Selecting the Number of Communities for Weighted Degree-Corrected Stochastic Block Models

Yucheng Liu Xiaodong Li

University of California, Davis

Introduction

- Weighted networks often reveal more refined community structure than corresponding unweighted ones.
- We want to study how to consistently estimate the number of communities in weighted networks under mild conditions.

Contributions:

1. We propose the weighted DCSBM, which is a generic model for weighted networks without modeling the likelihood.
2. We propose a stepwise testing procedure in selecting the number of communities under the weighted DCSBM and prove the consistency of the procedure under mild conditions.
3. We generalize the Nonsplitting Property of SCORE to weighted DCSBM.
4. Simulations on both synthetic and real-world weighted networks show empirical consistency of our proposed procedure.

Weighted DCSBM

- **Key difference from standard DCSBM:** Instead of specifying the exact likelihood, the model only specifies the first two moments.
- **Parameters:** θ_i is the heterogeneity parameter; $\boldsymbol{\pi}_i \in \mathbb{R}^K$ indicates the community belonging of node i; $\boldsymbol{B}$ is the $K \times K$ symmetric community connectivity matrix.
- **First moment:** $\mathbb{E}[A_{ij}] \coloneqq M_{ij} = \theta_i\theta_j\boldsymbol{\pi}_i^\top\boldsymbol{B}\boldsymbol{\pi}_j$.
- **Second moment:** $V_{ij} \coloneqq \mathrm{var}(A_{ij}) = \nu(M_{ij})$, where ν is known from the underlying distribution.

Estimation of parameters:

- Node community belonging is estimated by SCORE.
- $\hat{\theta}_i^{(m)} \coloneqq \frac{\sqrt{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A}\hat{\mathbf{1}}_k^{(m)}}}{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A}\mathbf{1}_n} d_i$, where d_i is the degree of node i.
- $\widehat{B}_{kl}^{(m)} \coloneqq \frac{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A}\hat{\mathbf{1}}_l^{(m)}}{\sqrt{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A}\hat{\mathbf{1}}_k^{(m)}}\sqrt{(\hat{\mathbf{1}}_l^{(m)})^\top \boldsymbol{A}\hat{\mathbf{1}}_l^{(m)}}}$.

Assumption 1

Denote $\theta_{\max} = \max\{\theta_1, \dots, \theta_n\}$ and $\theta_{\min} = \min\{\theta_1, \dots, \theta_n\}$. c_0 is a small constant. We assume the following conditions hold:

- [Fixed rank] The true number of communities K is fixed.
- [Balancedness]
$$\min_{1\le k\le K} \frac{n_k}{n} \ge c_0 \quad \text{and} \quad \frac{\theta_{\min}}{\theta_{\max}} \ge c_0.$$
- [Sparseness]
$$\frac{1}{c_0} \ge \theta_{\max} \ge \theta_{\min} \ge \frac{\log^3 n}{\sqrt{n}}.$$
- [Community connectivity] The $K \times K$ matrix $\boldsymbol{B}$ is fixed, and its entries and eigenvalues satisfy
$$\begin{cases} B_{kk} = 1 \quad \text{for} \;\; k = 1, \dots, K, \\ c_0 \le B_{kl} \le 1 \;\; \text{for} \; 1 \le k, l \le K, \\ \lambda_1(\boldsymbol{B}) > |\lambda_2(\boldsymbol{B})| \ge \cdots \ge |\lambda_K(\boldsymbol{B})| \ge c_0 > 0. \end{cases}$$
- [Variance-mean function] The function $\nu(\cdot)$ satisfies
$$c_0\mu \le \nu(\mu) \le \mu/c_0 \quad \text{and} \quad \nu(\cdot) \text{ is } 1/c_0\text{-Lipschitz}.$$
- [Bernstein condition] For any $i \le j$ and any integer $p \ge 2$, there holds
$$\mathbb{E}\left[|A_{ij} - M_{ij}|^p\right] \le \binom{p!}{2} R(c_0)^{p-2}\nu(M_{ij}),$$
where $R(c_0)$ is a constant only depending on c_0.

Our Algorithm

SVPS: Stepwise Variance Profile Scaling

For $m = 1, 2, \dots$:

1. Group nodes into m distinct communities using SCORE.
2. Obtain estimated mean adjacency matrix $\widehat{\boldsymbol{M}}^{(m)}$ by fitting DCSBM and derive the estimated variance profile matrix $\widehat{\boldsymbol{V}}^{(m)}$ using the variance-mean relationship.
3. Find the scaling matrix $\widehat{\boldsymbol{\Psi}}^{(m)}$ such that $\widehat{\boldsymbol{\Psi}}^{(m)}\widehat{\boldsymbol{V}}^{(m)}\widehat{\boldsymbol{\Psi}}^{(m)}$ is doubly stochastic (every row sum equals 1).
4. Obtain test statistic $T_{n,m} = \left|\lambda_{m+1}\left(\left(\widehat{\boldsymbol{\Psi}}^{(m)}\right)^{\frac{1}{2}}\boldsymbol{A}\left(\widehat{\boldsymbol{\Psi}}^{(m)}\right)^{\frac{1}{2}}\right)\right|$.

   Stop the procedure if $T_{n,m} < 2 + \epsilon$ and obtain $\widehat{K} = m$.

Theoretical Results

Theorem (Null Case). If we implement SVPS with $m = K$ and SCORE for spectral clustering, then for any fixed $c_0 > 0$ in Assumption 1, as $n \to \infty$, we have $T_{n,m} \le 2 + o_P(1)$.

Theorem (Underfitting Case). If we implement SVPS with $m < K$ and SCORE for spectral clustering, then for any fixed $c_0 > 0$ in Assumption 1, as $n \to \infty$, we have $T_{n,m} \xrightarrow{P} \infty$.

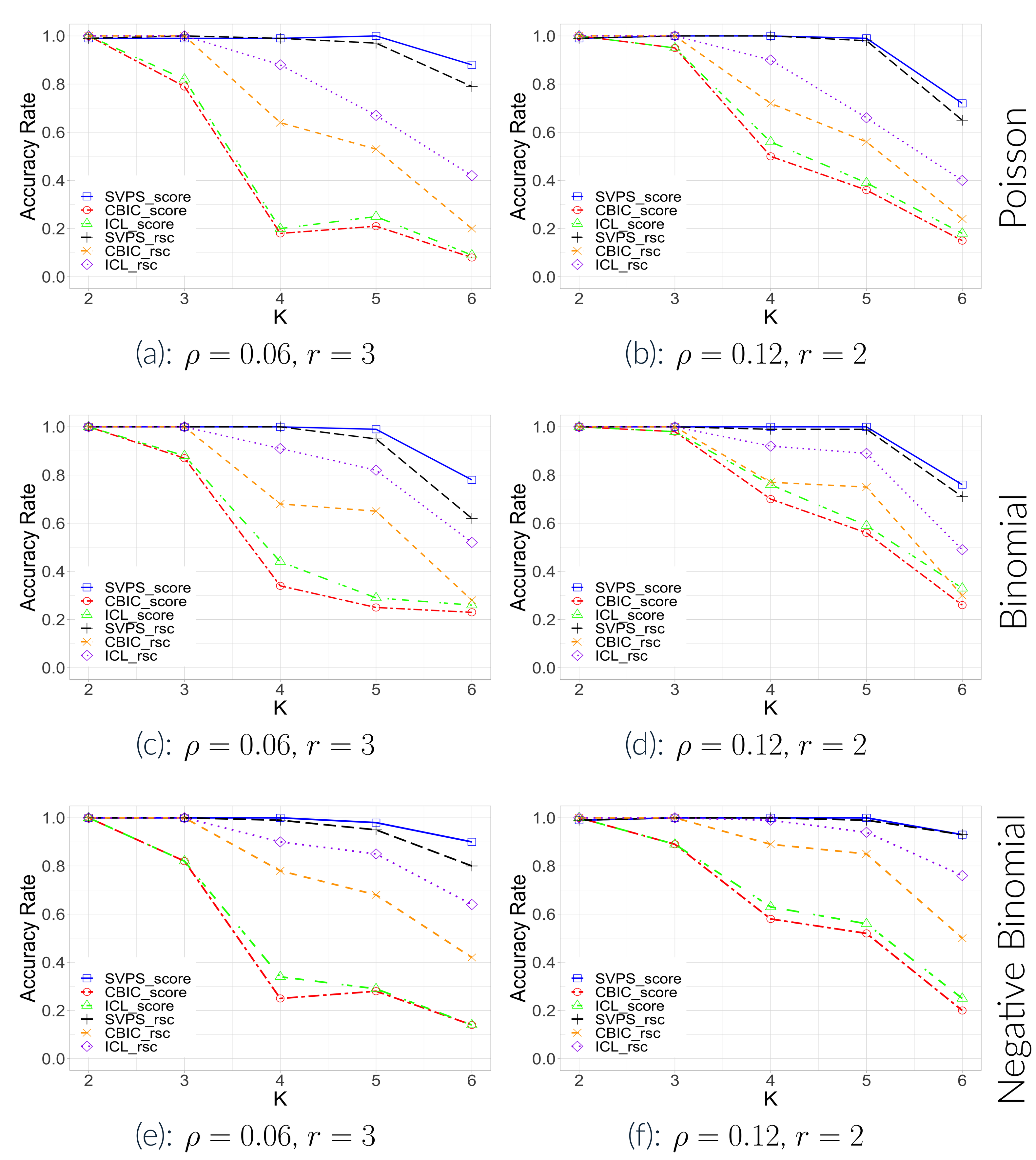
Definition (Nonsplitting Property (Jin et al., 2022)). The estimated communities of a network satisfy the Nonsplitting Property if the true communities are a refinement of the estimated ones.

Lemma (Nonsplitting Property). Under Assumption 1, with any fixed $c_0 > 0$, for any fixed $m \le K$, SCORE satisfies the NSP with probability $1 - O(n^{-3})$.

Experiments

Generating mechanism of synthetic networks:

- Underlying generating distributions: Poisson, binomial and negative binomial.
- $B_{kl} = \rho\left(1 + r \times \mathbf{1}_{\{k=l\}}\right)$.



(a): $\rho = 0.06$, $r = 3$ (b): $\rho = 0.12$, $r = 2$

(c): $\rho = 0.06$, $r = 3$ (d): $\rho = 0.12$, $r = 2$

(e): $\rho = 0.06$, $r = 3$ (f): $\rho = 0.12$, $r = 2$

Figure 1: Accuracy rate of SVPS and other methods for comparison.