On the Effect of Misspecifying the Embedding Dimension in Low-rank Network Models

Roddy Taing and Keith Levin

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

Abstract

As networks have become ubiquitous in the sciences, there has been growing interest in network models whose structure is driven by latent node-level variables in a geometric space. These "latent positions" are often estimated via embeddings, whereby the nodes of a network are mapped to points in Euclidean space so that "similar" nodes are mapped to nearby points. Under certain model assumptions, these embeddings are consistent estimates of the latent positions, but most such results require the embedding dimension to be chosen correctly. Embedding dimension selection methods have been studied extensively, but little is known about the behavior of embeddings when the dimension is misspecified. We provide a theoretical description of the effects of dimension misspecification under the random dot product graph, a class of latent space network models. We show that when the dimension is chosen too large, consistent estimation still holds, albeit at a slower rate than when the embedding dimension is chosen correctly. On the other hand, when the dimension is chosen too small, there is a fundamental estimation error lower bound that need not go to zero in the large-network limit. Our main technical result is a generalization of earlier work showing that all non-signal eigenvectors of a low-rank matrix subject to additive noise are delocalized.

Background

- Latent space models are a popular choice to describe real-world networks.
- These models assume network structure is driven by node-level, latent geometric structure.
- The adjacency spectral embedding (ASE) can be used to estimate these latent positions.
- Constructing the ASE requires that a practitioner choose an embedding dimension, but choosing this dimension from data is known to be difficult.
- It is important to understand the consequences of incorrectly choosing the embedding dimension for the ASE.

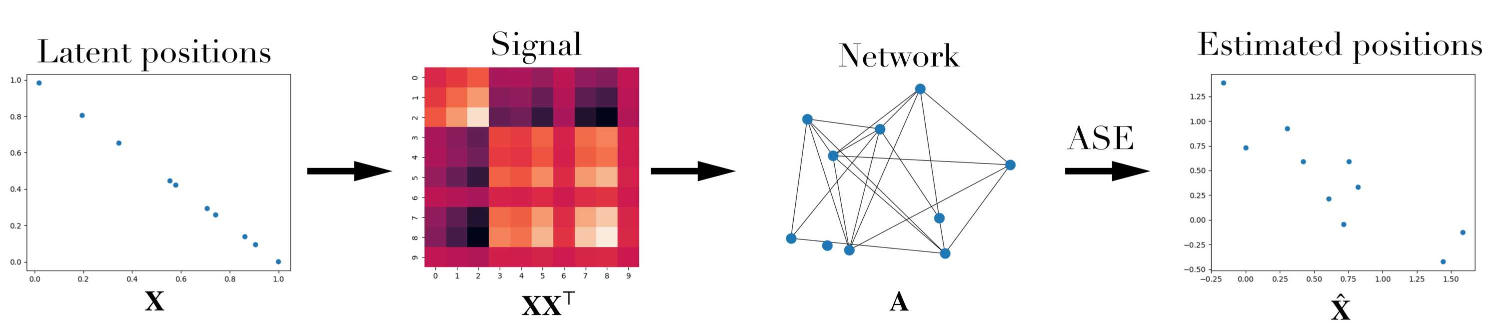


Figure 1. Data generating process and estimation.

The Model

Consider the signal-plus-noise model

$$\mathbf{A} = \rho_n \mathbf{X}\mathbf{X}^\top + \mathbf{E} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times r}$ has rows drawn i.i.d. from some underlying distribution, F, on $\mathbb{R}^r$, and $\rho_n \in [0,1]$ is a signal strength parameter. If, conditional on $\mathbf{X}$, $\mathbf{E}$ has independent and symmetric entries, then we say $\mathbf{A}$ is distributed as an r-dimensional weighted RDPG with latent positions $\mathbf{X}$ and write $(\mathbf{A},\mathbf{X}) \sim \mathrm{RDPG}(F,n)$. Assume the spectral decomposition

$$\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{U}}^\top$$

We estimate $\mathbf{X}$ using the d-dimensional **adjacency spectral embedding (ASE)**:

$$\hat{\mathbf{X}}_{1:d} = \hat{\mathbf{U}}_{1:d}|\hat{\mathbf{S}}|^{1/2}_{1:d} \tag{2}$$

Model Misspecification

- Results showing the consistency of the ASE often assume the true embedding dimension is known. It is estimated in practice.
- Practitioners can choose an incorrect embedding dimension when estimating.
- We show the ASE is consistent, even when the embedding dimension is chosen too large.

Notice by triangle inequality for when $d - r > 0$

$$\left\|\hat{\mathbf{X}}_{1:d} - \rho_n^{1/2}\mathbf{X}_{1:d}\right\|_{2,\infty} \le \underbrace{\left\|\hat{\mathbf{X}}_{1:r} - \rho_n^{1/2}\mathbf{X}_{1:r}\right\|_{2,\infty}}_{\text{Controlled via existing techniques}} + \underbrace{\left\|\hat{\mathbf{X}}_{r+1:d}\right\|_{2,\infty}}_{\text{Goal}},$$

where for some $\mathbf{B}$, $\|\mathbf{B}\|_{2,\infty}$ is the maximum Euclidean norm of the rows of $\mathbf{B}$.

Weighted Networks

Assume entries of $\mathbf{E}$ are independent (up to symmetry) and satisfy the moment conditions

$$\mathbb{E}\left[E_{ij}\right] = 0, \quad \mathbb{E}\left[|E_{ij}|^2\right] = \sigma^2, \quad \mathbb{E}\left[|E_{ij}|^p\right] \le \frac{C_1^p}{(q_n^*)^{(p/2-1)}} \quad (p > 2) \tag{3}$$

where $q_n^* \in \left[n^{-1}(\log n)^6, 1\right]$.

Theorem (Eigenvectors of A are delocalized)

Let $\mathbf{A} = \rho_n\mathbf{X}\mathbf{X}^\top + \mathbf{E}$ be as in Equation (1), and suppose Equation (3) holds. Denote eigenvectors of $\mathbf{A}$ as $\hat{u}$. Under mild conditions on r and ρ_n, the eigenvalues and eigenvectors of $\rho_n\mathbf{X}\mathbf{X}^\top$, we have, for some $\gamma \ge 0$, for all $j \in [n]$,

$$\max_{\alpha > r} |\hat{u}_{j\alpha}| \lesssim \frac{r^2(\log n)^{4+6\gamma}}{\sqrt{n}}. \tag{4}$$

Theorem (Estimation Bounds for Weighted Networks)

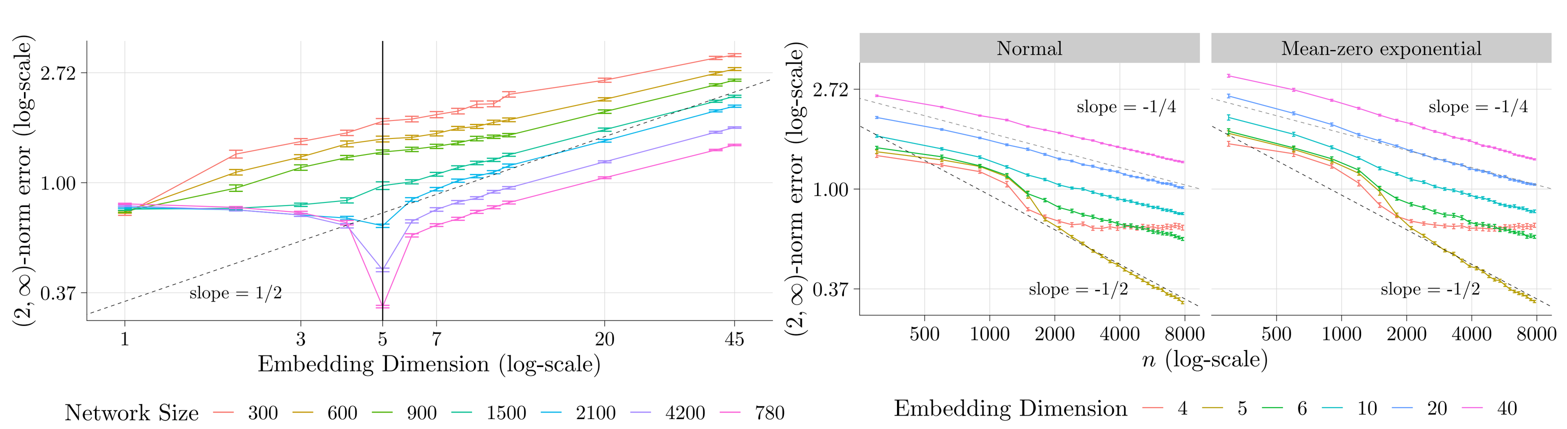
Suppose $(\mathbf{A},\mathbf{X}) \sim \mathrm{RDPG}(F,n)$ is an r-dimensional weighted RDPG with signal strength ρ_n, and suppose Equation (3). Let $\hat{\mathbf{X}}_{1:d}$ be as in Equation (2). Then under mild conditions on r, ρ_n, the eigenvalues and eigenvectors of $\rho_n\mathbf{X}\mathbf{X}^\top$, for $d - r > 0$,

$$\left\|\hat{\mathbf{X}}_{1:d} - \rho_n^{1/2}\mathbf{X}_{1:d}\mathbf{W}\right\|_{2,\infty} \lesssim \left\|\hat{\mathbf{X}}_{1:r}\mathbf{W}^* - \rho_n^{1/2}\mathbf{X}_{1:r}\right\|_{2,\infty} + \sqrt{\sigma^2(d-r)}\frac{r^2(\log n)^{5+6\gamma}}{n^{1/4}}.$$

On the other hand, for $d - r < 0$,

$$\min_{\mathbf{W}\in\mathbb{O}_r} \left\|\left[\hat{\mathbf{X}}_{1:d}\ \mathbf{0}_{d+1:r}\right]\mathbf{W} - \rho_n^{1/2}\mathbf{X}_{1:r}\right\|_{2,\infty} \gtrsim \sqrt{|d-r|\rho_n}.$$

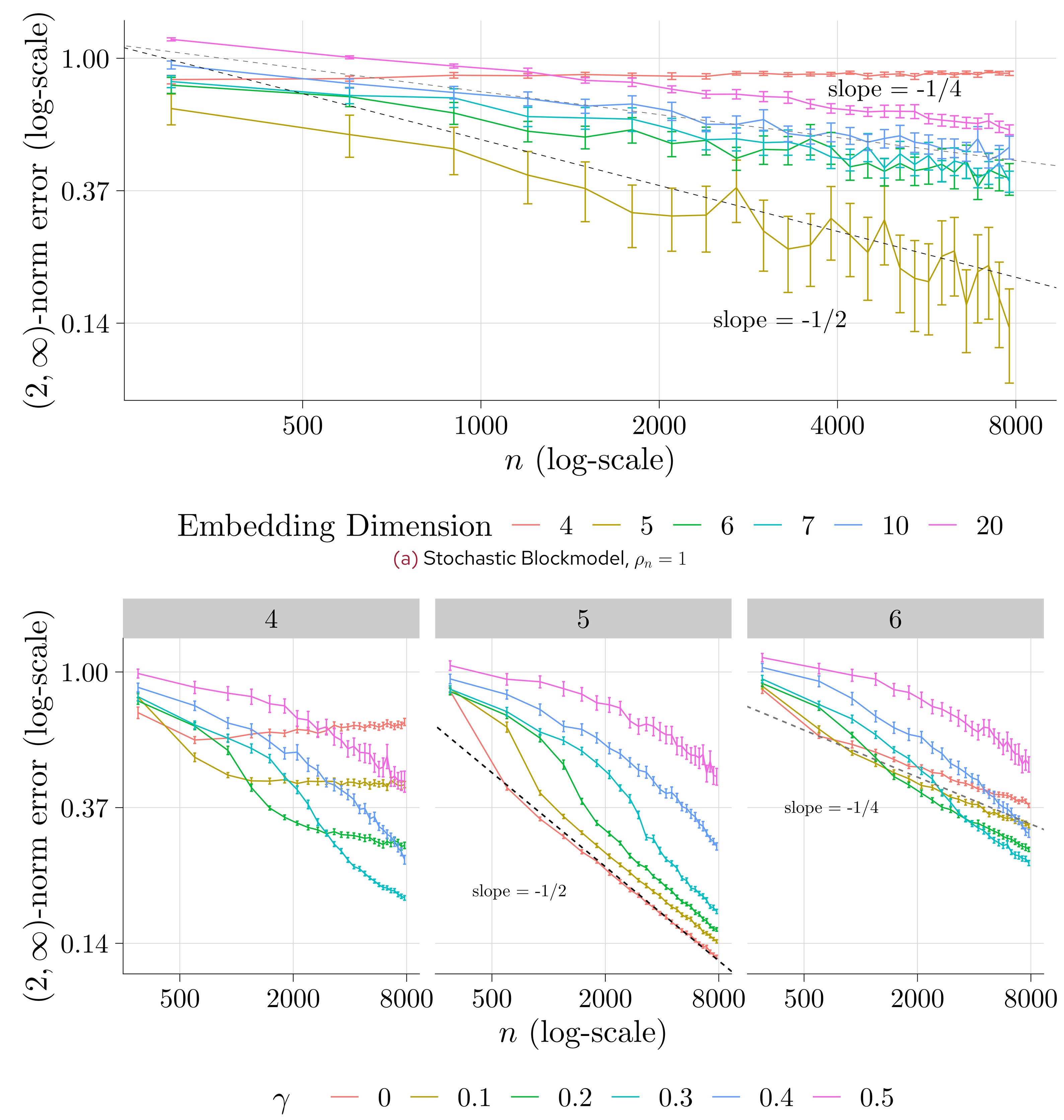
Here, $\mathbf{W}, \mathbf{W}^*$ are orthogonal matrices that account for the non-identifiability in recovering $\mathbf{X}$ from $\mathbf{A}$.



(a) Weighted RDPG, $\rho_n = 1$ (b) Weighted RDPG, $\rho_n = 1$

Figure 2. Estimation error of the ASE in $(2,\infty)$-norm for weighted networks.

Binary Networks and Sparsity



(a) Stochastic Blockmodel, $\rho_n = 1$

(b) Binary RDPG, $\rho_n = n^{-\gamma}$

Figure 3. Estimation error of the ASE in $(2,\infty)$-norm for binary matrices.

Discussion

- Our theory and experiments show that the ASE has three asymptotic behaviors, depending on embedding dimension:
  - When $d - r < 0$, the estimation error of the ASE is lower bounded by the signal strength parameter ρ_n, and it is inconsistent for non-vanishing ρ_n.
  - When $d - r = 0$, consistent estimation at a $n^{-1/2}$ rate (up to log factors), as in (Levin et al., 2022; Rubin-Delanchy et al., 2022).
  - When $d - r > 0$ consistent estimation, albeit at a slower rate, due to delocalization of eigenvectors associated with perturbed null space (Equation (4)). Our theory closes a conjecture in Athreya et al. (2017).
- Our results lend support to the belief that it is better to err on the side of choosing model rank too large.
- Future work: determine if model misspecfication affects downstream tasks (visualization, community detection, and hypothesis testing).

References

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning*, 18(1):8393–8484, 2017.

K. Levin, A. Lodhia, and E. Levina. Recovering shared structure from multiple networks with unknown edge distributions. *Journal of Machine Learning Research*, 23(3):1–48, 2022.

P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4): 1446–1473, 06 2022.