Random Graphs

Hanwen Hu
STA6557 Final Presentation

April 21, 2022

Table of Contents

Graph

- A graph G is an ordered pair of (V, E), where V is the vertex set, and E is the set of edges, also a subset of the Cartesian product of $V \times V$.
- If a graph G has n vertices (i.e. $|V| = n$), we denote $V = \{1, 2, \ldots, n\}$, and we say there is a connection between vertex i and j if $(i, j) \in E$.
- The adjacency matrix A provides a compact representation of G:

$$A_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E, \\ 0, & \text{o.w..} \end{cases}$$

- For a random graph G, the probability of connection could be denoted by a probability matrix P, where $P_{ij} = P((i,j) \in E)$.
- In practice, P is not observable, instead we observe A, a noisy version of P.

Random Dot Product Graph (RDPG) Model

- Let F be a probability distribution whose support is given by $\mathcal{X}_d \subset \mathbb{R}^d$, then it is a d-**dimensional inner product distribution** on $\mathbb{R}^d$, if $\forall x, y \in \mathcal{X}^d$, we have $\langle x, y \rangle \in [0, 1]$.
- Let F be a d-dimensional inner product distribution with $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$, collected in the rows of the matrix

$$\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T \in \mathbb{R}^{n \times d}.$$

- For example, thinking of the vertices as members of a social network, the vectors together with the dot product encode semantically the idea of differing "interests" and varying levels of "talkativeness". The more two members share the same interest, the more possible that they will build a link between each other.

Random Dot Product Graph (RDPG) Model

- Suppose A is a random adjacency matrix given by

$$\mathrm{P}(A|\boldsymbol{X}) = \prod_{i<j} (\langle X_i, X_j \rangle)^{A_{ij}} (1 - \langle X_i, X_j \rangle)^{1-A_{ij}},$$

  then we write $(A, \boldsymbol{X}) \sim \mathrm{RDPG}(F, n)$, and say that A is the adjacency matrix of a Random Dot Product Graph (RDPG) of dimension at most d with latent positions given by $X_1, X_2, \ldots, X_n$.
- Moreover, if given fixed latent positions $\boldsymbol{X}$, a graph G is generated according to the distribution above, we say A is a realization of a RDPG with latent positions $\boldsymbol{X}$ and denote that $A \sim \mathrm{RDPG}(\boldsymbol{X})$.

Non-identifiability

- Given a graph G distributed as an RDPG, the natural task is to recover the latent position X that generates G. However, the RDPG model has an inherently non-identifiability: Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be the latent positions and $W \in \mathbb{R}^{d \times d}$ be an orthonormal matrix, Then we have

$$\boldsymbol{X}\boldsymbol{X}^T = (\boldsymbol{X}W)(\boldsymbol{X}W)^T,$$

which implies that $\boldsymbol{X}$ and $\boldsymbol{X}W$ will give rise to the same distribution over the graphs.

Adjacency Spectral Embedding

- Given a symmetric, positive semi-definite matrix $Q \in \mathbb{R}^{n\times n}$, the spectral decomposition of Q is given by

$$Q = U_Q S_Q U_Q^T,$$

where $U_Q \in \mathbb{R}^{n\times n}$ is orthonormal, and S_Q is diagonal with the eigenvalues of Q.

- Let $|A| = (A^T A)^{1/2}$. Given a positive integer $d \geq 1$ and an adjacency matrix A of n vertices, the **Adjacency Spectral Embedding (ASE)** of A into $\mathbb{R}^d$ is given by $\hat{\boldsymbol{X}} = U_0 S_0^{1/2}$, where

$$|A| = [U_0 | U_0^{\perp}][S_0 \oplus S_0^{\perp}][U_0 | U_0^{\perp}]^T$$

is the spectral decomposition of $|A|$, S_0 is the diagonal matrix with d largest eigenvalues of $|A|$, and each column of $U_0 \in \mathbb{R}^{n\times d}$ is the corresponding eigenvector.

Laplacian Spectral Embedding

- On the other hand, we may define the **Laplacian Spectral Embedding (LSE)** of A in the following way:
- Given an adjacency matrix A, let $\mathcal{L}(A) = D^{-1/2}AD^{-1/2}$ denote the normalized Laplacian of A, where D is the diagonal matrix whose diagonal entries $D_{ii} = \sum_{j \neq i} A_{ij}$.
- Given a positive integer $d \geq 1$ and an adjacency matrix A of n vertices, the LSE of A into $\mathbb{R}^d$ is given by $\breve{\mathbf{X}} = U_1 S_1^{1/2}$, where

$$|\mathcal{L}(A)| = [U_1 | U_1^{\perp}][S_1 \oplus S_1^{\perp}][U_1 | U_1^{\perp}]^T$$

is the spectral decomposition of $|\mathcal{L}(A)|$, S_1 is the diagonal matrix with d largest eigenvalues of $|\mathcal{L}(A)|$, and each column of $U_1 \in \mathbb{R}^{n \times d}$ is the corresponding eigenvector.

Consistency of Embeddings

Theorem (Consistency of ASE (Lyzinski, Vince et al. 2016))

Let $A_m \sim RDPG(\boldsymbol{X}^m)$ for $m \geq 1$ be a sequence of RDPGs, where $\boldsymbol{X}^m$ is assumed to be of rank d for all sufficiently large m. And let $\hat{\boldsymbol{X}}^m$ be the ASE of A_m, and let $\boldsymbol{X}_i^m, \hat{\boldsymbol{X}}_i^m$ be the i-th row of $\boldsymbol{X}^m, \hat{\boldsymbol{X}}^m$. Then as $m \to \infty$, the probability that there exists $W_m \in O(d)$ such that

$$\max_{1 \leq i \leq m} \|\hat{\boldsymbol{X}}_i^m - W_m \boldsymbol{X}_i^m\| \leq \frac{Cd^{1/2} \log^2 m}{\delta^{1/2}(P^m)}$$

goes to 1, i.e. this event occurs asymptotically almost surely. $C > 0$ is some fixed constant, $P^m = \boldsymbol{X}^m(\boldsymbol{X}^m)^T$, and $\delta(P) = \max_i \sum_{j=1}^m P_{ij}$.

Distributional Results

Theorem (CLT for Rows of ASE (Athreya, Avanti, et al. 2016))

Let $(A^m, \boldsymbol{X}^m) \sim RDPG(F)$ be a sequence of adjacency matrices and associated latent positions of a d-dimensional RDPG according to a inner product distribution F supported on $\mathcal{X}^d \subset \mathbb{R}^d$. Let $\Phi(x, \Sigma)$ denote the CDF of a multivariate Gaussian with mean 0 and covariance matrix Σ evaluated at $x \in \mathbb{R}^d$, then there exists a sequence of $(W_m)_{m=1}^{\infty} \subset O(d)$, such that for any $z \in \mathbb{R}^d$ and fixed index i,

$$\lim_{m\to\infty} P(\sqrt{m}(\hat{\boldsymbol{X}}_i^m - W_m \boldsymbol{X}_i^m) \le z) = \int_{\mathcal{X}^d} \Phi(z, \Sigma(x)) dF(x),$$

where

$$\Sigma(x) = \Delta^{-1}\mathbb{E}[(x^T X_1 - (x^T X_1)^2) X_1 X_1^T]\Delta^{-1}, \text{ and } \Delta = \mathbb{E}[X_1 X_1^T].$$

Distributional Results

Theorem (CLT for Rows of LSE (Tang, Priebe 2018))

Let $(A^m, \boldsymbol{X}^m) \sim RDPG(F)$ be a sequence of adjacency matrices and associated latent positions of a d-dimensional RDPG according to a inner product distribution F supported on $\mathcal{X}^d \subset \mathbb{R}^d$. Then there exists a sequence of $(W_m)_{m=1}^{\infty} \subset O(d)$, s.t. for any $z \in \mathbb{R}^d$ and fixed index i,

$$\lim_{m\to\infty} P(m(\breve{\boldsymbol{X}}_i^m - W_m \frac{\boldsymbol{X}_i^m}{\sqrt{\sum_j (\boldsymbol{X}_i^m)^T \boldsymbol{X}_j^m}}) \leq z) = \int_{\mathcal{X}^d} \Phi(z, \tilde{\Sigma}(x)) dF(x),$$

$$\tilde{\Sigma}(x) = \mathbb{E}[(\frac{\tilde{\Delta}^{-1} X_1}{X_1^T \mu} - \frac{x}{2x^T \mu})(\frac{X_1^T \tilde{\Delta}^{-1}}{X_1^T \mu} - \frac{x^T}{2x^T \mu})\frac{x^T X_1 - x^T X_1 X_1^T x}{x^T \mu}],$$

$$\mu = \mathbb{E}[X_1] \in \mathbb{R}^d, \ \tilde{\Delta} = \mathbb{E}[\frac{X_1 X_1^T}{X_1^T \mu}] \in \mathbb{R}^{d \times d}.$$

Semiparametric Hypothesis Test for Graph Data

- Given two adjacency matrices A, B for two graphs with the same number of nodes, we want to conduct a hypothesis test on whether they share the same latent position, up to an orthonormal transformation. That is, if we assume $A \sim \mathrm{RDPG}(X)$ and $B \sim \mathrm{RDPG}(Y)$, then the null hypothesis $\mathbb{H}_0$ is given as

$$X =_W Y, \text{ i.e. } \exists W \in \mathrm{O}(d), X = YW.$$

- A Bootstrapping hypothesis test procedure is proposed by Tang et.al. (2017).

Semiparametric Hypothesis Test for Graph Data

Algorithm 1: Bootstrapping procedure for the test: $\mathbb{H}_0 : X =_W Y$

Input: Embedding dimension d, Number of bootstrap samples b.

procedure Bootstrap(X, T, b)

$d \leftarrow \text{ncol}(X)$; $\mathcal{S}_X \leftarrow \emptyset$.

for i = 1:b **do**

$A_i \leftarrow \text{RDPG}(\hat{X})$; $B_i \leftarrow \text{RDPG}(\hat{X})$

$\hat{X}_i \leftarrow \text{ASE}(A_i, d)$; $\hat{Y}_i \leftarrow \text{ASE}(B_i, d)$

$T_i \leftarrow \min_W \|\hat{X}_i - \hat{Y}_i W\|_F$; $\mathcal{S}_X \leftarrow \mathcal{S}_X \cup T_i$

end for

return $p \leftarrow (|\{s \in \mathcal{S}_X : s - T \geq 0\}| + 0.5)/b$

end procedure

1. $\hat{X} \leftarrow \text{ASE}(A, d)$; $\hat{Y} \leftarrow \text{ASE}(B, d)$; $T \leftarrow \min_W \|\hat{X} - \hat{Y} W\|_F$
2. $p_X \leftarrow \text{Bootstrap}(\hat{X}, T, b)$, $p_Y \leftarrow \text{Bootstrap}(\hat{Y}, T, b)$
3. $p \leftarrow \max\{p_X, p_Y\}$

Output: p-value of the hypothesis test.

Application Example

- Consider the neural imaging graphs obtained from the test-retest diffusion MRI and magnetization-prepared rapid acquisition gradient echo (MPRAGE) data from Landman et al. (2011). It consists of 42 images, one pair from each of 21 subjects.
- The scans are converted into spatially aligned graphs with $n = 70$ vertices, in which each vertex corrresponds to a particular voxel in a reference coordinate system to which the image is registered. The graphs are then embedded into $\mathbb{R}^4$.
- Pairwise comparisons are done between 42 graphs.

Application Example

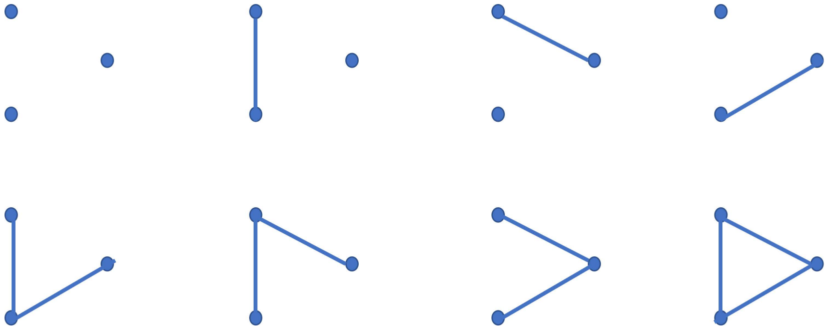
- In general, the test procedure fails to reject the null hypothesis when the two graphs are for the same subject.
- Besides, it also frequently reject the null hypothesis when the two graphs are from scans of different subjects.

Graph Space as a Stratified Space - preliminary work

- Suppose a graph has n vertices, then there are $\binom{n}{2}$ possible edges in the graph. If we only care about edges between vertices, then there are $2^{\binom{n}{2}}$ possible types of graphs in total.
- Moreover, if we generalize to the graphs with **weighted edges**, then a graph with k edges determines a stratum with k positive parameters over the points of an open k-dimensional orthant.
- Therefore, a graph space consisting of all graphs with n vertices is a stratified space with $\binom{n}{2}+1$ strata, where a coordinate in each dimension may, for example account for the distance of a data point on the corresponding edges, from a staring vertex, assuming a directed graph.
- For the tree space version in Omar's final presentation see Billera et al.(2001).
- The dimension of this graph space is $\binom{n}{2}$. In particular, there is one top dimensional stratum with dimension $\binom{n}{2}$. And for any $1 < k < \binom{n}{2}$, there will also be one co-dimension k strata with

Graph Space as a Stratified Space

- The graph space G_3 consisting of all graphs of 3 vertices with weighted edges consists of 4 strata and has a dimension of 3.
- Below, each stratum has a dimension of 0, 1, 2 and 3.



Combining RDPG into Stratified Space Data

- In the process of multi-graph inferences, if graph data with n vertices and arbitrarily connected weighted edges are given, we can model the connection of edges by inferring the latent positions of each vertex. This provides an estimate of probability distribution for a graph to lie on each stratum.
- Furthermore, The location of a graph on each stratum is determined by the weights of each of its edges.

Conclusion

- This presentation introduces the main idea of Random Dot Product Graph model, including the inference of latent positions via ASE and LSE. Some asymptotic results about the embeddings are given. An example of graph hypothesis test procedure is given.
- Besides, we explores the way to model a graph space as a stratified space, to combine the idea of RDPG inference with the view of seeing graph space as a stratified space.

References

1. Athreya, Avanti, et al. "Statistical inference on random dot product graphs: a survey." The Journal of Machine Learning Research 18.1 (2017): 8393-8484.
2. Young, Stephen J., and Edward R. Scheinerman. "Random dot product graph models for social networks." International Workshop on Algorithms and Models for the Web-Graph. Springer, Berlin, Heidelberg, 2007.
3. Lyzinski, Vince, et al. "Community detection and classification in hierarchical stochastic blockmodels." IEEE Transactions on Network Science and Engineering 4.1 (2016): 13-26.
4. Athreya, Avanti, et al. "A limit theorem for scaled eigenvectors of random dot product graphs." Sankhya A 78.1 (2016): 1-18.
5. Tang, Minh, and Carey E. Priebe. "Limit theorems for eigenvectors of the normalized Laplacian for random graphs." The Annals of Statistics 46.5 (2018): 2360-2415.

References

6 B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. Lim, J. A. Farrell, et al. Multi-parametric neuroimaging reproducibility: a 3-t resource study. Neuroimage, 54: 2854-2866, 2011.

7 M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random dot product graphs. Journal of Computational and Graphical Statistics, 26:344-354, 2017a.

8 Barden, Dennis, Huiling Le, and Megan Owen. "Central limit theorems for Fréchet means in the space of phylogenetic trees." Electronic journal of probability 18 (2013): 1-25.

9 Billera, L. J., Holmes, S. P. and Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv. in Appl. Math. 27 (2001), 733–767. MR-1867931