Law of Large Graphs

September 7, 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 graphs represented as adjacency matrices, A_m, each having N vertices with known correspondence. The graphs we consider are undirected and unweighted with no self-loops, thus each A_m is a binary ($\in \{0,1\}$) symmetric matrix with zeros along the diagonal. An example scenario of this arises in the field of connectomics, where functional brain imaging data for each subject can be represented as a graph, with each vertex having a defined anatomical correspondence, and an edge between two regions is defined to exist if correlation in activity between the regions reaches a certain threshold. In this setting, we may consider each graph to be a random graph. The general model for such graphs is the Independent Edge Model with parameter $P \in [0,1]^{N \times N}$. Each P_{ij} refers to the probability that an edge exists between vertex i and vertex j. We aim to estimate P with our finite sample of M graphs.

2.1 Entry-Wise Least Squares Estimate

The most intuitive approach in this scenario is the element-wise mean among the adjacency matrices:

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

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 examined in isolation, to estimate the mean graph P one should take the element-wise mean, $\bar{A}$.

2.2 Model Assumption: Stochastic Block Model

Since the number of independent edges increases on the order of N^2 with respect to the number of vertices, it becomes necessary to make structural assumptions on the vertices to better estimate the matrix P. A first structural assumption is the stochastic block model (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. This imposes the idea of *structural equivalence*, where vertices are defined to be structurally equivalent if there connections to other nodes are similar. In the stochastic block model, groups of vertices, or blocks, are then structurally equivalent since the vertices contained have equal likelihood in their connections among the blocks. An example of block structure can be thought to exist in functional brain imaging, for instance the structures in the basal ganglia will likely behave similarly in their connections and may be considered a block.

The SBM is formally defined by the parameters k, ρ, and B. In this model each vertex is assigned to one of k blocks and the fraction of vertices belonging to the i th block is designated as ρ_i. The connection probabilities of this block structure are stored in the symmetric $k \times k$ block matrix B, where B_{ij} represents the probability of an edge existing between a vertex of block i and one of block j.

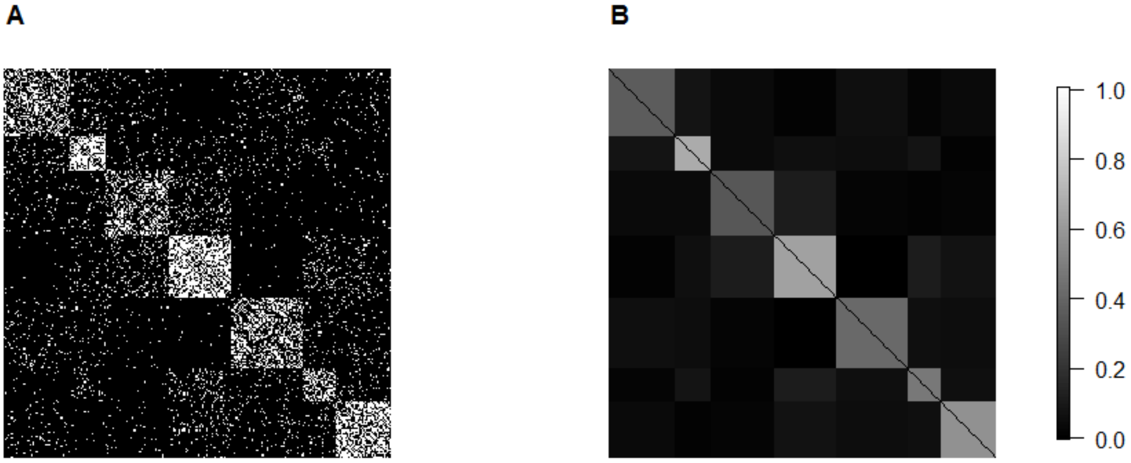

Figure 1: Example illustrating the SBM. (a) Adjacency matrix generated from the (b) edge-wise probability matrix that follows a SBM with k = 7 blocks

2.3 Latent Position Model

In general, the SBM is too strict for applications since vertices, though they may behave similarly, often do not behave identically. A model for random graphs proposed by Hoff et. al. that can loosen this restriction 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. [?]. For example in the functional connectomics, components of the latent vectors may refer the relative importance of an anatomical region among a set of tasks. The magnitude then may refer to how active the region is generally. Therefore, active regions vital for a similar task are more likely to be functionally connected.

In a latent position model representing a SBM, all vertices in the same block would have identical latent positions. In the context of applications, it is unrealistic to have such a strict assumption. However when representing the graph as an RDPG, we can relax this structure by assuming that the latent positions of vertices within a certain block are gaussian distributed around the true block latent position. This now allows for a flexible model that can capture block-like structure in the graph while maintaining subtle distinctions among vertices within each block.

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

In order to expose the underlying block structure within a graph, Sussman et. al. studied adjacency spectral embedding (ASE) to enforce a low rank$-k$ approximation on the the probability matrix P [?]. 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 of the adjacency matrix. With this embedding, X, each row is then a latent vector for the corresponding vertex.

In this work, we use ASE to embed the mean matrix $\bar{A}$, rather than the adjacency matrix alone. By making the assumption that there is an underlying block-distibuted RDPG 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), in mean squared error (MSE), among the two defined as:

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

3 Results

3.1 Theoretical Results

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

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

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

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

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 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 $MSE(\hat{P}_{ij})$. Figure 2a demonstrates that as N increases, the $MSE(\hat{P}_{ij})$ for this block model converges to our estimate, represented as the dotted lines. Further Figure 2b illustrates that the result holds independently of the value of M, and that high M is not a necessary condition for our estimate.

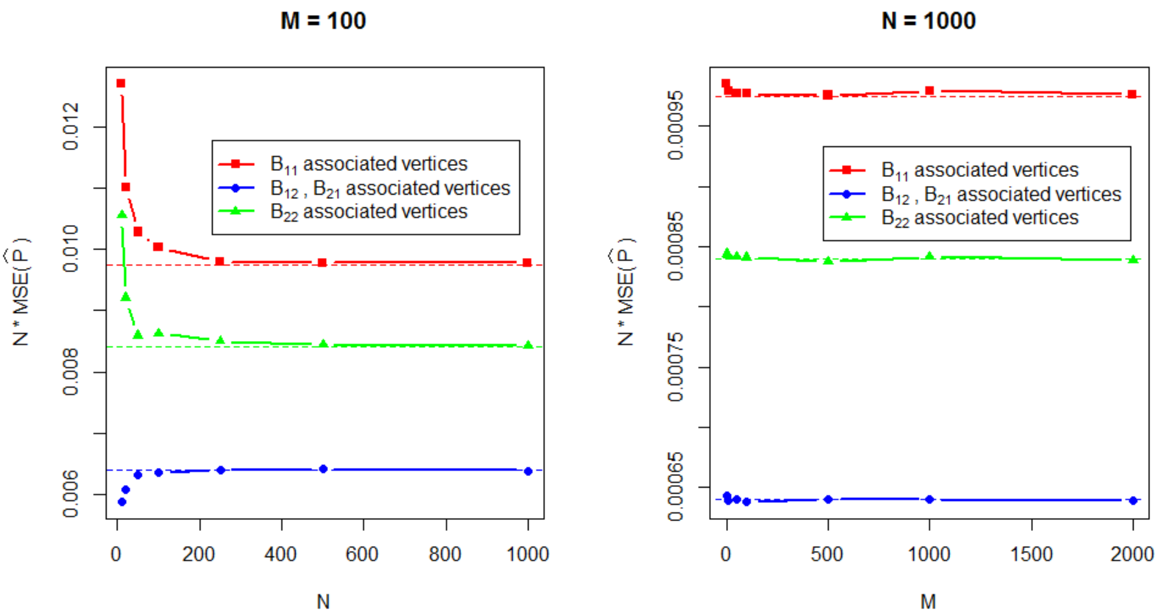

Figure 2: Simulation results. (a) N*MSE($\hat{P}$) and (b) M*MSE($\hat{P}$) calculated from edges with associated edge probabilities, while increasing N and M, respectively. Observe that the simulated values converge asymptotically, in N, to the predictions represented by the dotted lines.

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 3 demonstrates, for large N equation X holds true.

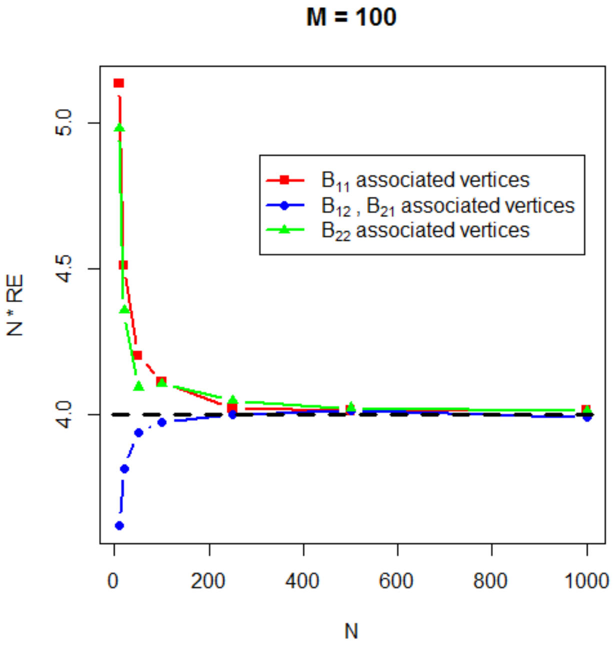

Figure 3: N*RE calculated from edges with associated edge probabilities. Observe that the simulated values asymptotically converge to the predictions, represented by the dotted line.

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 setting $N = 500$, $M = 100$ and using the same B matrix used above.

Figure 4 demonstrates the effect of different block membership, ρ, values on both MSE($\hat{P}$) and RE. These simulated results again match well for the predictions from equation x and x, with a mean deviation of 2.4e-7, and 1.1e-4, respectively.

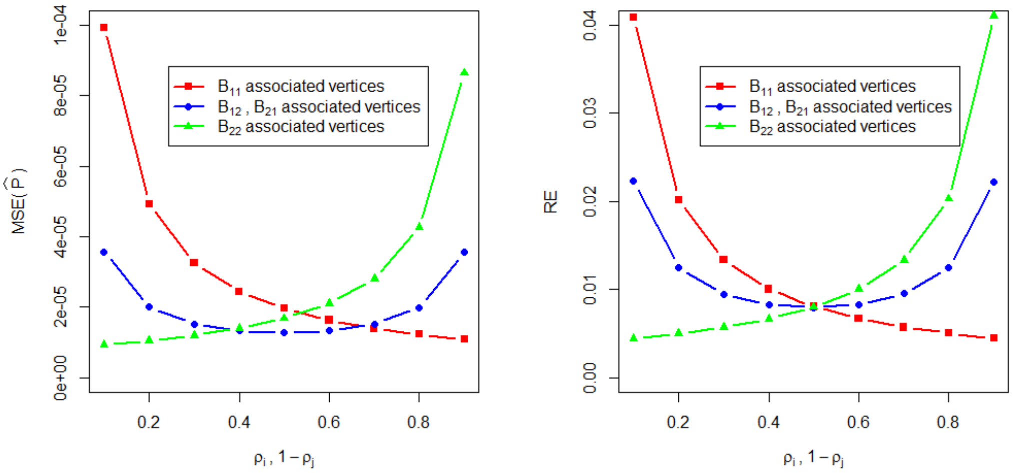

Figure 4: Simulated results for (a) Var($\hat{P}$) and (b) RE calculated from edges with associated edge probabilities. The simulated values for the variance and RE measurements deviated from the predictions with a mean of 2.4e-7, and 1.1e-4, respectively.

3.3 CoRR Brain Graphs: Cross-Validation

To demonstrate that the $\hat{P}$ estimate is valid under data that does not perfectly follow a 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 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 calculate 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.

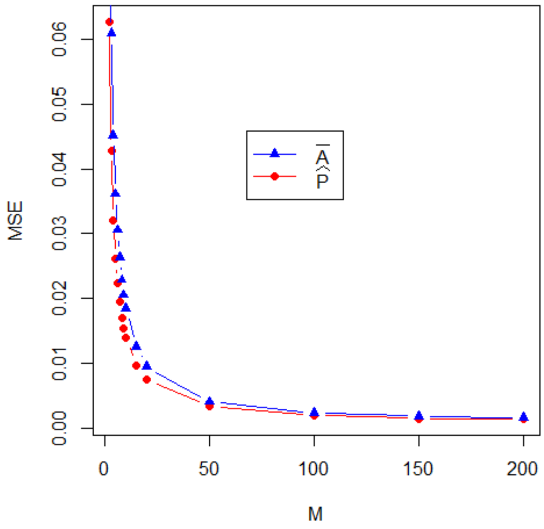

Figure 5: Mean squared error for $\bar{A}$ and $\hat{P}$, calculated through cross-validation, in estimating the mean graph on the CoRR brain graphs.

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 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 and Corr data set effect of M on k)

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 leads to 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 Outline for Proof of Relative Efficiency

Here we provide an outline of the proof for the MSE($\hat{P}$) result presented in section 3.1.

When comparing two estimators, the first thing we need to consider is consistency. It is easy to see that $\bar{A}$ is unbiased as an estimate of P. Moreover, since two latent positions are conditionally asymptotically independent by extended version of Corollary 4.11 in Athreya et al. (2013), we know $\hat{P}$ is consistent, as well as $\bar{A}$.

Thus the relative efficiency between $\hat{P}$ and $\bar{A}$, which is equivalent to the ratio of mean square errors in this case, is a good indicate in comparison. Since $\hat{P}_{ij} = \hat{X}_i^T \hat{X}_j$ is a noisy version of the dot product of $\nu_s^T \nu_t$, by Equation 5 in Brown and Rutemiller (1977), combined with asymptotic independence between $\hat{X}_i$ and $\hat{X}_j$, and the covariance matrices given by extended version of Corollary 4.11 in Athreya et al. (2013), we have the variance of $\hat{P}_{ij}$ converges to $\left(1/\rho_{\tau_i} + 1/\rho_{\tau_j}\right) P_{ij}(1-P_{ij})/(n\cdot m)$ as $n \to \infty$. Since the variance of $\bar{A}_{ij}$ is $P_{ij}(1-P_{ij})/m$, the relative efficiency between $\hat{P}_{ij}$ and $\bar{A}_{ij}$ is approximately $(\rho_{\tau_i}^{-1}+\rho_{\tau_j}^{-1})/n$ when n is sufficiently large.

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.