ON SPECTRAL EMBEDDING PERFORMANCE AND ELUCIDATING NETWORK STRUCTURE IN STOCHASTIC BLOCKMODEL GRAPHS

Joshua Cape, Minh Tang, Carey E. Priebe

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

We characterize the information-theoretic relative performance of Laplacian spectral embedding (LSE) and adjacency spectral embedding (ASE) for block assignment recovery in stochastic blockmodel graphs via Chernoff information. We investigate the relationship between spectral embedding performance and underlying network structure (e.g., homogeneity, core-periphery, (un)balancedness) via a comprehensive treatment of the two-block stochastic blockmodel and the class of K-block models exhibiting homogeneous balanced affinity structure. Our findings support the claim that, for a particular notion of sparsity, loosely speaking, "Laplacian spectral embedding favors relatively sparse graphs, whereas adjacency spectral embedding favors not-too-sparse graphs." We also provide evidence in support of the claim that "adjacency spectral embedding favors core-periphery network structure."

Background and Overview

Statistical inference on graphs often proceeds via spectral methods involving low-dimensional embeddings of matrix-valued graph representations, such as the graph Laplacian or adjacency matrix. Within the statistics literature, substantial attention has been paid to the class of K-block SBMs with positive-semidefinite block edge probability matrix $\mathbf{B} \in (0,1)^{K \times K}$. This is due in part to the extensive study of the *random dot product graph* (RDPG) *model* [2, 8], a latent position random graph model which includes positive-semidefinite SBMs as a special case. Notably, it was recently shown that for the random dot product graph model, both Laplacian spectral embedding and adjacency spectral embedding behave approximately as random samples from Gaussian mixture models [3, 7]. In tandem with these limit results, the concept of Chernoff information [4] was employed in [7] to demonstrate that neither Laplacian nor adjacency spectral embedding dominates the other for subsequent inference as a spectral embedding method when the underlying inference task is to recover vertices' latent block assignments. In doing so, the results in [7] clarify and complete the groundbreaking work in [6] on normalization in spectral clustering by demonstrating that, for certain blockmodel regimes, K-means clustering is inferior to Gaussian mixture modeling for spectral clustering.

Our work synthesizes recent advances in random graph limit theory [5, 7] in order to extend existing, preliminary Chernoff-based embedding analysis to provide a detailed comparison of two popular spectral embedding procedures. We undertake an extensive investigation of network structure for stochastic blockmodel graphs by considering sub-models exhibiting various functional relationships, symmetries, and geometric properties within the inherent parameter space consisting of block membership probabilities and block edge probabilities. We depict relative spectral embedding performance as a function of the stochastic block model parameter space for a range of sub-models exhibiting various forms of network structure (e.g., homogeneous community structure, core-periphery structure, (un)balanced block sizes).

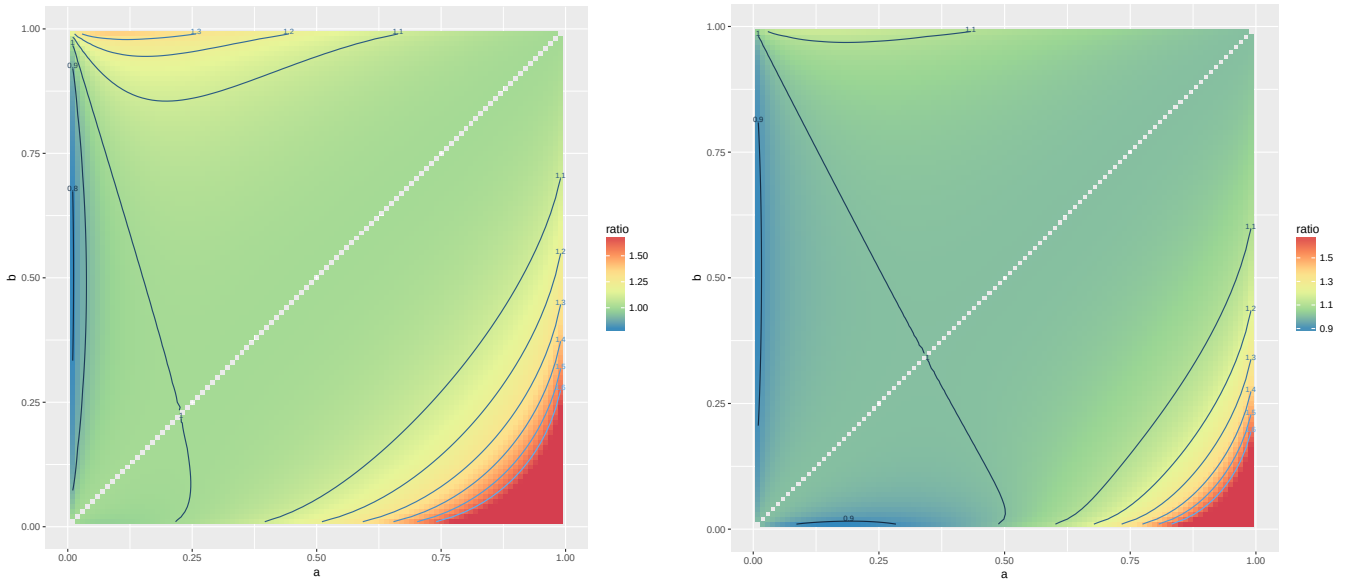
Spectral embedding performance

We focus on the following two models which have garnered widespread interest in the literature—see [1] and the references therein.

1. The two-block SBM with $\mathbf{B} = \left[\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right]$ and $\boldsymbol{\pi} = (\pi_1, 1-\pi_1)^\top$ where $a, b, c, \pi_1 \in (0,1)$;

2. The $K \geq 2$ block SBM exhibiting homogeneous balanced affinity structure, i.e., $\mathbf{B}_{ij} = a$ for all $i = j$, $\mathbf{B}_{ij} = b$ for all $i \neq j$, $0 < b < a < 1$, and $\boldsymbol{\pi} = (\frac{1}{K}, \ldots, \frac{1}{K})^\top$.

Consider large n-vertex graphs from the K-block stochastic blockmodel with symmetric block edge probability matrix $\mathbf{B}$ and block probability vector $\boldsymbol{\pi}$ exhibiting block sizes $n_k = \pi_k n$ for each $k = 1, \ldots, K$. Using the concept of Chernoff information together with recent advances in random graph limit theory, we establish an



(a) The ratio $\rho^\star$ for $\mathbf{B} = \left[\begin{smallmatrix} a & b \\ b & b \end{smallmatrix}\right]$, $\boldsymbol{\pi} = (\frac{1}{2}, \frac{1}{2})^\top$.

(b) The ratio $\rho^\star$ for $\mathbf{B} = \left[\begin{smallmatrix} a & b \\ b & b \end{smallmatrix}\right]$, $\boldsymbol{\pi} = (\frac{1}{4}, \frac{3}{4})^\top$.

Figure 1: Core-periphery network structure and embedding performance in the two-block stochastic blockmodel.

information-theoretic summary characteristic (ratio quantity) $\rho^\star \equiv \rho^\star(\mathbf{B}, \boldsymbol{\pi})$ with the interpretation that the cases $\rho^\star > 1$, $\rho^\star < 1$, and $\rho^\star = 1$ correspond to comparative large-sample embedding performance summarized as ASE $>$ LSE, ASE $<$ LSE, and ASE $=$ LSE, respectively.

Core-periphery network structure

For the collection of two-block SBMs exhibiting core-periphery structure with $\mathbf{B} \equiv \mathbf{B}(a, b)$ as specified in the above sub-captions, Figure 1(a) and Figure 1(b) show $\rho^\star$ evaluated over the parameter space $a, b \in (0, 1)$ in the balanced (block size) regime and in an unbalanced regime, respectively. The empty diagonal depicts the Erdős–Rényi model singularity when $a = b$.

Acknowledgments

This work is partially supported by the XDATA and D3M programs of the Defense Advanced Research Projects Agency (DARPA) and by the Acheson J. Duncan Fund for the Advancement of Research in Statistics at Johns Hopkins University. Part of this work was completed during visits by JC and CEP to the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[2] A. Athreya, D. E. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

[3] A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, 2016.

[4] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23(4):493–507, 1952.

[5] P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2017.

[6] P. Sarkar and P. J. Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *Annals of Statistics*, 43(3):962–990, 2015.

[7] M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Annals of Statistics*, 46(5):2360–2415, 2018.

[8] S. Young and E. Scheinerman. Random dot product graph models for social networks. *Algorithms and Models for the Web-Graph*, pages 138–149, 2007.