Fitting Networks with a Cancellation Trick

Jiashun Jin (Carnegie Mellon University)
Jingming Wang (University of Virginia)

April 25, 2025

Overview

- Compared with linear network models, nonlinear models may have some additional values, but are much harder to analyze
- We address the problem by discovering a cancellation trick, which can convert a nonlinear model approximately to a linear model
- We showcase this with the logit-DCBM, but the idea may be useful in much broader settings

Social networks

Data. Adjacency matrix for an undirected network with n nodes:

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ and } j \text{ have an edge,} \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(self-edges not allowed)}$$

- There are K non-overlapping communities

$$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$$

- All upper triangular entries of A are independent Bernoulli

$$A = \underbrace{\Omega - \mathrm{diag}(\Omega)}_{\text{signal}} + \underbrace{W}_{\textit{noise}}, \qquad \Omega_{ij} = \mathbb{P}(A_{ij} = 1)$$

The logit-DCBM network model

$$\Omega_{ij} = \frac{1}{1 + \theta_i\theta_j\pi_i'P\pi_j} \cdot \theta_i\theta_j\pi_i'P\pi_j \equiv N_{ij} \cdot \theta_i\theta_j\pi_i'P\pi_j$$

- θ_i: degree heterogeneity parameter
- π_i: community labels

$$\pi_i = (0, \ldots, 0, \underbrace{1,}_{k\text{-th entry}} 0, \ldots, 0)' \iff i \in \mathcal{C}_k$$

- P: baseline connectivity between communities (a $K \times K$ matrix)

A special latent space model (Hoff 2005); reduces to DCBM (Karrer & Newman 2011) if we remove the nonlinear part in red

Two cancellation tricks

Prob 1. If $\Omega_{ij} = \frac{1}{1+\theta_i\theta_j}\theta_i\theta_j$, how to recover θ_i's using Ω?

Trick 1. For any odd integer $m \geq 3$,

$$\frac{\sum_{i_2,\ldots,i_m \in S_{i_1}(dist)} \Omega_{i_1 i_2}(1-\Omega_{i_2 i_3})\ldots\Omega_{i_{m-2} i_{m-1}}(1-\Omega_{i_{m-1} i_m})\Omega_{i_m i_1}}{\sum_{i_2,\ldots,i_m \in S_{i_1}(dist)}(1-\Omega_{i_1 i_2})\Omega_{i_2 i_3}\ldots(1-\Omega_{i_{m-2} i_{m-1}})\Omega_{i_{m-1} i_m}(1-\Omega_{i_m i_1})} = \theta_{i_1}^2.$$

Prob 2. If $\Omega_{ij} = x_0 \cdot \frac{1}{1+\theta_i\theta_j x_0}\theta_i\theta_j$, how to recover x_0 using Ω and θ_i's?

Trick 2.

$$\frac{\sum_{i,j}\Omega_{ij}}{\sum_{ij}\theta_i\theta_j(1-\Omega_{ij})} = x_0$$

Community detection by R-SCORE

$$\Omega_{ij} = \frac{1}{1 + \theta_i\theta_j\pi_i' P\pi_j} \cdot \theta_i\theta_j\pi_i' P\pi_j \equiv N_{ij} \cdot \theta_i\theta_j\pi_i' P\pi_j$$

- **Goal**. Cluster all n nodes into K different community
- **Background**. Without the red part, logit-DCBM reduces to DCBM, in which case, SCORE is an effective algorithm
- **Proposal**. A recursive algorithm called the R-SCORE
  - 1. Apply SCORE by pretending the underlying model is DCBM
  - 2. Fitting all parameters by adapting the two cancellation tricks
  - 3. Obtain an approximated DCBM:

$$\frac{1}{\widehat{N}_{ij}}\,\Omega_{ij} \approx (1 + \theta_i\theta_j\pi_i' P\pi_j)\,\Omega_{ij} = \theta_i\theta_j\pi_i' P\pi_j$$

  - Iterate steps 1-3

Parameter-fitting details

Idea. Let $\widehat{\mathcal{C}}_k$ be the k-th estimated community by SCORE.

- Restrict our attention to each $\widehat{\mathcal{C}}_k$
- Fit parameters by mimicking the cancellation tricks

$$\hat{\theta}_i = \sqrt{\frac{\sum_{j \neq k \in \widehat{\mathcal{C}_{\pi_i}}} A_{ij}(1 - A_{jk})A_{ki}}{\sum_{j \neq k \in \widehat{\mathcal{C}_{\pi_i}}} (1 - A_{ij})A_{jk}(1 - A_{ki})}}$$

$$\widehat{P}_{k\ell} = \frac{\sum_{i \in \widehat{\mathcal{C}}_k, j \in \widehat{\mathcal{C}}_\ell} A_{ij}}{\sum_{i \in \widehat{\mathcal{C}}_k, j \in \widehat{\mathcal{C}}_\ell} \hat{\theta}_i \hat{\theta}_j (1 - A_{ij})}$$

Theoretical Results

Consider Hamming error: $r_n(\widehat{\Pi}) = n^{-1} \min_{\mathcal{P}} \sum_{i=1}^{n} \mathbf{1}_{(\hat{\pi} \neq \mathcal{P}\pi_i)}$

- Rate of SCORE

$$r_n(\widehat{\Pi}^{score}) \lesssim \left[\|(N - \mathbf{1}_n\mathbf{1}_n') \circ \widetilde{\Omega}\|^2 + \lambda_1(\widetilde{\Omega})\right] / \lambda_K^2(\widetilde{\Omega}) =: \delta_n$$

- Rate of SCORE after parameter-fitting

$$r_n(\widehat{\Pi}) \lesssim \left[\|(N \oslash \widehat{N} - \mathbf{1}_n\mathbf{1}_n') \circ \widetilde{\Omega}\|^2 + \tau_n^2 + \lambda_1(\widetilde{\Omega})\right] / \lambda_K^2(\widetilde{\Omega})$$

τ_n relies on the rates of $\widehat{P}$ and $\widehat{N}$, is small under certain regularity conditions.

Theorem (Error Rate of R-SCORE)

Under some regularity conditions, the error rate of R-SCORE is given by

$$r_n(\widehat{\Pi}^{rscore}) \lesssim \left[\lambda_1(\widetilde{\Omega}) + n\bar{\theta}^4 \log(n) + n^2\bar{\theta}^2{\delta_n}^2 + n^2\bar{\theta}^6\delta_n\right] / \lambda_K^2(\widetilde{\Omega})$$

Rate Comparison: SCORE vs. R-SCORE

In the case that $|\lambda_{\min}(P)| \geq C$ and $n\bar{\theta}^4 \to \infty$,

$$r_n(\widehat{\Pi}^{score}) \lesssim \frac{1}{n\bar{\theta}^2} + \bar{\theta}^4 \,,$$

$$r_n(\widehat{\Pi}^{rscore}) \lesssim \frac{1}{n\bar{\theta}^2} + \bar{\theta}^6 + \frac{\log(n)}{n} \,.$$

Let $\bar{\theta} = n^{-\beta}$, where $0 < \beta < 1/2$ (required for SNR $\to \infty$).

$$r_n(\widehat{\Pi}^{score}) \lesssim n^{-a_0(\beta)}, \qquad r_n(\widehat{\Pi}^{rscore}) \lesssim n^{-a_1(\beta)}.$$