Two-sample hypothesis testing for random dot product graphs

Minh Tang

Department of Applied Mathematics and Statistics
Johns Hopkins University

JSM 2014

Joint work with Avanti Athreya, Vince Lyzinski,
Carey E. Priebe and Daniel L. Sussman.

Introduction and Overview

1. The problem of deciding whether two give graphs are the “same” has applications in e.g., neuroscience, social networks.
2. We propose a valid and consistent test for the above under a random graph model.
3. The test proceeds by embedding the graphs into Euclidean space followed by computing a distance between a kernel density “estimate” of the embedded points.

Random dot product graphs

Let Ω be a subset of $\mathbb{R}^d$ such that, for all $\omega, \omega' \in \Omega$, $0 \leqslant \langle \omega, \omega' \rangle \leqslant 1$. Let F be a distribution taking values in Ω.

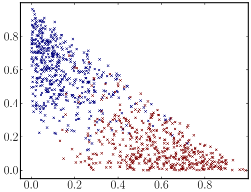
1. Let $\{X_i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} F$.
2. $A_n \sim \text{RDPG}(F)$ is the adjacency matrix of a graph associated with $\{X_i\}_{i=1}^n$. The upper diagonal entries of A_n are independent Bernoulli random variables with $\mathbb{P}[X_i \sim X_j] = \langle X_i, X_j \rangle$, i.e.,

$$\mathbb{P}[A_n \,|\, \{X_i\}_{i=1}^n] = \prod_{i<j} \langle X_i, X_j \rangle^{A_n(i,j)} (1 - \langle X_i, X_j \rangle)^{1-A_n(i,j)}$$

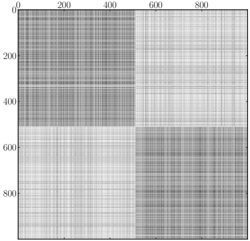
See Young and Scheinerman (2007).

- Random dot product graphs are an example of *latent position graphs* (Hoff et al., 2002), in which each vertex is associated with a latent position.
- Random dot product graphs are related to stochastic blockmodels Holland et al. (1983), degree-corrected stochastic block models Karrer and Newman (2011), and mixed membership block models Airoldi et al. (2008).
- Non-identifiability: For any distribution F and orthogonal matrix W, the graphs $A \sim \mathrm{RDPG}(F)$ and $B \sim \mathrm{RDPG}(F \circ W)$ are identically distributed.

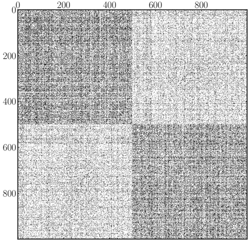
$X = \{X_i\}_{i=1}^{n} \subset \mathbb{R}^d$
original latent vectors



$P = XX^T \in [0, 1]^{n \times n}$
probability matrix



$A = \mathrm{Bern}(K)$
adjacency matrix



Observation

A looks like P (at least at rough scale).

Problem Statement

Given $A \sim \mathrm{RDPG}(F)$ and $B \sim \mathrm{RDPG}(G)$, consider the following test:

$$\mathbb{H}_0 \colon F =_W G \quad \text{against} \quad \mathbb{H}_1 \colon F \neq_W G$$

where $F =_W G$ denotes that there exists an orthogonal $d \times d$ matrix W such that $F = G \circ W$ and $F \neq_W G$ denotes that $F \neq G \circ W$ for any orthogonal W.

Adjacency spectral embedding

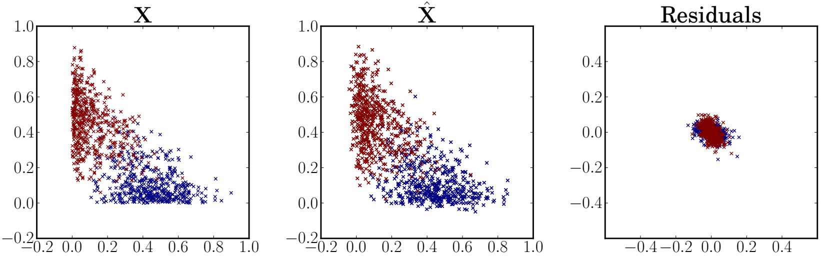
Definition

Let A be an $n \times n$ adjacency matrix and denote by $|A|$ the matrix $(A^{\top}A)^{1/2}$. Let $d \geqslant 1$ and consider the following spectral decomposition of $|A|$

$$|A| = [U_A|\tilde{U}_A][S_A \oplus \tilde{S}_A][U_A|\tilde{U}_A]$$

where $U_A \in \mathbb{R}^{n \times d}$, $\tilde{U}_A \in \mathbb{R}^{n \times d}$. The columns of U_A correspond to the d largest eigenvalues of $|A|$. The adjacency spectral embedding of A into $\mathbb{R}^d$ is then the $n \times d$ matrix $\hat{X} = U_A S_A^{1/2}$.

$\hat{X}$ is close to X



Modicum of consistency I

Theorem

Suppose $(A, X) \sim \mathrm{RDPG}(F)$ *is a graph on* n *vertices. Denote by* $\hat{X}$ *the adjacency spectral embedding of* A *into* $\mathbb{R}^d$*. Let* $\eta > 0$ *be arbitrary. Then for sufficiently large* n *there exists a* $d \times d$ *orthogonal matrix* W *such that, with probability at least* $1 - 3\eta$*,*

$$\left| \|\hat{X} - XW\|_F - C_1(F) \right| \leqslant \frac{C_2(F) d^{3/2} \log(n/\eta)}{\sqrt{n}} \tag{1}$$

where $C_1(F)$ *and* $C_2(F)$ *are constants depending only on* F*.*

Two-sample testing via maximum mean discrepancy

Let κ be a kernel on Ω with reproducing kernel Hilbert space $\mathcal{H}$. Denote by $\mathcal{F}$ the unit ball $\mathcal{F} = \{h \in \mathcal{H} \colon \|h\|_{\mathcal{H}} \leqslant 1\}$.

For a distribution F taking values in Ω the map $\mu[F]$ defined by

$$\mu[F] := \int_{\Omega} \kappa(\omega, \cdot)\, dF(\omega).$$

belongs to $\mathcal{H}$. If κ is a *universal kernel*, then μ is an injective map.

Let F and G be probability distributions taking values in Ω; $X, X' \sim F$ and $Y, Y' \sim G$. Then

$$\begin{aligned} \|\mu[F] - \mu[G]\|_{\mathcal{H}}^2 &= \sup_{h \in \mathcal{F}} |\mathbb{E}_F[h] - \mathbb{E}_G[h]|^2 \\ &= \mathbb{E}[\kappa(X, X')] - 2\mathbb{E}[\kappa(X, Y)] + \mathbb{E}[\kappa(Y, Y')]. \end{aligned} \tag{2}$$

is an integral probability metric, termed the *maximum mean discrepancy* Gretton et al. (2012).

Denote by $\Phi\colon \Omega \mapsto \mathcal{H}$ the canonical feature map

$$\Phi(X) = \kappa(\cdot, X)$$

of κ. Given $\{X_i\} \overset{\text{i.i.d}}{\sim} F$ and $\{Y_i\} \overset{\text{i.i.d}}{\sim} G$, the quantity $V_{n,m}(X, Y)$

$$\begin{aligned}
V_{n,m}(X, Y) &= \Big\| \frac{1}{n} \sum_{i=1}^{n} \Phi(X_i) - \frac{1}{m} \sum_{k=1}^{n} \Phi(Y_k) \Big\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{k=1}^{n} \kappa(X_i, Y_k) \\
&+ \frac{1}{m^2} \sum_{k=1}^{m} \sum_{l=1}^{m} \kappa(Y_k, Y_l).
\end{aligned}$$

is a *consistent estimate* of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$.

Test statistic

Denote by $\hat{X} = \{\hat{X}_1, \ldots, \hat{X}_n\}$ and $\hat{Y} = \{\hat{Y}_1, \ldots, \hat{Y}_m\}$ the adjacency spectral embedding of A and B, respectively. Assume that κ is a unitarily invariant kernel, e.g., a radial kernel. Define the test statistic $V_{n,m}(\hat{X}, \hat{Y})$ as follows:

$$V_{n,m}(\hat{X}, \hat{Y}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(\hat{X}_i, \hat{X}_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \kappa(\hat{X}_i, \hat{Y}_k) + \frac{1}{m^2} \sum_{l=1}^{m} \sum_{k=1}^{m} \kappa(\hat{Y}_k, \hat{Y}_l)$$

Modicum of consistency II

Theorem

Let $(X, A) \sim \mathrm{RDPG}(F)$ *and* $(Y, B) \sim \mathrm{RDPG}(G)$ *be independent random dot product graphs with latent position distributions* F *and* G *satisfying distinct eigenvalues assumption. Consider the hypothesis test*

$$\mathbb{H}_0 \colon F =_W G \quad \textit{against} \quad \mathbb{H}_1 \colon F \neq_W G$$

Suppose $m, n \to \infty$ *and* $m/(m+n) \to \rho \in (0,1)$*. Then under the null*

$$(m+n)(V_{n,m}(\hat{X}, \hat{Y}) - V_{n,m}(X, YW)) \xrightarrow{\text{a.s.}} 0 \tag{3}$$

where W *is any orthogonal matrix such that* $F = G \circ W$*.*

Sketch of argument

Eq. (3) that

$$(m+n)(V_{n,m}(\hat{X},\hat{Y}) - V_{n,m}(X, YW)) \xrightarrow{\text{a.s.}} 0$$

follows from the following bound

$$\sup_{f\in\mathcal{F}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(W\hat{X}_i) - f(X_i) \big) \Big| \xrightarrow{\text{a.s.}} 0$$

established via Taylor's expansion and a covering number argument.

Limiting distribution of $V_{n,m}(\hat{X}, \hat{Y})$.

Hence under the null hypothesis of $F =_{\mathcal{W}} G$, evoking previous results of Anderson et al. (1994) and Gretton et al. (2012) for $V_{n,m}(X, Y)$, one has

$$(m+n)V_{n,m}(\hat{X}, \hat{Y}) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l \chi_{1l}^2 \tag{4}$$

where $\{\chi_{1l}^2\}$ are independent χ^2 random variables with one degree of freedom and $\{\lambda_l\}$ are the eigenvalues of the integral operator

$$I_{F,\tilde{\kappa}}(\phi) = \int_{\Omega} \phi(y)\tilde{\kappa}(x, y)dF(y)$$

Simulation Results

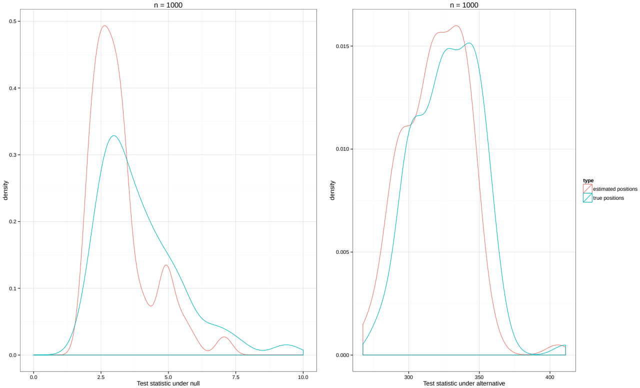


Figure 1 : Distribution of test statistics under null and alternative as computed from the latent positions and those estimated from adjacency spectral embedding for testing the null hypothesis $F =_W G$.

	$\epsilon = 0.02$		$\epsilon = 0.05$		$\epsilon = 0.1$	
n	$\{X, Y\}$	$\{\hat{X}, \hat{Y}\}$	$\{X, Y\}$	$\{\hat{X}, \hat{Y}\}$	$\{X, Y\}$	$\{\hat{X}, \hat{Y}\}$
100	0.07	0.06	0.07	0.09	0.21	0.27
200	0.06	0.09	0.11	0.17	0.89	0.83
500	0.08	0.1	0.37	0.43	1	1
1000	0.1	0.14	1	1	1	1

Table 1 : Power estimates for testing the null hypothesis $F =_{\mathcal{W}} G$ at a significance level of $\alpha = 0.05$ using bootstrap permutation tests for $V_{n,m}(\hat{X}, \hat{Y})$ and $V_{n,m}(X, Y)$. In each bootstrap test, $B = 200$ bootstrap samples were generated. Each estimate of power is based on 1000 Monte Carlo replicates of the corresponding bootstrap test.

References I

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

V. Alba Fernández, M. D. Jiménez Gamero, and J. Muñoz García. A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics and Data Analysis*, 52:3730–3748, 2008.

A. Gretton, K. M. Borgwadt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13: 723–773, 2012.

P. Hall, F. Lombard, and C. J. Potgieter. A new approach to function-based hypothesis testing in location-scale families. *Technometrics*, 55:215–223, 2013.

References II

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

P. W. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 10, 2011.

S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149, 2007.