2] [3] [4]
[7]

2 Model

We present, with theory, a comparison of two estimators for the mean of a collection of graphs. This work considers the scenario of having M unweighted and undirected hollow graphs stored as adjacency matrices, A_m, each having N vertices with known correspondence. In this setting, we consider each graph to be a random graph, in that it is sampled from a common edge-wise probability matrix P, which we aim to estimate using our finite sample of graphs.

2.1 Maximum Likelihood Estimation

Maximum likelihood estimation is a common approach to parameter estimation and is the basis for using $\bar{A}$ when estimating the mean graph from a sample.

$$\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A_m \tag{1}$$

In this approach each element of the adjacency matrix A_{ij} is treated as an independent Bernoulli random variable with probability P_{ij}. Therefore, with each element treated independently, to estimate the mean graph P the maximum likelihood approach indicates that one should take the element-wise mean, $\bar{A}$.

2.2 Model Assumption: Stochastic Block Model

It is often useful when analyzing graphs, particularly in the case of large N, to make the assumption that many of the vertices do not behave independently and that there is a community structure where collections of vertices behave similarly in their connections. The model we use in this paper to depict this community structure the stochastic block model (SBM). The SBM is defined by the parameters k, ρ, and B. In this model each vertex is assigned to one of k communities with the probability of being assigned to the i th community designated as ρ_i. The properties of this community structure is stored in the symmetric $k \times k$ block matrix B, where B_{ij} represents the probability of an edge existing between a vertex of community i and one of community j.

(Add picture of a block model adjacency matrix for clarification)

2.3 Latent Position Model

A model for random graphs proposed by Hoff et. al. is the latent position model. In this model, each vertex has an associated latent vector, and the probability of a edge being present between two vertices is dependent only on their latent vectors. [?] A specific instance of this model that we will examine is the random dot product graph (RDPG) in which the probability of an edge being present between two nodes is the dot product of their latent vectors. [?]

2.4 Adjacency Spectral Embedding: $\hat{P}$

In order to expose the underlying block structure within a graph, Sussman et. al. proposed adjacency spectral embedding (ASE) to enforce a low rank$-k$ approximation on the adjacency matrix [?]. This embedding creates a RDPG representation of the adjacency matrix from its low rank eigen decomposition. The latent vectors are stored in the $N \times k$ matrix X, where the columns are comprised of the eigenvectors associated with the k largest eigenvalues, in absolute value, of the adjacency matrix. With this embedding, X, each row is then a latent vector for its corresponding vertex.

In this work, we extend ASE to embed the mean matrix $\bar{A}$, rather than the adjacency matrix alone. By making the assumption that there is an underlying SBM structure to graphs, enforcing this low rank approximation on $\bar{A}$ will provide a better estimate for the true mean matrix, P. We will refer to this new estimate as $\hat{P} = XX^T$. Details of this algorithm are presented in section 5.

2.5 Performance Evaluation: Relative Efficiency

To compare the performance between $\hat{P}$ and $\bar{A}$, we examine the relative efficiency (RE) among the two defined as:

$$RE_{ij} = \frac{Var(\hat{P}_{ij})}{Var(\bar{A}_{ij})} \tag{2}$$

This comparison is valid as long as both $\bar{A}$ and $\hat{P}$ are consistent estimators in that they are asymptotically unbiased. It is known that $\bar{A}$ is unbiased, being simply the mean of M Bernoulli trials. As for $\hat{P}$, we know it to be a low row approximation of $\bar{A}$ and thus its expectation is $\bar{A}$, which is an unbiased estimator for P. Thus the expectation for $\hat{P}$ is P as well, satisfying the condition.

3 Results

3.1 Main Result

We show that, given large N, for an SBM the relative efficiency between $\hat{P}$ and $\bar{A}$ is:

$$RE_{ij} = \frac{Var(\hat{P}_{ij})}{Var(\bar{A}_{ij})} = \frac{1/\rho_i + 1/\rho_j}{N} \tag{3}$$

This comes from a proof (outlined in section 5.X) for the variance of $\hat{P}_{ij}$ under the condition that N is large:

$$Var(\hat{P}_{ij}) = \frac{(1/\rho_i + 1/\rho_j)P_{ij}(1 - P_{ij})}{NM} \tag{4}$$

Further, knowing that $\bar{A}_{ij}$ is the estimator of the Bernoulli parameter P_{ij} from M samples, with variance $P_{ij}(1 - P_{ij})/M$, this yields the above result.

This result implicates then that for large graphs that follow a stochastic block model, a better estimate for the mean graph, under mean squared error (MSE), is the $\hat{P}$ estimate.

3.2 Validation with Simulations

To demonstrate the above result for the variance of $\hat{P}$ and relative efficiency (Equations x.x and x.x), we simulate random graphs from an SBM with parameters:

$$B = \begin{bmatrix} .42 & .2 \\ .2 & .7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} .5 & .5 \end{bmatrix}$$

From this model we perform simulated studies by sampling M adjacency matrices with N vertices to calculate both $\bar{A}$ and $\hat{P}$. With these estimators for P, we can calculate the MSE of each block region in the model, defined as the elements of the adjacency matrix that have the same edge-wise probability. This simulation is then repeated 1000 times and the estimates for variance and relative efficiency are averaged over the trials. We then confirm that these simulation errors match with our predictions.

Using this model, we first aim to verify equation XX for $Var(\hat{P}_{ij})$. Figure 1a demonstrates that as N increases, the $Var(\hat{P}_{ij})$ for this block model converges to our estimate, represented as the dotted lines. Further Figure 1b illustrates that the result holds independently of the value of M, and that high M is not a necessary condition for our estimate.

To verify the result for relative efficiency, we again use the previous model for simulations to compare the estimators $\bar{A}$ and $\hat{P}$. As Figure 2 demonstrates, for large N equation X holds true.

To confirm that equations x and x hold with respect to ρ values, we now examine simulations where the ρ vector for the SBM is varied, while fixing N and M.

$$B = \begin{bmatrix} .42 & .2 \\ .2 & .7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} \rho_1 & \rho_2 \end{bmatrix}, \qquad N = 500, \qquad M = 100$$

Figure 3 demonstrates the effect of different block membership, ρ, values on both Var($\hat{P}$) and RE. These simulated results again match well for the predictions from equation x and x, with a mean deviation of 3.7e-7, and 1.6e-4, respectively.

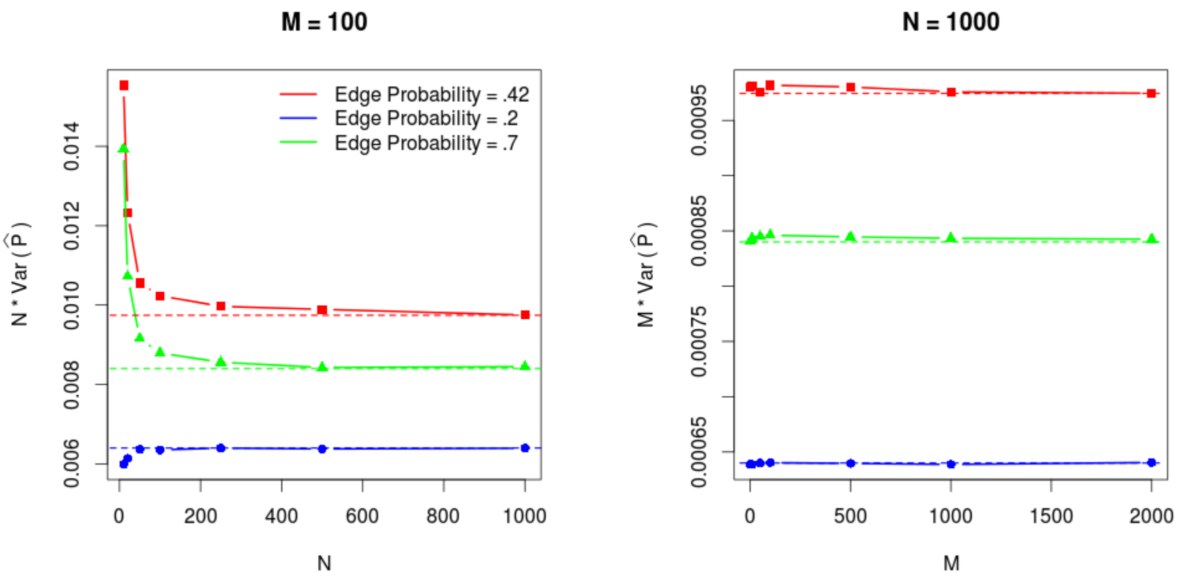

Figure 1: Simulation results. (a) N*Var($\hat{P}$) and (b) M*Var($\hat{P}$) calculated from edges with associated edge probabilities, while increasing N and M, respectively. Observe that the simulated values asymptotically, in N, converge to the predictions, represented by the dotted lines.

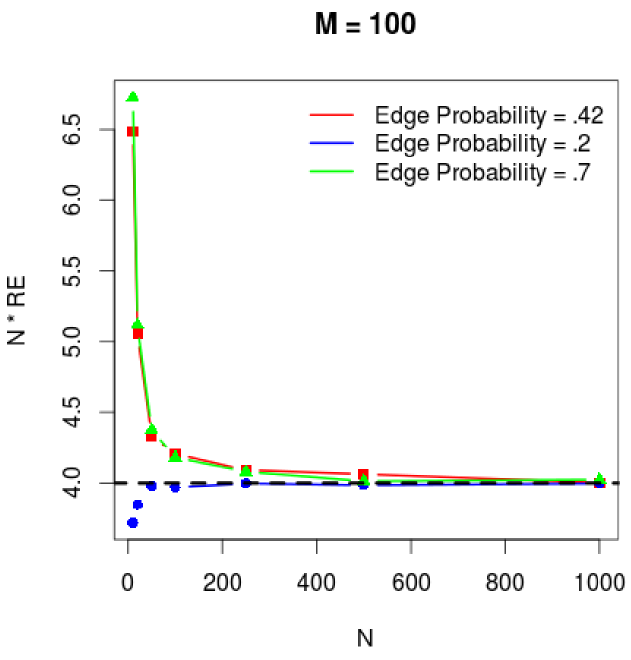

Figure 2: N*RE calculated from edges with associated edge probabilities. Observe that the simulated values asymptotically converge to the predictions, represented by the dotted line.

3.3 CoRR Brain Graphs: Cross-Validation

To demonstate that the $\hat{P}$ estimate is valid under data that does not perfectly follow an SBM, we examine a set of 464 brain connectomes generated from fMRI scans available at the Consortium for Reliability and Reproducibility (CoRR). Details on this dataset and connectiome generation can be seen in section x.x. The connectomes generated each

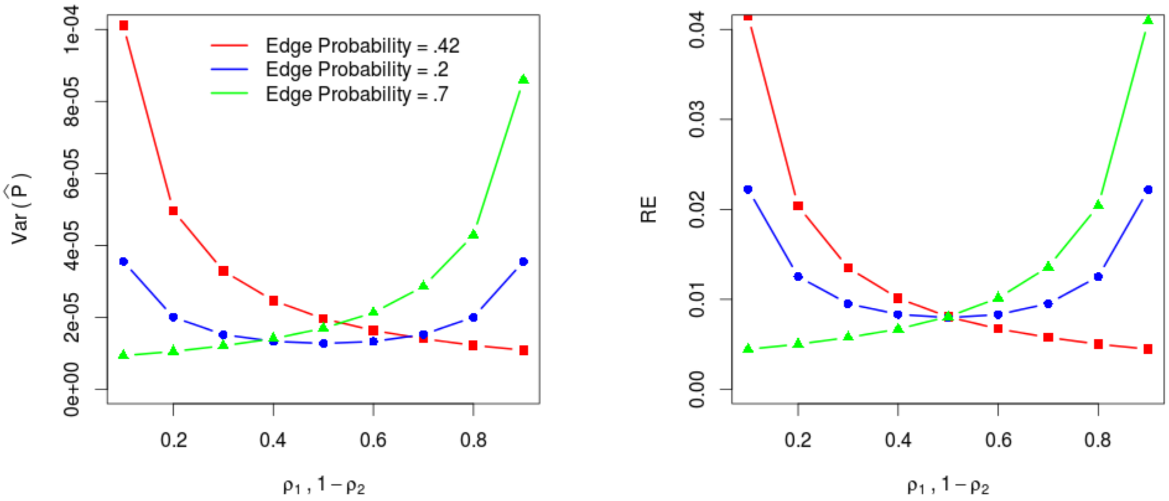

Figure 3: Simulated results for (a) Var($\hat{P}$) and (b) RE calculated from edges with associated edge probabilities. The simulated values for the variance and RE measurements deviated from the predictions with a mean of 3.7e-7, and 1.6e-4, respectively.

have 788 vertices, with anatomical correspondence. To compare $\bar{A}$ and $\hat{P}$ we perform a cross-validation study to examine the impact of the number of available graphs, M. For each sample size M, we randomly sample M adjacency matrices from the CoRR data set and estimate the mean with both $\bar{A}$ and $\hat{P}$. We then calculated the MSE of these estimators compared to the test mean, defined to be the $\bar{A}$ estimate of the $(464 - m)$ remaining samples.

Figure 4 demonstrates that for this dataset, the $\hat{P}$ estimate outperforms $\bar{A}$ in MSE, justifying that the $\hat{P}$ is a valid and likely more accurate estimate of P even when the data does not perfectly follow an SBM.

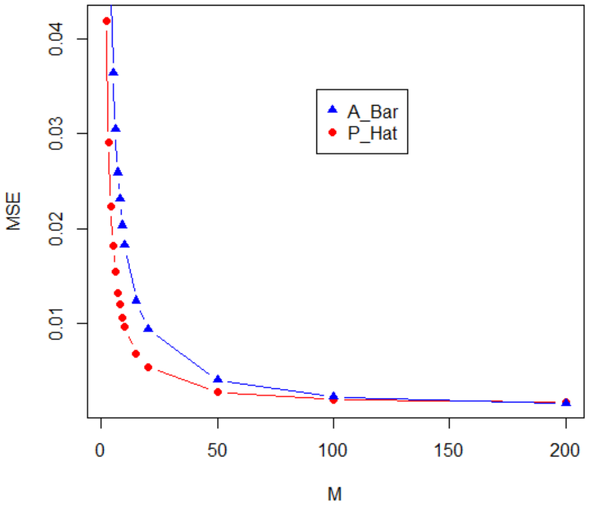

Figure 4: Mean squared error, calculated through cross-validation, in estimating the mean graph on the CoRR brain graphs.

4 Discussion

Given the popularity of large N connectome datasets that have known vertex correspondence, perhaps ASE should be chosen over ABar when estimating a group averaged graph.

NMF may be even better estimate than ASE. [5]

5 Methods

5.1 Algorithm: $\hat{P}$

Input: $A_1, A_2, ..., A_M$, with each $A_i \in \{0,1\}^{N \times N}$ having vertex correspondence

1. Calculate $\bar{A} = \frac{1}{M} \sum\limits_{m=1}^{M} A_m$
2. Estimate SBM parameter k (see section 5.2)
3. Form the matrix $X \in R^{n \times k}$ with the columns in X consisting of the eigenvectors corresponding to the largest, in absolute value, eigenvalues of $\bar{A}$, with the diagonal entries augmented (see section 5.3).
4. $\hat{P} = XX^T$

5.2 Choosing Dimension

Often in dimensionality reduction techniques, the choice for dimension, k, relies on visually analyzing a plot of the ordered eigenvalues, looking for a ”gap” or ”elbow” in this scree-plot. Zhu and Ghodsi [**?**] present an automated method for finding this gap in the scree-plot that takes only the ordered eigenvalues as an input. In order to prevent underestimating k, which is much more harmful than over-estimating, we initialize $k_0 = 0$ and iterate over the Zhu and Ghodsi algorithm by removing the first k_{i-1} eigenvalues from calculation at each iteration to determine the location of the ”next elbow”. For the experiments performed in this work, we choose k to be k_3 under this approach.

(Show the scree plot for a connectome here)

5.3 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 creates a bias in the calculation of the eigenvectors. We minimize this bias by using an iterative method developed by Scheinerman and Tucker [8]. In this method, steps 3 and 4 of the $\hat{P}$ algorithm are repeated, each time replacing the diagonal component of $\bar{A}$ with the diagonal of $\hat{P}$, until $\hat{P}$ converges.

5.4 Dataset Description

The connectomes analyzed were created from resting state functional MRI (fMRI) and Diffusion Tensor Imaging (DTI) scans from the Consortium for Reliability and Reproducibility (CoRR) and are available via the International Neuroimaging Data-sharing

Initiative (INDI). The SWU 4 - Southwest University image collection was used to generate 464 connectomes with 788 anatomically corresponding vertices. (Need to describe how graphs were made with reference, will ask eric bridgeford)

5.5 Source code and data

5.6 Proof of Var($\hat{P}$)

Here we provide an outline of the proof for the Var($\hat{P}$) result presented in section 3.1. The full proof is provided at ...

References

[1] Avanti Athreya, Vince Lyzinski, David J Marchette, Carey E Priebe, Daniel L Sussman, and Minh Tang. A central limit theorem for scaled eigenvectors of random dot product graphs arXiv : 1305 . 7388v2 [math . ST] 23 Dec 2013. pages 1–24, 2013.

[2] Andressa Cerqueira, Daniel Fraiman, Claudia D Vargas, and Florencia Leonardi. A test of hypotheses for random graph distributions built from EEG data. pages 1–17, 2015.

[3] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. 2012.

[4] Charles Freer. Functional Neuroimaging. *Journal of neurology, neurosurgery, and psychiatry*, 59:220, 2014.

[5] N D Ho. Nonnegative matrix factorization algorithms and applications. *Thesis*, (June):185, 2008.

[6] By Vince Lyzinski, Daniel L Sussman, Minh Tang, Avanti Athreya, and Carey E Priebe. PERFECT CLUSTERING FOR STOCHASTIC. pages 1–13, 2000.

[7] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. page 46, 2009.

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