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

Yucheng Liu
University of California, Davis

Xiaodong Li
University of California, Davis

Abstract

We investigate how to select the number of communities for weighted networks without a full likelihood modeling. First, we propose a novel weighted degree-corrected stochastic block model (DCSBM), where the mean adjacency matrix is modeled in the same way as in the standard DCSBM, while the variance profile matrix is assumed to be related to the mean adjacency matrix through a given variance function. Our method of selecting the number of communities is based on a sequential testing framework. In each step, the weighted DCSBM is fitted via some spectral clustering method. A key component of our method is matrix scaling on the estimated variance profile matrix. The resulting scaling factors can be used to normalize the adjacency matrix, from which the test statistic is then obtained. Under mild conditions on the weighted DCSBM, our proposed procedure is shown to be consistent in estimating the true number of communities. Numerical experiments on both simulated and real-world network data demonstrate the desirable empirical properties of our method.

1 INTRODUCTION

In network analysis, community detection is regarding how to partition the vertices of a graph into clusters with similar connection patterns. This problem has numerous applications in a variety of fields, including applied physics, sociology, economics, and biology; see the survey paper Fortunato (2010). An important problem in community detection is selecting the number of communities. To answer this question, various methods have been proposed for unweighted networks, such as sequential testing (Lei, 2016; Jin et al., 2022; Han et al., 2023), spectral thresholding (Le and Levina, 2022), and model comparison based methods (Daudin et al., 2008; Wang and Bickel, 2017; Saldana et al., 2017; Chen and Lei, 2018; Li et al., 2020; Hu et al., 2020; Ma et al., 2021), etc. In particular, sequential testing methods are usually based on goodness-of-fit testing methods (e.g., Gao and Lafferty, 2017; Jin et al., 2018, 2021; Hu et al., 2021).

Most existing methods for selecting the number of communities are only applicable to unweighted networks. In contrast, the problem of determining the number of communities for weighted networks is little studied. In fact, weighted networks are common in practice and often reveal more refined community structures; see, for example, Newman and Girvan (2004), Ball et al. (2011) and Newman and Reinert (2016). A challenge for rank selection in weighted networks lies in the modeling of the networks, since likelihood based methods could be very restrictive for this problem.

In this paper, we propose a generic model for weighted networks without modeling the likelihood. To be specific, we only model the mean structure in the form of the degree-corrected stochastic block model (DCSBM) as in Karrer and Newman (2011), and the entrywise variances as functions of the corresponding means. Due to its connection to the standard DCSBM, our proposed model is referred to as the weighted DCSBM. To see why the variance profile matrix is helpful for selecting the number of communities, note that the adjacency matrix can be decomposed into $\boldsymbol{A} = \boldsymbol{M} + (\boldsymbol{A} - \boldsymbol{M})$, where $\boldsymbol{M}$ is the mean adjacency matrix. If we denote K as the number of communities, then $\boldsymbol{M}$ is of rank K, and $\boldsymbol{A}$ can be viewed as a perturbation of a rank-K matrix. Denote the nonzero eigenvalues of $\boldsymbol{M}$ as $|\lambda_1(\boldsymbol{M})| \geq \cdots \geq |\lambda_K(\boldsymbol{M})| > 0$, and all eigenvalues of $\boldsymbol{A}$ as $|\lambda_1(\boldsymbol{A})| \geq \cdots \geq |\lambda_n(\boldsymbol{A})|$. Assume that we know an explicit and tight upper bound $\|\boldsymbol{A}-\boldsymbol{M}\| \leq \tau$. Then, we have $|\lambda_{K+1}(\boldsymbol{A})| \leq \|\boldsymbol{A} - \boldsymbol{M}\| \leq \tau$. Furthermore,

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

if we assume the signal-to-noise ratio of the weighted DCSBM is large enough, which is $|\lambda_K(\boldsymbol{M})| \gg \tau$, then $|\lambda_K(\boldsymbol{A})| \geq |\lambda_K(\boldsymbol{M})| - \|\boldsymbol{A}-\boldsymbol{M}\| \gg \tau$. Then in step $m = 1, 2, ...$ of a sequential testing procedure, we can choose the test statistic as $T_{n,m} = |\lambda_{m+1}(\boldsymbol{A})|$, which satisfies $T_{n,m} \gg \tau$ for $m \leq K-1$, and $T_{n,m} \leq \tau$ for $m = K$. Consequently, we can use $T_{n,m} \leq \tau$ as the stopping rule to select the number of communities.

However, such an explicit and tight bound $\|\boldsymbol{A}-\boldsymbol{M}\| \leq \tau$ is generally not available. One natural idea is to consider a normalized version of the noise matrix, which allows us to derive a tight and explicit upper bound for its spectral norm. Here we borrow an idea from Landa et al. (2022), which studies how to reveal the rank of a Poisson data matrix by spectral truncation with the assistance of matrix scaling on the variance profile matrix. Returning to our problem, if the variance profile matrix $\boldsymbol{V}$ is known, its entries are $V_{ij} = \mathrm{var}(A_{ij}) = \mathrm{var}(A_{ij} - M_{ij})$. It is known that there exists a diagonal matrix $\boldsymbol{\Psi} = \mathrm{Diag}(\psi_1, \ldots, \psi_n)$, such that $\boldsymbol{\Psi}\boldsymbol{V}\boldsymbol{\Psi}$ is doubly stochastic, i.e., every row sum of $\boldsymbol{\Psi}\boldsymbol{V}\boldsymbol{\Psi}$ is 1 (e.g., Knight et al., 2014). With this diagonal scaling matrix $\boldsymbol{\Psi}$, recent results (e.g., Latała et al., 2018) in random matrix theory guarantee that $\left\|\boldsymbol{\Psi}^{\frac{1}{2}}(\boldsymbol{A}-\boldsymbol{M})\boldsymbol{\Psi}^{\frac{1}{2}}\right\| \leq 2+\epsilon$ for some small positive constant ϵ.

Based on the above heuristic, we propose the following sequential testing procedure. For each $m = 1, 2, 3, \ldots$, we first group the nodes into m distinct communities using some spectral clustering method, e.g., SCORE proposed in Jin (2015) or the regularized spectral clustering (RSC) procedure proposed in Amini et al. (2013). With the estimated groups, we obtain the estimated mean adjacency matrix $\widehat{\boldsymbol{M}}^{(m)}$ by fitting the DCSBM, and further derive the estimated variance profile matrix $\widehat{\boldsymbol{V}}^{(m)}$ using the variance-mean relationship. Next, we find a diagonal scaling matrix $\widehat{\boldsymbol{\Psi}}^{(m)}$ such that $\widehat{\boldsymbol{\Psi}}^{(m)}\widehat{\boldsymbol{V}}^{(m)}\widehat{\boldsymbol{\Psi}}^{(m)}$ is doubly stochastic. The test statistic is defined as $T_{n,m} = \left|\lambda_{m+1}\left(\left(\widehat{\boldsymbol{\Psi}}^{(m)}\right)^{\frac{1}{2}}\boldsymbol{A}\left(\widehat{\boldsymbol{\Psi}}^{(m)}\right)^{\frac{1}{2}}\right)\right|$. For each m, we stop the iterative procedure if $T_{n,m} < 2+\epsilon$, and then report the estimated number of communities as $\widehat{K} = m$.

1.1 Related Work

The stochastic block model (SBM) by Holland et al. (1983) and its variants, such as the degree-corrected stochastic block model (DCSBM) by Karrer and Newman (2011), have been widely used to model community structures in networks. Many community detection methods have been proposed in the literature, with examples including modularity and likelihood based methods (e.g., Newman and Girvan, 2004; Bickel and Chen, 2009; Zhao et al., 2012; Amini et al., 2013; Bickel et al., 2013; Le et al., 2016; Zhang and Zhou, 2016, 2020), spectral methods (e.g., Rohe et al., 2011; Amini et al., 2013; Jin, 2015; Lei and Rinaldo, 2015; Joseph and Yu, 2016; Gulikers et al., 2017; Gao et al., 2018; Abbe et al., 2020; Zhang et al., 2020; Deng et al., 2021; Li et al., 2022; Lei et al., 2020), convex optimization methods (e.g., Cai and Li, 2015; Guédon and Vershynin, 2015; Abbe et al., 2015; Chen et al., 2018; Li et al., 2021), and many other methods such as Bayesian approaches.

Matrix scaling has been used in Landa et al. (2022) to reveal the rank of a Poisson data matrix, where the variance profile matrix is the same as the mean adjacency matrix and thus can be estimated by the observed count data matrix. In contrast, our work is not restricted to the Poisson data matrix. In fact, we have a parametric model for the mean structure in our weighted DCSBM, whereas there is no such parametric mean modeling in Landa et al. (2022).

Sequential testing based on stepwise SBM/DCSBM fitting has been commonly used in the literature of rank selection for standard binary networks (e.g., Lei, 2016; Wang and Bickel, 2017; Hu et al., 2020; Ma et al., 2021; Hu et al., 2021; Jin et al., 2022). However, these methods either rely on the binary structure of the unweighted network or on modeling the complete likelihood of the random network, and thus cannot be straightforwardly extended to our setup.

To address some technical challenges in the underfitting case $m < K$, we generalize the Nonsplitting Property established in Jin et al. (2022) for the spectral clustering approach SCORE proposed in Jin (2015) to the weighted DCSBM. This is the primary reason why our main results are established using SCORE for spectral clustering, although extensions to other spectral methods may be possible, which we leave for future investigation.

1.2 Organization of the Paper

This paper is organized as follows. In Section 2, we extend the standard DCSBM to model the mean structure of weighted networks, and link the mean adjacency matrix and the variance profile matrix with a variance function. Section 3 introduces a stepwise DCSBM fitting procedure with a spectral test statistic based on variance profile matrix scaling. The main results on the consistency of our procedure are presented in Section 4 under certain assumptions on the weighted DCSBM. In Section 5, we demonstrate the empirical properties of our proposed procedure with both synthetic and real-world weighted networks. The proofs of our main results are given in Appendix B. Prelim-

inary results and proofs of the technical lemmas are also included in the supplementary material.

2 WEIGHTED DEGREE-CORRECTED STOCHASTIC BLOCK MODELS

Our first task is to extend the standard DCSBM proposed in Karrer and Newman (2011) to a weighted DCSBM which accommodates networks with nonnegative weights. The edges in our weighted DCSBM are assumed to be independent. Rather than specifying the exact likelihood of the random graph with weighted edges as in the standard DCSBM, the weighted DCSBM relies on specifying the first two moments, namely, the mean adjacency matrix and the variance profile matrix.

We begin by introducing some notation for the network. Let $\boldsymbol{A}$ be the adjacency matrix of the weighted network with n nodes, which belong to K separate communities. Denote $\mathcal{N}_1, \ldots, \mathcal{N}_K$ as the underlying communities, with respective cardinalities $n_1, \ldots, n_K$, and thereby $n = n_1 + \cdots + n_K$. Let $\phi : [n] \to [K]$ be the community membership function of node, such that $\phi(i) = k$ if and only if node i belongs to community $\mathcal{N}_k$. We also use the vector $\boldsymbol{\pi}_i \in \mathbb{R}^K$ to indicate the community belonging of node i by letting $\pi_i(k) = 1$ if $i \in \mathcal{N}_k$ and $\pi_i(k) = 0$ otherwise. Denote $\boldsymbol{\Pi} = [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n]^\top \in \mathbb{R}^{n \times K}$ as the community membership matrix.

The mean adjacency matrix is modeled similarly as in the standard DCSBM. Note that the network is allowed to have self-loops for simplicity of analysis without loss of generality. Specifically, let $\boldsymbol{B}$ be a $K \times K$ symmetric community connectivity matrix with positive entries and diagonal entries $B_{kk} = 1$ for $k = 1, \ldots, K$ to ensure identifiability. Also, let $\theta_1, \ldots, \theta_n > 0$ be the heterogeneity parameters. Denote $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ and $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_n)$. Then, the entries of the expected mean matrix are parameterized as

$$\mathbb{E}[A_{ij}] \coloneqq M_{ij} \coloneqq \theta_i \theta_j B_{\phi(i)\phi(j)} = \theta_i \theta_j \boldsymbol{\pi}_i^\top \boldsymbol{B} \boldsymbol{\pi}_j, \quad (1)$$

for $1 \le i \le j \le n$. In matrix form, the mean adjacency matrix can be represented as

$$\boldsymbol{M} = \boldsymbol{\Theta}\boldsymbol{\Pi}\boldsymbol{B}\boldsymbol{\Pi}^\top\boldsymbol{\Theta}. \quad (2)$$

In modeling the second order moments, we denote $V_{ij} = \mathrm{var}(A_{ij})$ and assume that it is a function of the corresponding mean by some variance function $\nu(\cdot)$, i.e., $V_{ij} = \nu(M_{ij})$. By denoting the variance profile matrix $\boldsymbol{V} = [V_{ij}]_{i,j=1}^n$, we can also represent the variance-mean relationship as $\boldsymbol{V} = \nu(\boldsymbol{M})$, where ν is viewed as an entrywise operation.

The variance function $\nu(\cdot)$ is known for some common distributions, such as $\nu(\mu) = \mu(1-\mu)$ for the Bernoulli case and $\nu(\mu) = \mu$ for the Poisson case. In the present work, we assume ν is known. A potential direction for future study is how to estimate ν in a data-dependent manner, but we leave this for future investigation.

3 METHODOLOGY

In this section, we will outline the details of our stepwise procedure for estimating the number of communities in weighted networks, referred to as Stepwise Variance Profile Scaling (SVPS). SVPS is a stepwise method that relies on fitting the weighted DSCBM with an increasing sequence of candidate numbers of communities. In other words, for each positive integer $m = 1, 2, \ldots$, we aim to test $H_0 : K = m$, where K is the true rank of the weighted DCSBM.

3.1 Stepwise Weighted DCSBM Fitting

With each hypothetical number of communities m, we apply a standard community detection method, such as SCORE (Jin, 2015) or RSC (Amini et al., 2013; Joseph and Yu, 2016), to obtain m estimated communities $\widehat{\mathcal{N}}_1^{(m)}, \ldots, \widehat{\mathcal{N}}_m^{(m)}$. Subsequently, the DCSBM parameters $\boldsymbol{\theta}$, $\boldsymbol{B}$ and $\boldsymbol{W}$ can be directly estimated in a plug-in manner.

The estimates of $\boldsymbol{\theta}$ and $\boldsymbol{B}$ are the same as in the standard DCSBM, and here we give a review of their derivations in Jin et al. (2022). Let us first see how to represent $\boldsymbol{\theta}$ and $\boldsymbol{B}$ with the mean adjacency matrix $\boldsymbol{M}$, the true communities $\mathcal{N}_1, \ldots, \mathcal{N}_K$, and the expected degrees. Decompose $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = \boldsymbol{\theta}_1 + \cdots + \boldsymbol{\theta}_K$, where $\boldsymbol{\theta}_k \in \mathbb{R}^n$ for $k = 1, \ldots, K$ such that $\boldsymbol{\theta}_k(i) = \theta_i$ if $i \in \mathcal{N}_k$ and $\boldsymbol{\theta}_k(i) = 0$ otherwise. We can similarly decompose the n-dimensional all-one vector into $\mathbf{1}_n = \mathbf{1}_1 + \cdots + \mathbf{1}_K$ such that $\mathbf{1}_k(j) = 1$ if $j \in \mathcal{N}_k$ and $\mathbf{1}_k(j) = 0$ otherwise. It is easy to verify that for $1 \le k, l \le K$, $\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_l = B_{kl} \|\boldsymbol{\theta}_k\|_1 \|\boldsymbol{\theta}_l\|_1$. By the assumption $B_{kk} = 1$ for $k = 1, \ldots, K$, the above equality implies $\|\boldsymbol{\theta}_k\|_1 = \sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}$, which further gives

$$B_{kl} = \frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_l}{\|\boldsymbol{\theta}_k\|_1 \|\boldsymbol{\theta}_l\|_1} = \frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_l}{\sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k} \sqrt{\mathbf{1}_l^\top \boldsymbol{M} \mathbf{1}_l}}. \quad (3)$$

Denote the degree of node $i \in \mathcal{N}_k$ as $d_i = \sum_{j=1}^n A_{ij}$. Its expectation, referred to as the population degree, is given by

$$\begin{aligned} d_i^* \coloneqq \mathbb{E}[d_i] &= \theta_i \left(B_{k1} \|\boldsymbol{\theta}_1\|_1 + B_{k2} \|\boldsymbol{\theta}_2\|_1 + \cdots + B_{kK} \|\boldsymbol{\theta}_K\|_1 \right) \\ &= \theta_i \left(\frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_1}{\sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}} + \cdots + \frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_K}{\sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}} \right) = \theta_i \frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_n}{\sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}}. \end{aligned}$$

This implies that the degree-correction parameter θ_i can be expressed as

$$\theta_i = \frac{\sqrt{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}}{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_n} d_i^*, \quad i \in \mathcal{N}_k. \tag{4}$$

With (3) and (4), we obtain plug-in estimates of $\boldsymbol{\theta}$ and $\boldsymbol{B}$ by replacing the true community partition $\mathcal{N}_1, \ldots, \mathcal{N}_K$ with the estimated partition $\widehat{\mathcal{N}}_1^{(m)}, \ldots, \widehat{\mathcal{N}}_m^{(m)}$, replacing $\boldsymbol{M}$ with $\boldsymbol{A}$, and replacing d_i^* with d_i. In analogy to $\mathbf{1}_n = \mathbf{1}_1 + \cdots + \mathbf{1}_K$, we decompose the all-one vector as the sum of indicator vectors corresponding to the estimated communities: $\mathbf{1}_n = \hat{\mathbf{1}}_1^{(m)} + \cdots + \hat{\mathbf{1}}_m^{(m)}$, where for each $j = 1, \ldots, n$ and $k = 1, \ldots, m$, $\hat{\mathbf{1}}_k^{(m)}(j) = 1$ if $j \in \widehat{\mathcal{N}}_k^{(m)}$ and $\hat{\mathbf{1}}_k^{(m)}(j) = 0$ otherwise. Then, the plug-in estimates are

$$\hat{\theta}_i^{(m)} \coloneqq \frac{\sqrt{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_k^{(m)}}}{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \mathbf{1}_n} d_i, \ k = 1, \ldots, m \text{ and } i \in \widehat{\mathcal{N}}_k^{(m)}, \tag{5}$$

and

$$\begin{aligned} \widehat{B}_{kl}^{(m)} &\coloneqq \frac{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_l^{(m)}}{\|\hat{\boldsymbol{\theta}}_k^{(m)}\|_1 \|\hat{\boldsymbol{\theta}}_l^{(m)}\|_1} \\ &= \frac{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_l^{(m)}}{\sqrt{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_k^{(m)}} \sqrt{(\hat{\mathbf{1}}_l^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_l^{(m)}}}, \ 1 \le k, l \le m. \end{aligned} \tag{6}$$

Then we have the estimated mean matrix

$$\widehat{M}_{ij}^{(m)} \coloneqq \hat{\theta}_i^{(m)} \hat{\theta}_j^{(m)} \widehat{B}_{kl}^{(m)} = \frac{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \hat{\mathbf{1}}_l^{(m)}}{(\hat{\mathbf{1}}_k^{(m)})^\top \boldsymbol{A} \mathbf{1}_n \cdot (\hat{\mathbf{1}}_l^{(m)})^\top \boldsymbol{A} \mathbf{1}_n} d_i d_j, \tag{7}$$

for any $1 \le i, j \le n$ with $i \in \widehat{\mathcal{N}}_k^{(m)}$ and $j \in \widehat{\mathcal{N}}_l^{(m)}$. By denoting the community membership matrix corresponding to $\widehat{\mathcal{N}}_1^{(m)}, \ldots, \widehat{\mathcal{N}}_m^{(m)}$ as $\widehat{\boldsymbol{\Pi}}^{(m)} \in \mathbb{R}^{n \times m}$, the matrix form of mean estimation is (omitting the superscripts)

$$\widehat{\boldsymbol{M}}^{(m)} = \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\Pi}} \widehat{\boldsymbol{B}} \widehat{\boldsymbol{\Pi}}^\top \widehat{\boldsymbol{\Theta}}. \tag{8}$$

With the estimated mean matrix $\widehat{\boldsymbol{M}}^{(m)}$, we have the residual matrix $\boldsymbol{R}^{(m)} = \boldsymbol{A} - \widehat{\boldsymbol{M}}^{(m)}$ and the estimated variance profile matrix $\widehat{\boldsymbol{V}}^{(m)} = \nu(\widehat{\boldsymbol{M}}^{(m)})$. In the case where the variance function is unknown, we substitute $\nu(\cdot)$ with the estimated variance function $\widehat{\nu}(\cdot)$.

3.2 Variance Profile Scaling and Spectral Statistic

In Section 1, we have introduced how to define the spectral test statistic based on variance profile matrix scaling. Here, we briefly introduce it again. For the variance profile matrix $\boldsymbol{V}$, assume $\boldsymbol{\Psi} = \mathrm{Diag}(\boldsymbol{\psi}) = \mathrm{Diag}(\psi_1, \ldots, \psi_n)$ is a diagonal matrix such that $\boldsymbol{\Psi V \Psi}$ is doubly stochastic, i.e.,

$$\sum_{i=1}^n V_{ij} \psi_i \psi_j = 1, \quad \forall j = 1, \ldots, n. \tag{9}$$

The uniqueness and existence of such $\boldsymbol{\Psi}$ and related algorithms are well-known in the literature; see, e.g., Sinkhorn (1967); Knight et al. (2014).

Analogously, in the m-th step of SVPS, we choose the scaling factors $\hat{\psi}_1^{(m)}, \ldots, \hat{\psi}_n^{(m)}$, such that $\widehat{\boldsymbol{\Psi}}^{(m)} \widehat{\boldsymbol{V}}^{(m)} \widehat{\boldsymbol{\Psi}}^{(m)}$ is doubly stochastic, where $\widehat{\boldsymbol{\Psi}}^{(m)} = \mathrm{Diag}(\hat{\psi}_1^{(m)}, \ldots, \hat{\psi}_n^{(m)})$. In other words,

$$\sum_{i=1}^n \widehat{V}_{ij} \hat{\psi}_i \hat{\psi}_j = 1, \quad \forall j = 1, \ldots, n. \tag{10}$$

Define the test statistic as

$$T_{n,m} = \left| \lambda_{m+1} \left(\left(\widehat{\boldsymbol{\Psi}}^{(m)} \right)^{\frac{1}{2}} \boldsymbol{A} \left(\widehat{\boldsymbol{\Psi}}^{(m)} \right)^{\frac{1}{2}} \right) \right|.$$

For $m = 1, 2, \ldots$, we stop the iterative procedure if $T_{n,m} < 2 + \epsilon$ for some prespecified small constant ϵ, and then obtain the estimated number of communities as $\widehat{K} = m$.

4 MAIN RESULTS

In this section, we aim to establish the consistency of SVPS in selecting the number of communities under the weighted DCSBM. Our main results consist of two parts: (1) in the null case $m = K$, we will show that $T_{n,m} \le 2 + o_p(1)$; (2) in the underfitting case $m < K$, we will show that $T_{n,m} \gg O_p(1)$. Before presenting the main results, we first introduce a sequence of conditions on the weighted DCSBM under which our consistency analysis is conducted.

4.1 Assumptions

Note that the conditions imposed on the weighted DCSBM are captured by a constant c_0.

Assumption 1. Consider the weighted DCSBM described in Section 2. Denote $\theta_{\max} = \max\{\theta_1, \ldots, \theta_n\}$ and $\theta_{\min} = \min\{\theta_1, \ldots, \theta_n\}$. Assume the following conditions hold:

- [Fixed rank] The true number of communities K is fixed.
- [Balancedness]

$$\min_{1 \le k \le K} \frac{n_k}{n} \ge c_0 \quad \text{and} \quad \frac{\theta_{\min}}{\theta_{\max}} \ge c_0. \tag{11}$$

- [Sparseness]

$$\frac{1}{c_0} \geq \theta_{\max} \geq \theta_{\min} \geq \frac{\log^3 n}{\sqrt{n}}. \tag{12}$$

- [Community connectivity] The $K \times K$ matrix $\boldsymbol{B}$ is fixed, and its entries and eigenvalues satisfy

$$\begin{cases} B_{kk} = 1 \quad \text{for} \quad k = 1, \ldots, K, \\ c_0 \leq B_{kl} \leq 1 \quad \text{for } 1 \leq k, l \leq K, \\ \lambda_1(\boldsymbol{B}) > |\lambda_2(\boldsymbol{B})| \geq \cdots \geq |\lambda_K(\boldsymbol{B})| \geq c_0 > 0. \end{cases} \tag{13}$$

- [Variance-mean function] The function $\nu(\cdot)$ satisfies

$$c_0\mu \leq \nu(\mu) \leq \mu/c_0 \quad \text{and} \quad \nu(\cdot) \text{ is } 1/c_0\text{-Lipschitz}. \tag{14}$$

- [Bernstein condition] For any $i \leq j$ and any integer $p \geq 2$, there holds

$$\mathbb{E}\left[|A_{ij} - M_{ij}|^p\right] \leq \left(\frac{p!}{2}\right) R(c_0)^{p-2} \nu(M_{ij}), \tag{15}$$

where $R(c_0)$ is a constant only depending on c_0.

In fact, (15) is the standard Bernstein condition, which holds for various common distributions. As an example, the following result shows that the Poisson distribution satisfies this condition.

Proposition 1. *Let $X \sim$ Poisson(λ), where $\lambda \leq C(c_0)$ and $C(c_0)$ is a constant only depending on $c_0 > 0$. Then, for any integer $p \geq 2$, there holds*

$$\mathbb{E}\left[|X - \lambda|^p\right] \leq \left(\frac{p!}{2}\right) R(c_0)^{p-2} \lambda,$$

where $R(c_0)$ is a constant only depending on c_0.

Proof. Since a Poisson random variable is known to be discrete log-concave, by noting that $\mathrm{Var}(X) = \lambda$, this lemma can be directly obtained from Lemma 7.5 and Definition 1.2 of Schudy and Sviridenko (2011). □

4.2 Consistency

In the theoretical results presented in this section, we use SCORE (Jin, 2015) for spectral clustering. Here, we briefly introduce its implementation. First, compute the m leading singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\boldsymbol{A}$ corresponding to the m largest eigenvalues in magnitude; next, construct a $n \times (m-1)$ matrix of entrywise ratios $\boldsymbol{R}^{(m)}$: $\boldsymbol{R}^{(m)}(i,k) = u_{k+1}(i)/u_1(i)$ for $1 \leq i \leq n$ and $0 \leq k \leq m-1$; finally, the rows of the ratio matrix $\boldsymbol{R}^{(m)}$ are clustered by the k-means algorithm, assuming there are m clusters. In future work, it would be interesting to establish consistency results for other commonly used spectral clustering methods, e.g., the regularized spectral clustering (RSC) methods proposed and studied in Amini et al. (2013); Joseph and Yu (2016).

Next, we introduce the main results that guarantee the consistency of SVPS in selecting the number of communities K, provided that Assumption 1 holds. The following two theorems correspond to two parts of the main results: the null case ($m = K$) and the underfitting case ($m < K$).

Theorem 4.1 (Null Case)**.** *If we implement the procedure in Section 3 with the candidate number of communities $m = K$ and SCORE as the spectral clustering method, then, for any fixed $c_0 > 0$ in Assumption 1, as $n \to \infty$, we have $T_{n,m} \leq 2 + o_P(1)$.*

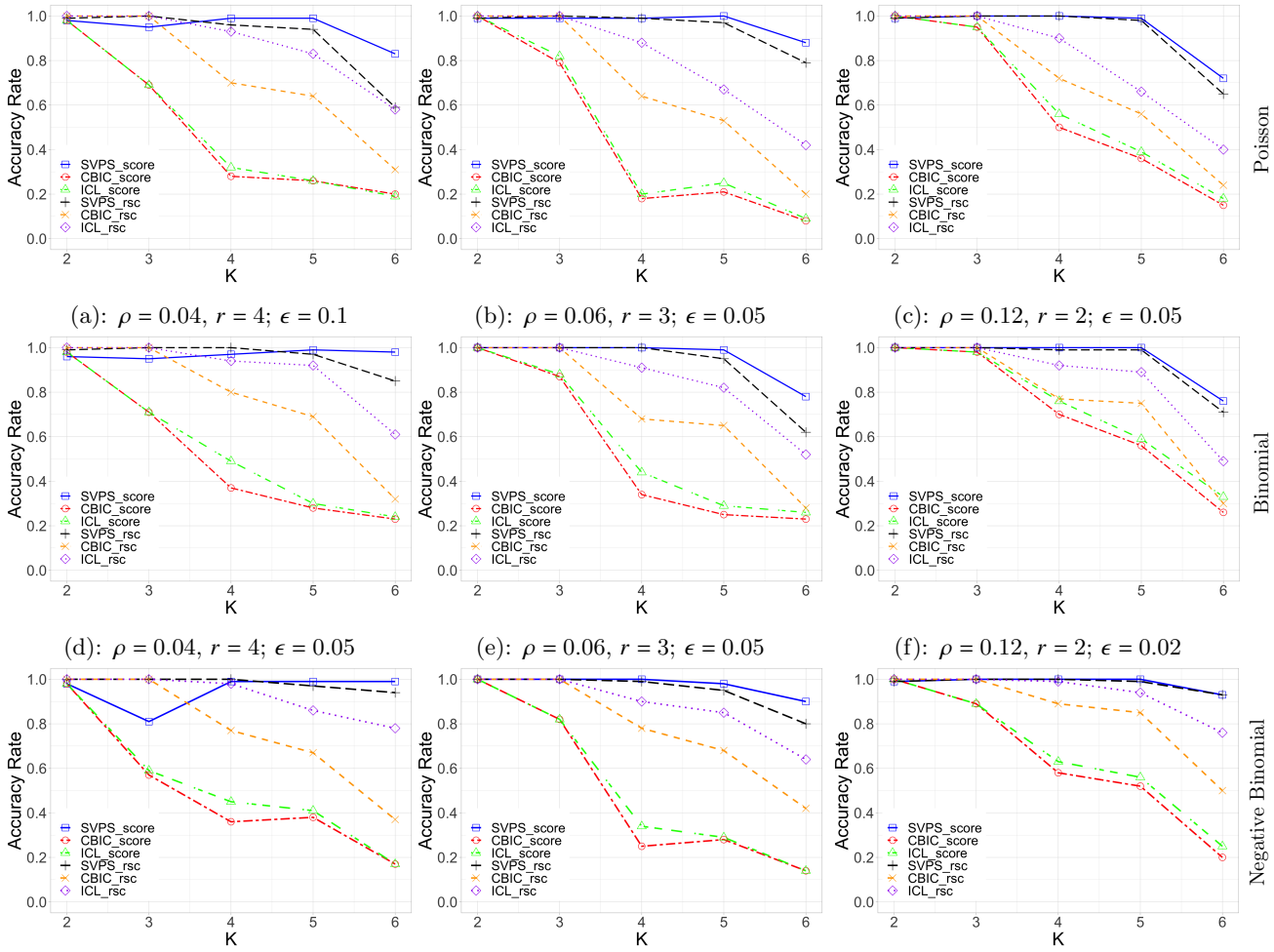
Theorem 4.2 (Underfitting Case)**.** *If we implement the procedure in Section 3 with the candidate number of communities $m < K$ and SCORE as the spectral clustering method, then, for any fixed $c_0 > 0$ in Assumption 1, as $n \to \infty$, we have $T_{n,m} \xrightarrow{P} \infty$.*

Obviously, combining Theorems 4.1 and 4.2 shows the consistency of SVPS. In (11), the sparsity assumption $\theta_{\min} \geq (\log^3 n)/\sqrt{n}$ might be suboptimal. In contrast, the corresponding assumption is typically $\theta_{\min} \geq C_0\sqrt{(\log n)/n}$ in the literature of unweighted network model selection (e.g., Jin et al., 2022). The extra logarithms for the weighted DCSBM arise from an application of the truncation technique in the proof, since existing spectral radius results for sparse and heterogeneous random matrices in the literature, e.g., Latała et al. (2018), usually require the entries to be uniformly bounded. It is an interesting question whether the logarithm factors can be further improved, but we leave this for future investigation.

5 NUMERICAL EXPERIMENTS

This section presents numerical experiments using both synthetic and real-world data to demonstrate the empirical behavior of SVPS. In the step of weighted DCSBM fitting in SVPS, we use either SCORE (Jin, 2015) or RSC (Amini et al., 2013; Joseph and Yu, 2016). If RSC is used, the regularization parameter is set to $0.25\,(\bar{d}/n)$, where $\bar{d}$ is the average node degree of the network. In the step of sequential testing, the threshold for comparing $T_{n,m}$ with is 2.02, 2.05 or 2.10. We will specify the exact choice in each experiment.

Throughout all our experiments, we select the variance function in SVPS as $\hat{\nu}(\mu) = \mu$. Note that this does not imply that the true variance function satisfies this



(a): $\rho = 0.04$, $r = 4$; $\epsilon = 0.1$ (b): $\rho = 0.06$, $r = 3$; $\epsilon = 0.05$ (c): $\rho = 0.12$, $r = 2$; $\epsilon = 0.05$

(d): $\rho = 0.04$, $r = 4$; $\epsilon = 0.05$ (e): $\rho = 0.06$, $r = 3$; $\epsilon = 0.05$ (f): $\rho = 0.12$, $r = 2$; $\epsilon = 0.02$

(g): $\rho = 0.04$, $r = 4$; $\epsilon = 0.1$ (h): $\rho = 0.06$, $r = 3$; $\epsilon = 0.1$ (i): $\rho = 0.12$, $r = 2$; $\epsilon = 0.1$

Figure 1: Accuracy rate comparison between SVPS, CBIC and ICL in different simulations. The top, middle and bottom rows correspond to the Poisson, binomial and negative binomial distribution, respectively.

relationship. In other words, model mismatching is allowed in our experiments. Therefore, the estimated variance profile matrix always satisfies $\widehat{\boldsymbol{V}}^{(m)} = \widehat{\boldsymbol{M}}^{(m)} = \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Pi}}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Pi}}^{\top}\widehat{\boldsymbol{\Theta}}$.

5.1 Synthetic Networks

5.1.1 Weighted DCSBM Generation

Now we discuss how to generate weighted networks based on certain weighted DCSBM for our simulations. We consider three types of distributions: Poisson, binomial, and negative binomial. In each case, the distribution of the weighted DCSBM is determined by its mean structure. In fact, to determine the mean structure, we only need to specify the $K \times K$ matrix $\boldsymbol{B}$, the vector of degree-correction parameters $\boldsymbol{\theta}$, and the block sizes. Therefore, we follow the experiment setting in Hu et al. (2020) and specify the above parameters as follows:

- For $\boldsymbol{B}$, let

$$B_{kl} = \rho\left(1 + r \times \mathbf{1}_{\{k=l\}}\right),$$

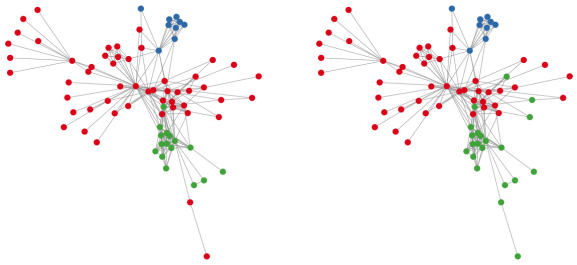
where ρ is a sparsity parameter and r is a contrast parameter. We consider the following combinations of (ρ, r): $(0.04, 4)$, $(0.06, 3)$ and $(0.12, 2)$ in our simulations.

- The block sizes are set according to the sequence $\boldsymbol{n}_{all} = (50, 100, 150, 50, 100, 150)$. For example, $(n_1, n_2) = (50, 100)$ when $K = 2$, $(n_1, n_2, n_3) = (50, 100, 150)$ when $K = 3$, etc.

- The degree-correction parameters are *i.i.d.* gener-

Table 1: Estimated K in the unweighted adjacency matrix of Les Misérables by different methods.

	BH	CBIC	ICL	StGoF
$\widehat{K}$	4	3	3	3



(a): RSC (b): SCORE

Figure 2: Estimated clusters of nodes in Les Misérables applying RSC and SCORE to the unweighted adjacency with $K = 3$.

ated from the following distribution:

$$\begin{cases} \text{Uniform}(0.6, 1.4), & \text{with probability } 0.8; \\ 0.5, & \text{with probability } 0.1; \\ 1.5, & \text{with probability } 0.1. \end{cases}$$

After generating $\boldsymbol{M}$, we consider the following simulations to generate $\boldsymbol{A}$:

- **Simulation 1:** $A_{ij} \overset{ind}{\sim} \text{Poisson}(M_{ij})$.
- **Simulation 2:** $A_{ij} \overset{ind}{\sim} \text{Binomial}(5, M_{ij}/5)$.
- **Simulation 3:** $A_{ij} \overset{ind}{\sim} \text{NB}(5, M_{ij}/5)$, where $\text{NB}(n, p)$ is the negative binomial distribution with n successful trials and the probability of success $1 - p$ in each trial.

5.1.2 Comparison with Score Based Methods

In the simulated experiments, SVPS is compared with two score based methods: corrected BIC (CBIC) in Hu et al. (2020) and integrated classification likelihood (ICL) method in Daudin et al. (2008), both of which are adapted to the weighted DCSBM with explicit likelihood function. To be specific, for each $m = 1, 2, ...,$ we calculate the following two scores

$$\text{CBIC}(m) = \log f(\boldsymbol{A}|\widehat{\boldsymbol{M}}^{(m)}) - \left[\lambda n \log m + \frac{m(m+1)}{2} \log n\right]$$

and

$$\text{ICL}(m) = \log f(\boldsymbol{A}|\widehat{\boldsymbol{M}}^{(m)}) - \left[\sum_{k=1}^{m} \hat{n}_k \log\left(\frac{n}{\hat{n}_k}\right) + \frac{m(m+2)}{2} \log n\right],$$

where λ is a tuning parameter, $\widehat{\boldsymbol{M}}^{(m)}$ is the estimated mean matrix, $\hat{n}_k$'s are the estimated block sizes according to $\widehat{\boldsymbol{M}}^{(m)}$, and $f(\boldsymbol{A}|\widehat{\boldsymbol{M}}^{(m)})$ is the likelihood function based on the actual generating distribution of the network. Throughout all our experiments, we fix $\lambda = 1$ which is the common choice used in Hu et al. (2020). Note that although ICL is derived under the SBM, we extend its usage to the weighted DCSBM, similar to the experiments in Hu et al. (2020). Ideally, $\widehat{\boldsymbol{M}}^{(m)}$ should be based on the maximum likelihood. However, for the sake of fair comparison with SVPS, we only consider estimating the mean adjacency matrix by spectral clustering as discussed in Section 3.1.

Next, we compare SVPS with CBIC and ICL in the three aforementioned simulations, experimenting with different combinations of (ρ, r) and $K = 2, 3, \ldots, 6$, where the spectral clustering method is either SCORE or RSC. In each simulation setup, we generate 100 independent networks and record the percentage of correctly estimating K of each method as the accuracy rate for comparison.

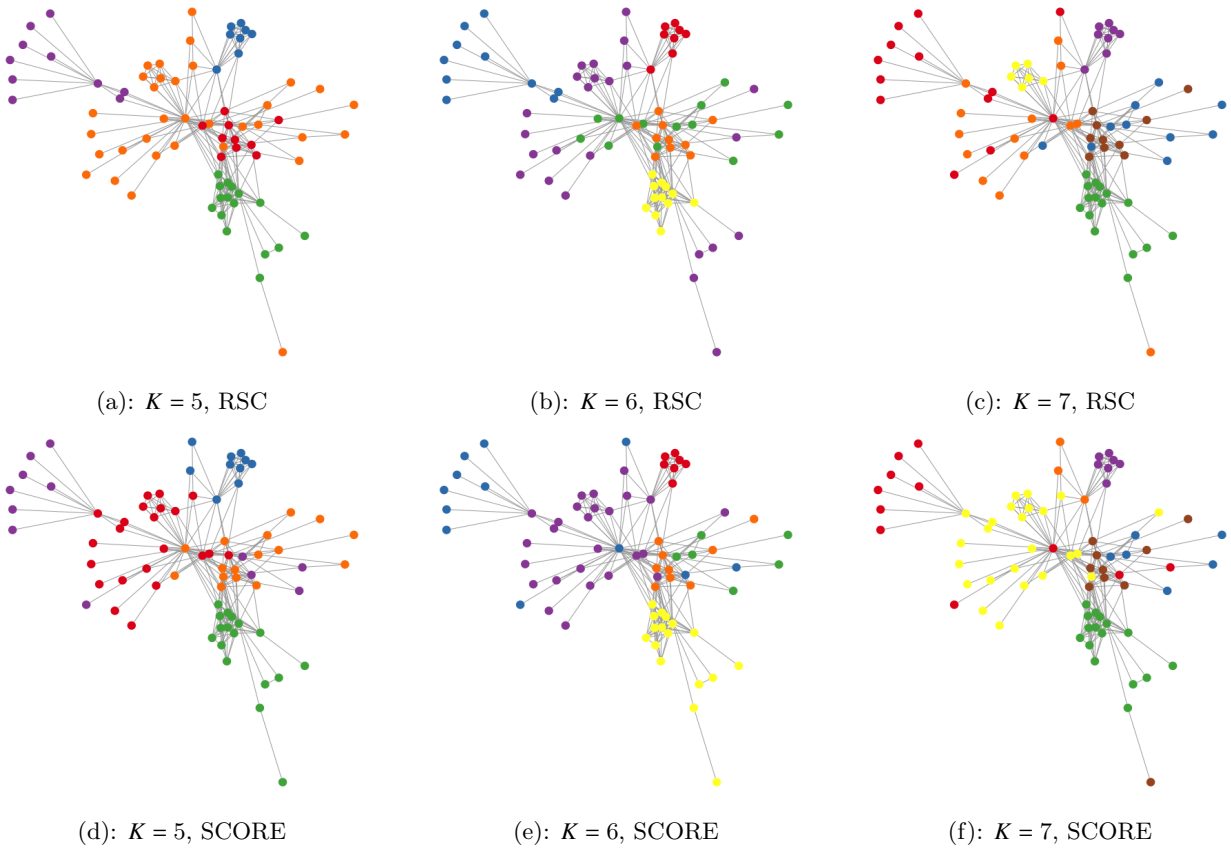
The plots of the empirical accuracy rates in the three simulations are shown in Figure 1, with figure captions indicating (ρ, r) and the choice of ϵ which determines the threshold in SVPS. As we can see, SVPS in general performs much better than the score based methods in these simulations, especially when using SCORE for spectral clustering. This result may seem surprising given that the score based methods utilize the likelihood information of the actual distribution, whereas SVPS does not even use the correct variance function. One possible explanation for the underperformance of the score based methods is that we only fit the models with spectral methods instead of maximum likelihood estimation with EM algorithms. Again, we choose spectral methods for fair comparison, since there is no likelihood as a guidance to fit the weighted DCSBM for SVPS. Another point to note is that in the first column of Figure 1, when $K = 3$, we observe a decrease in performance of SVPS with SCORE; however, the underlying reason remains unclear.

5.2 Les Misérables Network

In this subsection, we study the *Les Misérables* weighted network compiled in Knuth (1993), also analyzed in the literature of network analysis, see, e.g., Newman and Girvan (2004); Ball et al. (2011); Newman and Reinert (2016). In this network, any two characters (nodes) are connected by a weighted edge representing the number of co-occurrences between the pair in the same chapter of the book. The estimated number of communities by some model based approach in Newman and Reinert (2016) is 6, which may be rea-

Table 2: Estimated K in Les Misérables by applying SVPS to the regularized weighted adjacency with different τ values and CBIC, ICL to the original weighted adjacency with Poisson likelihood. The rows correspond to SCORE and RSC as the clustering method.

	SVPS ($\tau = 0.05$)	SVPS ($\tau = 0.1$)	SVPS ($\tau = 0.25$)	SVPS ($\tau = 0.5$)	CBIC	ICL
SCORE	7	6	7	6	7	7
RSC	7	7	6	6	7	5



(a): $K = 5$, RSC (b): $K = 6$, RSC (c): $K = 7$, RSC

(d): $K = 5$, SCORE (e): $K = 6$, SCORE (f): $K = 7$, SCORE

Figure 3: Estimated clusters in Les Misérables by applying RSC or SCORE to the weighted adjacency with $K = 5, 6, 7$.

sonable since corresponding major subplots are also identified in their analysis.

With this co-occurrence network, we investigate the difference between the estimated numbers of communities when the network is treated as unweighted and weighted. In the unweighted case, by replacing any positive weight with 1, we apply CBIC and ICL based on the standard DCSBM, as well as two popular methods for unweighted networks: the Bethe Hessian matrix (BH) proposed in Le and Levina (2022) and stepwise goodness-of-fit (StGoF) proposed in Jin et al. (2022) to the unweighted adjacency matrix to obtain the estimated number of communities $\widehat{K}$. We use SCORE for node clustering in CBIC, ICL and StGoF. Note that the original StGoF fails to stop in this case, so we choose $\widehat{K}$ as the number of communities which corresponds to the smallest test statistic, as suggested in Jin et al. (2022). The results of rank selection by these methods are listed in Table 1, which are either 3 or 4, much smaller than 6 as suggested in Newman and Reinert (2016). We apply RSC and SCORE with $K = 3$ to the unweighted network, and the clustering results are shown in Figure 2.

Coming back to the original weighted network, we compare the results of rank selection by SVPS, CBIC and ICL using both SCORE and RSC for spectral clustering. CBIC and ICL are based on the likelihood of the Poisson DCSBM. For the implementation of SVPS, we first preprocess the weighted network into a regu-

larized adjacency matrix $\boldsymbol{A}_\tau = \boldsymbol{A} + \tau \boldsymbol{J}_n$, where $\boldsymbol{J}_n$ is the $n \times n$ all-one matrix and $\tau = 0.05, 0.1, 0.25, 0.5$. The threshold for sequential testing in SVPS is set to 2.05. The comparison of the estimated K is summarized in Table 2, which shows some consistency among these methods. Additionally, Figure 3 displays the estimated clusters by applying SCORE and RSC to the weighted network with $K = 5, 6, 7$. It seems that RSC yields more interpretable results than SCORE.

6 DISCUSSION

This article studies a method for selecting the number of communities for weighted networks. First, we propose a generic weighted DCSBM, where the mean adjacency matrix is modeled as a DCSBM, and the variance profile is linked to the mean adjacency matrix through a variance function, without likelihood imposed. Next, we introduce a sequential testing approach for selecting the number of communities. In each step, the mean structure is fitted with some clustering method, and the variance profile matrix can be estimated through the variance function. A key component in constructing the test statistic is the scaling of the variance profile matrix, which is helpful in normalizing the adjacency matrix. The test statistic, obtained from the normalized adjacency matrix, is then used to determine the number of communities.

In theory, we show the consistency of our proposed procedure in selecting the number of communities under mild conditions on the weighted DCSBM. In particular, the network is allowed to be sparse. The consistency results include analysis of both the null and the underfitting cases. Additionally, the Nonsplitting Property studied in Jin et al. (2022) can be extended to the weighted DCSBM, which plays an essential role in our theoretical analysis.

For future work, a number of questions remain, both theoretically and methodologically. In theory, we assume that the weighted DCSBM satisfies some balanced conditions for analytical convenience, but this condition is very likely to be further relaxed. Our theoretical results are established for a known variance function $v(\cdot)$, and it would be interesting to have analogous results for an estimated variance function $\hat{v}(\cdot)$. Moreover, it would be valuable to study the asymptotic null distribution of $\boldsymbol{T}_{n,m}$, based on which the p-value can be calculated at each step. However, we are unable to derive distributional results for $\boldsymbol{T}_{n,m}$ based on the tools used in this paper.

A key assumption for our method is that the relationship between the variance profile matrix and the mean adjacency matrix can be characterized by a variance function. It would be interesting to see whether this assumption can be relaxed. Note that the estimation of the variance profile is for the purpose of matrix scaling so that the normalization matrix $\boldsymbol{\Psi}$ can be estimated under the null case. It would be valuable to investigate whether $\boldsymbol{\Psi}$ can be consistently estimated even if the variance profile matrix cannot be consistently estimated. We leave these aforementioned questions for future study.

Acknowledgments

Y. Liu and X. Li are partially supported by the NSF via the Career Award DMS-1848575. X. Li would like to thank Tracy Ke, Can Le, Haoran Li, Kaizheng Wang, and Ke Wang for their helpful discussions.

References

Abbe, E., Bandeira, A. S., and Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

Ball, B., Karrer, B., and Newman, M. E. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels1. *The Annals of Statistics*, 41(4):1922–1943.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Bosq, D. (2000). *Stochastic Processes and Random Variables in Function Spaces*, pages 15–42. Springer New York, New York, NY.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cai, T. T. and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059.

Chen, K. and Lei, J. (2018). Network cross-validation for determining the number of communities in net-

work data. *Journal of the American Statistical Association*, 113(521):241–251.

Chen, Y., Li, X., and Xu, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics*, 46(4):1573–1602.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

Deng, S., Ling, S., and Strohmer, T. (2021). Strong consistency, graph laplacians, and the stochastic block model. *The Journal of Machine Learning Research*, 22(1):5210–5253.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 3(486):75–174.

Gao, C. and Lafferty, J. (2017). Testing for global network structure using small subgraph statistics. *arXiv preprint arXiv:1710.00862*.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Guédon, O. and Vershynin, R. (2015). Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25.

Gulikers, L., Lelarge, M., and Massoulié, L. (2017). A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721.

Han, W. and Han, G. (2019). A new proof of hopf's inequality using a complex extension of the hilbert metric.

Han, X., Yang, Q., and Fan, Y. (2023). Universal rank inference via residual subsampling with application to large networks. *The Annals of Statistics*, 51(3):1109–1133.

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

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hu, J., Qin, H., Yan, T., and Zhao, Y. (2020). Corrected bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, 115(532):1771–1783.

Hu, J., Zhang, J., Qin, H., Yan, T., and Zhu, J. (2021). Using maximum entry-wise deviation to test the goodness of fit for stochastic block models. *Journal of the American Statistical Association*, 116(535):1373–1382.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89.

Jin, J., Ke, Z., and Luo, S. (2018). Network global testing by counting graphlets. In *International conference on machine learning*, pages 2333–2341. PMLR.

Jin, J., Ke, Z. T., and Luo, S. (2017). Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*.

Jin, J., Ke, Z. T., and Luo, S. (2021). Optimal adaptivity of signed-polygon statistics for network testing. *The Annals of Statistics*, 49(6):3408–3433.

Jin, J., Ke, Z. T., Luo, S., and Wang, M. (2022). Optimal estimation of the number of communities. *Journal of the American Statistical Association*, (just-accepted):1–41.

Joseph, A. and Yu, B. (2016). Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

Knight, P. A., Ruiz, D., and Uçar, B. (2014). A symmetry preserving algorithm for matrix scaling. *SIAM journal on Matrix Analysis and Applications*, 35(3):931–955.

Knuth, D. E. (1993). *The Stanford GraphBase: a platform for combinatorial computing*. AcM Press New York.

Landa, B. (2022). Scaling positive random matrices: concentration and asymptotic convergence. *Electronic Communications in Probability*, 27:1–13.

Landa, B., Zhang, T. T., and Kluger, Y. (2022). Biwhitening reveals the rank of a count matrix. *SIAM journal on mathematics of data science*, 4(4):1420–1446.

Latała, R., van Handel, R., and Youssef, P. (2018). The dimension-free structure of nonhomogeneous random matrices. *Inventiones mathematicae*, 214(3):1031–1080.

Le, C. M. and Levina, E. (2022). Estimating the number of communities by spectral methods. *Electronic Journal of Statistics*, 16(1):3315 – 3342.

Le, C. M., Levina, E., and Vershynin, R. (2016). Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics*, pages 373–400.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

Lei, L., Li, X., and Lou, X. (2020). Consistency of spectral clustering on hierarchical stochastic block models. *arXiv preprint arXiv:2004.14531*.

Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P. J., and Levina, E. (2022). Hierarchical community detection by recursive partitioning. *Journal of the American Statistical Association*, 117(538):951–968.

Li, T., Levina, E., and Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276.

Li, X., Chen, Y., and Xu, J. (2021). Convex relaxation methods for community detection. *Statistical science*, 36(1):2–15.

Ma, S., Su, L., and Zhang, Y. (2021). Determining the number of communities in degree-corrected stochastic block models. *Journal of Machine Learning Research*, 22(69):1–63.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Newman, M. E. and Reinert, G. (2016). Estimating the number of communities in a network. *Physical review letters*, 117(7):078301.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915.

Saldana, D. F., Yu, Y., and Feng, Y. (2017). How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181.

Schudy, W. and Sviridenko, M. (2011). Bernstein-like concentration and moment inequalities for polynomials of independent random variables: multilinear case. *arXiv preprint arXiv:1109.5193*.

Sinkhorn, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Wang, Y. R. and Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528.

Zhang, A. Y. and Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280.

Zhang, A. Y. and Zhou, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, 48(5):2575–2598.

Zhang, Y., Levina, E., and Zhu, J. (2020). Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

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

ORGANIZATION OF THE SUPPLEMENTARY MATERIAL

The supplementary material consists of four appendices. Appendix A contains the supporting lemmas of the main results. Appendix B presents the proofs of Theorem 4.1 and Theorem 4.2. Appendix C is the proof of Lemma A.10, which proves the Nonsplitting Property of SCORE under the weighted DCSBM. Appendix D presents additional synthetic simulations with larger values of K.

A SUPPORTING LEMMAS

We first cite two results from Landa (2022) regarding the sensitivity analysis of matrix scaling. Here we only state the results for symmetric matrix scaling with row sums equal to n.

Lemma A.1 (Landa (2022))**.** *Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with positive entries. Then, there exists a unique positive vector $\boldsymbol{x} \in \mathbb{R}^n$ satisfying*

$$x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = 1, \quad i = 1, \ldots, n.$$

Furthermore, denote $a_{\max} = \max_{1 \le i \le j \le n} A_{ij}$ and $a_{\min} = \min_{1 \le i \le j \le n} A_{ij}$. Then,

$$\frac{1}{\sqrt{n}} \frac{\sqrt{a_{\min}}}{a_{\max}} \le x_i \le \frac{1}{\sqrt{n}} \frac{\sqrt{a_{\max}}}{a_{\min}}, \quad i = 1, \ldots, n.$$

Lemma A.2 (Landa (2022))**.** *Let $\widetilde{\boldsymbol{A}}$ be a symmetric matrix with positive entries. Denote $\tilde{a}_{\max} = \max_{1 \le i \le j \le n} \widetilde{A}_{ij}$ and $\tilde{a}_{\min} = \min_{1 \le i \le j \le n} \widetilde{A}_{ij}$. Suppose there is a constant $\epsilon \in (0, 1)$ and a positive vector $\boldsymbol{x} \in \mathbb{R}^n$, such that*

$$\max_{1 \le i \le n} \left| \sum_{j=1}^{n} x_i \widetilde{A}_{ij} x_j - 1 \right| \le \epsilon.$$

Denote $x_{\min} = \min_{1 \le i \le n} x_i$. Then, there exists a positive vector $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{n} \tilde{x}_i \widetilde{A}_{ij} \tilde{x}_j = 1, \quad i = 1, \ldots, n,$$

and

$$\max_{1 \le i \le n} \left| \frac{\tilde{x}_i}{x_i} - 1 \right| \le \frac{\epsilon}{1 - \epsilon} + 4\epsilon \cdot \frac{n^{3/2}}{x_{\min}^3} \cdot \frac{\sqrt{\tilde{a}_{\max}}}{\tilde{a}_{\min}^2}.$$

As a consequence, if $\boldsymbol{x}$ satisfies the equations in Lemma A.1, there holds

$$\max_{1 \le i \le n} \left| \frac{\tilde{x}_i}{x_i} - 1 \right| \le \frac{\epsilon}{1 - \epsilon} + 4\epsilon \cdot \frac{a_{\max}^3}{a_{\min}^{3/2}} \cdot \frac{\tilde{a}_{\max}^{1/2}}{\tilde{a}_{\min}^2}.$$

The following lemma provides a tail bound of the sum of independent random variables satisfying the Bernstein condition.

Lemma A.3 (Bernstein's Inequality, Corollary 2.11 of Boucheron et al. (2013))**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \mathbb{E}[X_i^2] \le v$ and*

$$\sum_{i=1}^n \mathbb{E}\left[|X_i|^q\right] \le \frac{q!}{2} c^{q-2} v \quad \text{for all intergers } q \ge 3.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n \left(X_i - \mathbb{E}\left[X_i\right]\right) \ge t\right) \le \exp\left(-\frac{t^2}{2(v+ct)}\right).$$

The following lemma given in Latała et al. (2018) bounds the operator norm of heterogeneous random matrices.

Lemma A.4 (Remark 4.12 of Latała et al. (2018))**.** *Suppose that $\boldsymbol{X}$ is a random matrix with independent and centered upper-triangular entries. Define the quantities*

$$\sigma_\infty \coloneqq \max_i \sqrt{\sum_j \mathbb{E}\left[X_{ij}^2\right]}, \quad \sigma_\infty^* \coloneqq \max_{i,j} \left\|X_{ij}\right\|_\infty.$$

Then, for every $0 \le \epsilon \le 1$ and $t \ge 0$, we have

$$\mathbb{P}\left(\|\boldsymbol{X}\| \ge 2(1+\epsilon)\sigma_\infty + t\right) \le n \exp\left(-\frac{\epsilon t^2}{C\sigma_\infty^{*2}}\right),$$

where C is a universal constant.

Lemma A.5. *Under Assumption 1, there holds $|\lambda_K(\boldsymbol{M})| \ge c_0 \theta_{\min}^2 n$.*

Proof. Recall that $\boldsymbol{M} = \boldsymbol{\Theta\Pi B \Pi^\top \Theta}$, where $\boldsymbol{\Theta\Pi} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K]$. Denote $\sigma_k(\cdot)$ as the k-th singular value of a matrix. It is easy to observe

$$\sigma_K(\boldsymbol{\Theta\Pi}) = \min_k \|\boldsymbol{\theta}_k\|_2 \ge \theta_{\min}\sqrt{n}.$$

Then by the assumption (13), we have

$$|\lambda_K(\boldsymbol{M})| \ge |\lambda_K(\boldsymbol{B})| \cdot \sigma_K(\boldsymbol{\Theta\Pi})^2 \ge c_0 \theta_{\min}^2 n.$$

□

Lemma A.6. *Under Assumption 1, the scaling factors satisfy*

$$1/\left(C\theta_{\min}\sqrt{n}\right) \le \psi_i \le C/\left(\theta_{\min}\sqrt{n}\right), \quad i = 1, \ldots, n$$

for some constant C that only relies on c_0.

Proof. In light of Assumption 1, this result is implied by Lemma A.1 directly. □

Lemma A.7. *For any fixed $c_0 > 0$ in Assumption 1, there holds*

$$\left\|\boldsymbol{\Psi}^{\frac{1}{2}}(\boldsymbol{A} - \boldsymbol{M})\boldsymbol{\Psi}^{\frac{1}{2}}\right\| \le 2 + o_P(1) \quad \text{as } n \to \infty.$$

Proof. Denote $\boldsymbol{E} \coloneqq \boldsymbol{\Psi}^{\frac{1}{2}}(\boldsymbol{A} - \boldsymbol{M})\boldsymbol{\Psi}^{\frac{1}{2}}$ whose entries are $E_{ij} = (A_{ij} - M_{ij})\psi_i^{\frac{1}{2}}\psi_j^{\frac{1}{2}}$. By (9), we have

$$\sum_{j=1}^n \mathbb{E}\left[E_{ij}^2\right] = 1, \quad \forall\, i = 1, \ldots, n. \tag{16}$$

Define

$$\widehat{E}_{ij} = E_{ij}\mathbf{1}_{\{|E_{ij}| < 1/(\log n)\}} \quad \text{and} \quad \widetilde{E}_{ij} = \widehat{E}_{ij} - \mathbb{E}\left[\widehat{E}_{ij}\right].$$

The resulting matrices are denoted as $\widehat{\boldsymbol{E}}$ and $\widetilde{\boldsymbol{E}}$, respectively. Our proof strategy is to first bound $\|\widetilde{\boldsymbol{E}}\|$, and then $\|\widehat{\boldsymbol{E}}\|$, and finally $\|\boldsymbol{E}\|$.

Notice that $\widetilde{\boldsymbol{E}}$ has independent and mean-centered entries for $i \geq j$. Since

$$\left|\widehat{E}_{ij}\right| = \left|E_{ij}\right| \mathbf{1}_{\{|E_{ij}|<1/(\log n)\}} \leq \left|E_{ij}\right|,$$

almost surely we have

$$\mathbb{E}\left[\widetilde{E}_{ij}^2\right] = \mathrm{Var}\left(\widetilde{E}_{ij}\right) = \mathrm{Var}\left(\widehat{E}_{ij}\right) \leq \mathbb{E}\left[\widehat{E}_{ij}^2\right] \leq \mathbb{E}\left[E_{ij}^2\right].$$

This implies that

$$\sum_{j=1}^{n} \mathbb{E}\left[\widetilde{E}_{ij}^2\right] \leq 1, \quad \forall\, i = 1, \ldots, n.$$

Also, since $\|\widehat{E}_{ij}\|_\infty \leq 1/(\log n)$, we have $|\,\mathbb{E}[\widehat{E}_{ij}]| \leq 1/(\log n)$, and thus

$$\|\widetilde{E}_{ij}\|_\infty \leq \|\widehat{E}_{ij}\|_\infty + |\,\mathbb{E}\,\widehat{E}_{ij}| \leq 2/(\log n).$$

Define the quantities

$$\tilde{\sigma} \coloneqq \max_i \sqrt{\sum_j \mathbb{E}[\widetilde{E}_{ij}^2]} \leq 1, \quad \tilde{\sigma}_* \coloneqq \max_{i,j} \|\widetilde{E}_{ij}\|_\infty \leq 2/(\log n).$$

Then, apply Corollary A.4 to $\widetilde{\boldsymbol{E}}$. For any $t > 0$, there holds

$$\mathbb{P}\left(\|\widetilde{\boldsymbol{E}}\| \geq 2(1+\epsilon)\tilde{\sigma} + t\right) \leq n \exp\left(-\frac{\epsilon t^2}{C\tilde{\sigma}_*^2}\right).$$

By letting $\epsilon = t = \log^{-1/4} n$, with the above inequalities, we have

$$\|\widetilde{\boldsymbol{E}}\| \leq 2 + o_P(1) \quad \text{as } n \to \infty. \tag{17}$$

Next, we give an upper bound of $\|\widehat{\boldsymbol{E}}\|$. Since $\widetilde{\boldsymbol{E}} = \widehat{\boldsymbol{E}} - \mathbb{E}[\widehat{\boldsymbol{E}}]$, we have $\|\widehat{\boldsymbol{E}}\| \leq \|\widetilde{\boldsymbol{E}}\| + \|\,\mathbb{E}[\widehat{\boldsymbol{E}}]\|_F$. Notice that

$$\mathbb{E}\left[E_{ij}\mathbf{1}_{\{|E_{ij}|\geq 1/(\log n)\}}\right] = \mathbb{E}\left[E_{ij} - \widehat{E}_{ij}\right] = -\,\mathbb{E}\left[\widehat{E}_{ij}\right].$$

By Cauchy-Schwarz inequality,

$$\left(\mathbb{E}\left[\widehat{E}_{ij}\right]\right)^2 = \left(\mathbb{E}\left[E_{ij}\mathbf{1}_{\{|E_{ij}|\geq 1/(\log n)\}}\right]\right)^2 \leq \mathbb{E}[E_{ij}^2]\,\mathbb{P}\left(|E_{ij}| \geq 1/(\log n)\right). \tag{18}$$

By Lemma A.6, there holds

$$\begin{aligned}
\mathbb{P}\left(\left|E_{ij}\right| \geq 1/(\log n)\right) &= \mathbb{P}\left(\left|A_{ij} - M_{ij}\right| \geq \frac{1}{\psi_i^{1/2}\psi_j^{1/2}\log n}\right) \\
&\leq \mathbb{P}\left(\left|A_{ij} - M_{ij}\right| \geq \frac{\theta_{\min}\sqrt{n}}{C\log n}\right).
\end{aligned}$$

Then, by Assumption 1 and Lemma A.3, we have

$$\begin{aligned}
\max_{1\leq i,j\leq n} \mathbb{P}\left(\left|E_{ij}\right| \geq 1/(\log n)\right) &\leq \max_{1\leq i\leq j\leq n} 2\exp\left(-\frac{\left(\frac{\theta_{\min}\sqrt{n}}{C\log n}\right)^2}{2\left(M_{ij}/c_0 + R\frac{\theta_{\min}\sqrt{n}}{C\log n}\right)}\right) \\
&\leq 2\exp\left(-\frac{C\theta_{\min} n}{\theta_{\min}\log^2 n + \sqrt{n}\log n}\right) \\
&= O(n^{-10}),
\end{aligned} \tag{19}$$

where the last line is due to (12) in Assumption 1. Then, (16) and (18) imply

$$\left\|\mathbb{E}[\widehat{\boldsymbol{E}}]\right\|_F^2 = \sum_{1\le i,j\le n}\left(\mathbb{E}\,\widehat{E}_{ij}\right)^2 \le \sum_{1\le i,j\le n}\mathbb{E}[E_{ij}^2]\,\mathbb{P}\left(\left|E_{ij}\right|\ge 1/(\log n)\right) = O(n^{-9}).$$

Then (17) implies

$$\|\widehat{\boldsymbol{E}}\| \le \|\widetilde{\boldsymbol{E}}\| + \|\,\mathbb{E}[\widehat{\boldsymbol{E}}]\|_F \le 2 + o_P(1). \tag{20}$$

Finally, note that

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{E}}\neq\boldsymbol{E}\right) &= \mathbb{P}\left(\cup_{1\le i,j\le n}\{|E_{ij}|\ge 1/(\log n)\}\right)\\ &\le \sum_{1\le i,j\le n}\mathbb{P}(|E_{ij}|\ge 1/(\log n))\\ &\le n^2\max_{1\le i,j\le n}\mathbb{P}(|E_{ij}|\ge 1/(\log n)) = O(n^{-8}).\end{aligned}$$

Then (20) implies $\|\boldsymbol{E}\| \le 2 + o_P(1)$. □

Lemma A.8. *Under Assumption 1, there holds*

$$\|\boldsymbol{A}-\boldsymbol{M}\| = O_P(\theta_{\min}\sqrt{n}) \quad \text{as } n\to\infty.$$

Proof. This is a straightforward corollary of Lemmas A.6 and A.7. □

Lemma A.9. *Under Assumption 1, we have*

$$\max_{1\le i\le n}\left|\frac{d_i}{d_i^*}-1\right| = o_P(1) \quad \text{and} \quad \max_{1\le k,l\le K}\left|\frac{\mathbf{1}_k^\top\boldsymbol{A}\mathbf{1}_l}{\mathbf{1}_k^\top\boldsymbol{M}\mathbf{1}_l}-1\right| = o_P(1).$$

Proof. Note that

$$d_i - d_i^* = \sum_{j=1}^n (A_{ij}-M_{ij}),$$

in which the summands satisfy (15). Then, we have

$$\mathrm{var}(d_i-d_i^*) = \sum_{j=1}^n \nu(M_{ij}) \le \frac{1}{c_0}\sum_{j=1}^n M_{ij}.$$

By Lemma A.3, for any fixed $\epsilon>0$, we have

$$\begin{aligned}\sum_{i=1}^n \mathbb{P}\left(|d_i-d_i^*|\ge\epsilon d_i^*\right) &\le \sum_{i=1}^n 2\exp\left(-\frac{\epsilon^2 d_i^{*2}}{2\left(\frac{1}{c_0}(\sum_{j=1}^n M_{ij})+R\epsilon d_i^*\right)}\right)\\ &\le 2n\exp\left(-C\left(\frac{\epsilon^2}{1+\epsilon}\right)\theta_{\min}^2 n\right) = O(n^{-3}),\end{aligned}$$

where the last inequality is due to (12) in Assumption 1. Therefore, with probability $1-O(n^{-3})$, we have $\max_{1\le i\le n}\left|\frac{d_i}{d_i^*}-1\right|\le\epsilon$. Given ϵ can be chosen arbitrarily small, we have $\max_{1\le i\le n}\left|\frac{d_i}{d_i^*}-1\right| = o_P(1)$.

With similar arguments, we have $\max_{1\le k,l\le K}\left|\frac{\mathbf{1}_k^\top\boldsymbol{A}\mathbf{1}_l}{\mathbf{1}_k^\top\boldsymbol{M}\mathbf{1}_l}-1\right| = o_P(1)$. □

For unweighted networks, Jin et al. (2022) proved that SCORE enjoys the Nonsplitting Property (NSP) in both the underfitting and null cases. In other words, the true communities are refinements of the estimated communities with high probability. This property is essential for analyzing sequential testing approaches, as it reduces the number of possible DCSBM fittings. We give a formal definition of this property.

Definition 1 (Nonsplitting Property (Jin et al., 2022))**.** Let the ground truth of communities in a network be $\mathcal{N}_1, \ldots, \mathcal{N}_K$. Assume $m \le K$, and $\widehat{\mathcal{N}}_1^{(m)}, \ldots, \widehat{\mathcal{N}}_m^{(m)}$ are the estimated communities by certain clustering method. We say that these estimated communities satisfy the Nonsplitting Property (NSP), if the true communities $\mathcal{N}_1, \ldots, \mathcal{N}_K$ are a refinement of the estimated ones. In other words, for any $k = 1, \ldots, K$, there is exactly one $l = 1, \ldots, m$, such that $\mathcal{N}_k \cap \widehat{\mathcal{N}}_l^{(m)} \neq \emptyset$.

Then, we have the following lemma that guarantees the NSP of SCORE under the weighted DCSBM.

Lemma A.10. *Under Assumption 1 with any fixed $c_0 > 0$, for any fixed $m \le K$, SCORE satisfies the NSP with probability $1 - O(n^{-3})$.*

The proof of this lemma is deferred to Appendix C, which follows the idea of Jin et al. (2022). A crucial component of the proof is the row-wise bounds of eigenvector perturbations (Jin et al., 2017; Abbe et al., 2020).

B PROOFS OF THE MAIN RESULTS

Throughout this section, we use C to represent a constant that depends only on c_0, and its value may change from line to line.

B.1 Proof of Theorem 4.1

This subsection aims to prove Theorem 4.1, so we assume $m = K$. For notational convenience, we omit the superscripts in the estimators indicating the m-th step.

Note that

$$T_{n,m} = \left| \lambda_{m+1} \left(\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{A} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right) \right|$$

and

$$\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{A} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} = \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{M} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} + \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} (\boldsymbol{A} - \boldsymbol{M}) \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}},$$

where $\operatorname{rank}\left(\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{M} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right) \le K = m$. Then we have

$$T_{n,m} \le \left\| \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} (\boldsymbol{A} - \boldsymbol{M}) \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right\|.$$

On the other hand,

$$\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} (\boldsymbol{A} - \boldsymbol{M}) \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} = \left(\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{\Psi}^{-\frac{1}{2}} \right) \left(\boldsymbol{\Psi}^{\frac{1}{2}} (\boldsymbol{A} - \boldsymbol{M}) \boldsymbol{\Psi}^{\frac{1}{2}} \right) \left(\boldsymbol{\Psi}^{-\frac{1}{2}} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right).$$

In order to show that $T_{n,m} \le 2 + o_p(1)$, it suffices to show the following inequalities

$$\begin{cases} \left\| \boldsymbol{\Psi}^{\frac{1}{2}} (\boldsymbol{A} - \boldsymbol{M}) \boldsymbol{\Psi}^{\frac{1}{2}} \right\| \le 2 + o_p(1), \\ \left\| \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I}_n \right\| \le o_p(1). \end{cases}$$

The first inequality follows from Lemma A.7 and the second inequality can be shown by the following lemma.

Lemma B.1. *Under the null case $m = K$, Assumption 1 implies*

$$\max_{1 \le i \le n} \left| \frac{\hat{\theta}_i}{\theta_i} - 1 \right| = o_P(1), \quad \max_{1 \le k,l \le K} \left| \frac{\widehat{B}_{kl}}{B_{kl}} - 1 \right| = o_P(1),$$

$$\max_{1 \le i,j \le n} \left| \frac{\widehat{M}_{ij}}{M_{ij}} - 1 \right| = o_P(1), \quad \max_{1 \le k,l \le n} \left| \frac{\widehat{V}_{kl}}{V_{kl}} - 1 \right| = o_P(1).$$

Furthermore, we have

$$\left\| \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I}_n \right\| = o_P(1).$$

Proof. In the null case, the NSP given in Theorem A.10 implies strong consistency for the recovery of communities. Without loss of generality, denote $\widehat{\mathcal{N}}_k = \mathcal{N}_k$ and thereby $\mathbf{1}_k = \hat{\mathbf{1}}_k$ for $k = 1, \ldots, K$, with probability $1 - O(n^{-3})$. By combining (3), (4), (5) and (6), we get

$$\frac{\theta_i}{\hat{\theta}_i} = \frac{d_i^*}{d_i} \frac{\sum_{j \in \mathcal{N}_k} d_j}{\sum_{j \in \mathcal{N}_k} d_j^*} \sqrt{\frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}{\mathbf{1}_k^\top \boldsymbol{A} \mathbf{1}_k}},$$

and

$$\frac{B_{kl}}{\widehat{B}_{kl}} = \frac{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_l}{\mathbf{1}_k^\top \boldsymbol{A} \mathbf{1}_l} \sqrt{\frac{\mathbf{1}_k^\top \boldsymbol{A} \mathbf{1}_k}{\mathbf{1}_k^\top \boldsymbol{M} \mathbf{1}_k}} \sqrt{\frac{\mathbf{1}_l^\top \boldsymbol{A} \mathbf{1}_l}{\mathbf{1}_l^\top \boldsymbol{M} \mathbf{1}_l}}.$$

Then by Lemma A.9, we have

$$\max_{1 \le i \le n} \left| \frac{\hat{\theta}_i}{\theta_i} - 1 \right| = o_P(1), \quad \max_{1 \le k,l \le K} \left| \frac{\widehat{B}_{kl}}{B_{kl}} - 1 \right| = o_P(1).$$

Furthermore, $\max_{1 \le i,j \le n} \left| \frac{\widehat{M}_{ij}}{M_{ij}} - 1 \right| = o_P(1)$ follows from

$$\frac{\widehat{M}_{ij}}{M_{ij}} = \frac{\hat{\theta}_i \hat{\theta}_j \widehat{B}_{\phi(i)\phi(j)}}{\theta_i \theta_j B_{\phi(i)\phi(j)}}.$$

Then

$$\max_{1 \le i,j \le n} \left| \frac{\widehat{V}_{ij}}{V_{ij}} - 1 \right| = o_P(1) \tag{21}$$

follows from the assumption (14) on the variance-mean function.

On the other hand, notice that (21) implies

$$\max_{1 \le i \le n} \left| \frac{\sum_{j=1}^n \widehat{V}_{ij} \psi_i \psi_j}{\sum_{j=1}^n V_{ij} \psi_i \psi_j} - 1 \right| = o_P(1).$$

By (9), we further have

$$\max_{1 \le i \le n} \left| \sum_{j=1}^n \widehat{V}_{ij} \psi_i \psi_j - 1 \right| = o_P(1).$$

Denote $V_{\max} = \max_{1 \le i,j \le n} V_{ij}$ and $V_{\min} = \min_{1 \le i,j \le n} V_{ij}$. Additionally, $\widehat{V}_{\max}$ and $\widehat{V}_{\min}$ are defined similarly. Assumption 1 implies

$$V_{\max}/V_{\min} = O(1) \quad \text{and} \quad V_{\min}/V_{\max} = O(1).$$

Combined with (21), we have

$$\widehat{V}_{\max} \Big/ V_{\max} = O_P(1) \quad \text{and} \quad \widehat{V}_{\min} \Big/ V_{\min} = O_P(1).$$

Plug in the above inequalities to Lemma A.2, we have $\left\| \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I}_n \right\| = o_P(1)$. □

B.2 Proof of Theorem 4.2

This subsection aims to prove Theorem 4.2 with $m < K$. Again, we omit the superscripts in the estimators.

Note that

$$T_{n,m} = \left| \lambda_{m+1} \left(\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{A} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right) \right| \ge \left| \lambda_K \left(\widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \boldsymbol{A} \widehat{\boldsymbol{\Psi}}^{\frac{1}{2}} \right) \right| \ge |\lambda_K(\boldsymbol{A})| \hat{\psi}_{\min} \ge (|\lambda_K(\boldsymbol{M})| - \|\boldsymbol{A} - \boldsymbol{M}\|) \hat{\psi}_{\min}, \tag{22}$$

where $\hat{\psi}_{\min} = \min_{1 \le i \le n} \hat{\psi}_i$. Therefore, it suffices to find lower bounds for $|\lambda_K(\boldsymbol{M})|$ and $\hat{\psi}_{\min}$, as well as an upper bound for $\|\boldsymbol{A} - \boldsymbol{M}\|$.

Lemma B.2. *For the underfitting case $m < K$, denote $\widehat{M}_{\max} = \max_{1\le i,j\le n} \widehat{M}_{ij}$ and $\widehat{M}_{\min} = \min_{1\le i,j\le n} \widehat{M}_{ij}$. Also, denote $\widehat{V}_{\max}$ and $\widehat{V}_{\min}$ in a similar manner. Then under Assumption 1, there holds*

$$\max\left(\frac{\widehat{M}_{\max}}{\theta^2_{\min}}, \quad \frac{\theta^2_{\min}}{\widehat{M}_{\min}}\right) = O_P(1),$$

and

$$\max\left(\frac{\widehat{V}_{\max}}{\theta^2_{\min}}, \quad \frac{\theta^2_{\min}}{\widehat{V}_{\min}}\right) = O_P(1).$$

Furthermore, we have

$$\frac{1}{\theta_{\min}\hat{\psi}_{\min}\sqrt{n}} = O_P(1).$$

Proof. In the underfitting case $m < K$, the NSP given in Theorem A.10 implies that the true communities $\mathcal{N}_1, \dots, \mathcal{N}_K$ are refinements of the estimated communities $\widehat{\mathcal{N}}_1, \dots, \widehat{\mathcal{N}}_m$. For each $1 \le k \le m$, we assume that the number of true communities contained in $\widehat{\mathcal{N}}_k$ is $r_k \ge 1$, which implies that $r_1 + \cdots + r_m = K$. Then, we can represent the estimated communities as

$$\widehat{\mathcal{N}}_k = \mathcal{N}_{h_{k1}} \cup \cdots \cup \mathcal{N}_{h_{kr_k}}, \quad k = 1, \dots, m.$$

Here all indices h_{kj} for $k = 1, \dots, m$ and $j = 1, \dots, r_k$ are distinct over $1, \dots, K$. Similarly, we can decompose

$$\hat{\mathbf{1}}_k = \mathbf{1}_{h_{k1}} + \cdots + \mathbf{1}_{h_{kr_k}}, \quad k = 1, \dots, m.$$

Recall that

$$\widehat{M}_{ij} = \frac{(\hat{\mathbf{1}}_k)^\top \boldsymbol{A} \hat{\mathbf{1}}_l}{(\hat{\mathbf{1}}_k)^\top \boldsymbol{A} \mathbf{1}_n \cdot (\hat{\mathbf{1}}_l)^\top \boldsymbol{A} \mathbf{1}_n} d_i d_j.$$

Then by Lemma A.9, NSP and Assumption 1, it is easy to obtain

$$\max\left(\frac{\widehat{M}_{\max}}{\theta^2_{\min}}, \ \frac{\theta^2_{\min}}{\widehat{M}_{\min}}\right) = O_P(1).$$

Furthermore, by (14) in Assumption 1, we have

$$\max\left(\frac{\widehat{V}_{\max}}{\theta^2_{\min}}, \ \frac{\theta^2_{\min}}{\widehat{V}_{\min}}\right) = O_P(1).$$

Then by (10) and Lemma A.1, we have

$$\frac{1}{\theta_{\min}\hat{\psi}_{\min}\sqrt{n}} = O_P(1).$$

□

Proof of Theorem 4.2. By Assumption 1, we have

$$\theta_{\min}\sqrt{n} \ge \log^3 n \to \infty.$$

Combined with Lemma A.5 and A.8, Theorem 4.2 is proved from (22).

□

C PROOF OF THE NONSPLITTING PROPERTY

The proof of the NSP of SCORE, i.e., Lemma A.10, basically follows the arguments in Jin et al. (2017) and Jin et al. (2022). In other words, Lemma A.10 is an extension of Theorem 3.2 of Jin et al. (2022) to the case of weighted DCSBM. To carry out this extension, we first need some probabilistic results for subexponential random vectors and matrices.

C.1 Preliminaries

Lemma C.1 (Vector Bernstein-type Inequality, Theorem 2.5 of Bosq (2000))**.** *If $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent random vectors in a separable Hilbert space (where the norm is denoted by $\|\cdot\|$) with $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ and*

$$\sum_{i=1}^{n} \mathbb{E}\,\|\boldsymbol{X}_i\|^p \le \frac{p!}{2}\sigma^2 R^{p-2}, \quad p = 2, 3, 4, \ldots.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \boldsymbol{X}_i\right\| \ge t\right) \le 2\exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right).$$

The following lemma is an application of the vector Bernstein-type inequality supporting the proof of Lemma C.7.

Lemma C.2. *Let $X_1, \ldots, X_n$ be independent random variables satisfying*

$$\mathbb{E}[|X_i - \mu_i|^p] \le C_0\left(\frac{p!}{2}\right) R^{p-2}\mu_i,$$

where $\mu_{\max} \coloneqq \max_{1\le i\le n}\mu_i \le C_1$. Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n \in \mathbb{R}^d$ be fixed vectors. Then, with probability $1 - O(n^{-4})$, we have

$$\left\|\sum_{i=1}^{n}(X_i - \mu_i)\boldsymbol{w}_i\right\|_2 \le C\left(\sqrt{\mu_{\max}}\|\boldsymbol{W}\|_F\sqrt{\log n} + \|\boldsymbol{W}\|_{2\to\infty}(\log n)\right), \tag{23}$$

where $\boldsymbol{W}^\top = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]$, $\|\boldsymbol{W}\|_{2\to\infty} = \max_{1\le i\le n}\|\boldsymbol{w}_i\|_2$, and C is a constant that only relies on C_0, R, and C_1.

Proof. For any $p \ge 2$, we have

$$\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}\,\|(X_i - \mu_i)\boldsymbol{w}_i\|_2^p &= \sum_{i=1}^{n}(\mathbb{E}\,|X_i - \mu_i|^p)\|\boldsymbol{w}_i\|_2^p \\
&\le \sum_{i=1}^{n} C_0\left(\frac{p!}{2}\right) R^{p-2}\mu_i\|\boldsymbol{w}_i\|_2^p \\
&\le \sum_{i=1}^{n} C_0\left(\frac{p!}{2}\right) R^{p-2}C_1\|\boldsymbol{w}_i\|_2^2\|\boldsymbol{W}\|_{2\to\infty}^{p-2} \\
&= \left(\frac{p!}{2}\right)(R\|\boldsymbol{W}\|_{2\to\infty})^{p-2}\,(C_0C_1\|\boldsymbol{W}\|_F^2),
\end{aligned}$$

Then, by Lemma C.1, for any $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \lambda_i)\boldsymbol{w}_i\right\|_2 \ge t\right) \le 2\exp\left(-\frac{t^2}{C\mu_{\max}\|\boldsymbol{W}\|_F^2 + (R\|\boldsymbol{W}\|_{2\to\infty})t}\right).$$

Then we get (23) with probability $1 - O(n^{-4})$ for sufficiently large C. □

The following lemma is the subexponential case of the matrix Bernstein inequality.

Lemma C.3 (Theorem 6.2 of Tropp (2012))**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, symmetric matrices with dimension d. Assume that*

$$\mathbb{E}[\boldsymbol{X}_k] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[\boldsymbol{X}_k^p] \preceq \frac{p!}{2}R^{p-2}\boldsymbol{A}_k^2, \quad p = 2, 3, 4, \ldots.$$

Compute the variance parameter

$$\sigma^2 \coloneqq \left\|\sum_k \boldsymbol{A}_k^2\right\|.$$

Then, the following chain of inequalities holds for all $t \geq 0$:

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \boldsymbol{X}_k\right) \geq t\right) \leq d \exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right).$$

The following theorem from Abbe et al. (2020) provides perturbation bounds of eigenspaces which is crucial to the proof of Theorem A.10.

Lemma C.4 (Abbe et al. (2020))**.** *Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric random matrix, and let $\boldsymbol{A}^* = \mathbb{E}[\boldsymbol{A}]$. Denote the eigenvalues of $\boldsymbol{A}$ by $\lambda_1 \geq \cdots \geq \lambda_n$, and their associated eigenvectors by $\{\boldsymbol{u}_j\}_{j=1}^n$. Analogously for $\boldsymbol{A}^*$, the eigenvalues and eigenvectors are denoted by $\lambda_1^* \geq \cdots \geq \lambda_n^*$ and $\{\boldsymbol{u}_j^*\}_{j=1}^n$. For convenience, we also define $\lambda_0 = \lambda_0^* = \infty$ and $\lambda_{n+1} = \lambda_{n+1}^* = -\infty$. Note that we allow eigenvalues to be identical, so some eigenvectors may be defined up to rotations.*

Suppose r and s are two integers satisfying $1 \leq r \leq n$ and $0 \leq s \leq n - r$. Let $\boldsymbol{U} = [\boldsymbol{u}_{s+1}, \ldots, \boldsymbol{u}_{s+r}] \in \mathbb{R}^{n\times r}$, $\boldsymbol{U}^* = [\boldsymbol{u}_{s+1}^*, \ldots, \boldsymbol{u}_{s+r}^*] \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\Lambda}^* = diag(\lambda_{s+1}^*, \ldots, \lambda_{s+r}^*) \in \mathbb{R}^{r\times r}$. Define

$$\Delta^* = (\lambda_s^* - \lambda_{s+1}^*) \wedge (\lambda_{s+r}^* - \lambda_{s+r+1}^*) \wedge \min_{1\leq i\leq r} |\lambda_{s+i}^*|$$

and

$$\kappa = \max_{1\leq i\leq r} |\lambda_{s+i}^*| / \Delta^* .$$

Suppose for some $\gamma \geq 0$, the following assumptions hold:

- *A1 (Incoherence) $\|\boldsymbol{A}^*\|_{2\to\infty} \leq \gamma\Delta^*$.*
- *A2 (Row- and column-wise independence) For any $m \in [n]$, the entries in the m-th row and column of $\boldsymbol{A}$ are independent with other entries: namely, $\{A_{ij}, i = m \text{ or } j = m\}$ are independent of $\{A_{ij} : i \neq m, j \neq m\}$.*
- *A3 (Spectral norm concentration) $32\kappa \max\{\gamma, \varphi(\gamma)\} \leq 1$ and for some $\delta_0 \in (0, 1)$,*

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{A}^*\| \leq \gamma\Delta^*\right) \geq 1 - \delta_0.$$

- *A4 (Row concentration) Suppose $\varphi(x)$ is continuous and non-decreasing in $\mathbb{R}_+$ with $\varphi(0) = 0$, $\varphi(x)/x$ is non-increasing in $\mathbb{R}_+$, and $\delta_1 \in (0, 1)$. For any $i \in [n]$ and $\boldsymbol{W} \in \mathbb{R}^{n\times r}$,*

$$\mathbb{P}\left(\|(\boldsymbol{A} - \boldsymbol{A}^*)_{i\cdot}\boldsymbol{W}\|_2 \leq \Delta^*\|\boldsymbol{W}\|_{2\to\infty}\varphi\left(\frac{\|\boldsymbol{W}\|_F}{\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty}}\right)\right) \geq 1 - \frac{\delta_1}{n}.$$

Under Assumptions (A1)—(A4), with probability at least $1 - \delta_0 - 2\delta_1$, there exists an orthogonal matrix $\boldsymbol{Q}$ such that

$$\|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{A}\boldsymbol{U}^*(\boldsymbol{\Lambda}^*)^{-1})\|_{2\to\infty} \lesssim \kappa(\kappa + \varphi(1))(\gamma + \varphi(\gamma))\|\boldsymbol{U}^*\|_{2\to\infty} + \gamma\|\boldsymbol{A}^*\|_{2\to\infty} \,/\Delta^* . \tag{24}$$

Here $\lesssim$ only hides absolute constants.

C.2 Spectral Properties of the Adjacency Matrix

Now, we introduce some spectral properties of the mean adjacency matrix $\boldsymbol{M}$ and the observed weighted adjacency matrix $\boldsymbol{A}$, which are useful in proving the NSP of SCORE. We begin by studying the eigenvalues and eigenvectors of $\boldsymbol{M}$. The population adjacency matrix $\boldsymbol{M}$ is a rank-K matrix, with nonzero eigenvalues $\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*$ sorted in descending magnitude and their corresponding unit-norm eigenvectors $\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_K^*$. Notice that by Perron's theorem (Horn and Johnson, 2012), λ_1^* is positive with multiplicity 1, and we can choose $\boldsymbol{u}_1^*$ such that all its entries are strictly positive. Denote $\boldsymbol{U}^* = [\boldsymbol{u}_2^*, \ldots, \boldsymbol{u}_K^*] \in \mathbb{R}^{n\times(K-1)}$ and $\left(\boldsymbol{U}_{i\cdot}^*\right)^\top$ as its i-th row. Denote $\boldsymbol{\Lambda}^* = \mathrm{diag}(\lambda_2^*, \ldots, \lambda_K^*) \in \mathbb{R}^{(K-1)\times(K-1)}$.

Lemma C.5 (Lemma B.1 of Jin et al. (2017))**.** *Under Assumption 1, the following statements are true:*

1. *$|\lambda_k^*| \asymp \|\boldsymbol{\theta}\|_2^2$, $1 \leq k \leq K$.*

2. $\lambda_1^* - |\lambda_2^*| \asymp \lambda_1^*$.

Proof. This result is basically the same as Lemma B.1 of Jin et al. (2017) under Assumption 1. In particular, $\lambda_1^* - |\lambda_2^*| \asymp \lambda_1^*$ can be straightforwardly obtained from Han and Han (2019). □

Also, we have the following properties of the eigenvectors of $\boldsymbol{M}$:

Lemma C.6 (Lemma B.2 of Jin et al. (2017))**.** *Under Assumption 1, the following statements are true:*

1. *If we choose the sign of $\boldsymbol{u}_1^*$ such that $\sum_{i=1}^n u_1^*(i) > 0$, then the entries of $\boldsymbol{u}_1^*$ are positive satisfying $C^{-1}\theta_i/\|\boldsymbol{\theta}\|_2 \le u_1^*(i) \le C\theta_i/\|\boldsymbol{\theta}\|_2$, $1 \le i \le n$.*
2. *$\|\boldsymbol{U}_{i\cdot}^*\|_2 \le C\sqrt{K}\theta_i/\|\boldsymbol{\theta}\|_2$, $1 \le i \le n$.*

Here, C is a constant that depends only on c_0.

Now, we come back to community detection with SCORE. Let $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_K|$ be the leading K eigenvalues of $\boldsymbol{A}$ in magnitude, with corresponding unit-norm eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$. Denote $\boldsymbol{U} = [\boldsymbol{u}_2, \ldots, \boldsymbol{u}_K] \in \mathbb{R}^{n\times(K-1)}$. The proof of the following lemma is similar to that of Lemma 2.1 in Jin et al. (2017), but requires some adaptions for the weighted DCSBM. We defer the proof to Section C.4.

Lemma C.7. *Under Assumption 1, with probability $1 - O(n^{-3})$, the following statements are true:*

1. *We can select $\boldsymbol{u}_1$ such that $\|\boldsymbol{u}_1 - \boldsymbol{u}_1^*\|_\infty \le \frac{C(c_0)}{\sqrt{n\log n}}$.*
2. *$\|\boldsymbol{U}\boldsymbol{Q}^\top - \boldsymbol{U}^*\|_{2\to\infty} \le \frac{C(c_0)}{\sqrt{n\log n}}$ for some $\boldsymbol{Q} \in \boldsymbol{O}_{K-1}$.*

C.3 Proof of Lemma A.10

Using the above lemmas, we can prove Lemma A.10 by following the arguments in Jin et al. (2022). Here, we give a brief outline.

Define $\boldsymbol{R}^{(K)}$ as an $n\times(K-1)$ matrix constructed from the eigenvectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, ..., $\boldsymbol{u}_K$ by taking entrywise ratios of $\boldsymbol{u}_2$,..., $\boldsymbol{u}_K$ to $\boldsymbol{u}_1$, i.e., $R^{(K)}(i,k) = u_{k+1}(i)/u_1(i)$ for $1 \le i \le n$ and $0 \le k \le K-1$. For any $2 \le m \le K$, let $\boldsymbol{R}^{(m)}$ be an $n\times(m-1)$ matrix consists of the first $m-1$ columns of $\boldsymbol{R}^{(K)}$. For any candidate number of clusters m, SCORE amounts to performing k-means clustering on the rows of $\boldsymbol{R}^{(m)}$. Therefore, we need to study the statistical properties of the rows of $\boldsymbol{R}^{(m)}$.

Next, we define the population counterparts. Denote $\boldsymbol{O}_{K-1}$ as the space of $(K-1)\times(K-1)$ orthogonal matrices. For any $\boldsymbol{Q} \in \boldsymbol{O}_{K-1}$ and any $2 \le k \le K$, let $\boldsymbol{u}_k^*(\boldsymbol{Q})$ be the $(k-1)$-th column of $[\boldsymbol{u}_2^*, \ldots, \boldsymbol{u}_K^*]\boldsymbol{Q}$. Define $\boldsymbol{R}^{*(K)}(\boldsymbol{Q}) \in \mathbb{R}^{n\times(K-1)}$ such that its $(k-1)$-th column is the entrywise ratio of $\boldsymbol{u}_k^*(\boldsymbol{Q})$ to $\boldsymbol{u}_1^*$. For any $2 \le m \le K$, let $\boldsymbol{R}^{*(m)}(\boldsymbol{Q}) \in \mathbb{R}^{n\times(m-1)}$ consist of the first $m-1$ columns of $\boldsymbol{R}^{*(K)}(\boldsymbol{Q})$. The following lemma provides a uniform upper bound on the difference between the corresponding rows of $\boldsymbol{R}^{(m)}$ and $\boldsymbol{R}^{*(m)}(\boldsymbol{Q})$.

Lemma C.8 (Row-wise deviation bound)**.** *For $2 \le m \le K$, denote $\left(\boldsymbol{r}_i^{(m)}\right)^\top$ and $\left(\boldsymbol{r}_i^{*(m)}(\boldsymbol{Q})\right)^\top$ as the i-th row of $\boldsymbol{R}^{(m)}$ and $\boldsymbol{R}^{*(m)}(\boldsymbol{Q})$, respectively. Under Assumption 1, with probability $1-O(n^{-3})$, there exists a $(K-1)\times(K-1)$ orthogonal matrix $\boldsymbol{Q}$, such that*

$$\|\boldsymbol{r}_i^{(m)} - \boldsymbol{r}_i^{*(m)}(\boldsymbol{Q})\|_2 \le \|\boldsymbol{r}_i^{(K)} - \boldsymbol{r}_i^{*(K)}(\boldsymbol{Q})\|_2 \le \frac{C(c_0)}{\sqrt{\log n}}, \tag{25}$$

for each $2 \le m \le K$ and $1 \le i \le n$.

This lemma is a straightforward corollary of Lemmas C.6 and C.7 by following the proof of Lemma 4.1 in Jin et al. (2022), so we omit the details.

Combining Lemma C.8 with Lemma 4.2, Lemma 4.3 and Theorem 4.1 in Jin et al. (2022), Lemma A.10 is proved by exactly the same arguments as in Jin et al. (2022).

C.4 Proof of Lemma C.7

Proof. Divide $\lambda_1^*, \ldots, \lambda_K^*$ into three groups: (i) λ_1^*, (ii) positive values in $\lambda_2^*, \ldots, \lambda_K^*$, and (iii) negative values in $\lambda_2^*, \ldots, \lambda_K^*$. We shall apply Lemma C.4 to all three groups. For succinctness, we only show in detail the application to group (ii), while the proofs for the other two groups are similar and thus omitted.

Denote K_1 as the number of eigenvalues in group (ii). Define $\boldsymbol{\Lambda}_1^*$ as the diagonal matrix consisting of eigenvalues in group (ii), and $\boldsymbol{U}_1^*$ as the matrix whose columns are the associated eigenvectors. Define the empirical counterparts of the two matrices as $\boldsymbol{\Lambda}_1$ and $\boldsymbol{U}_1$. To show the second statement in the lemma, we aim to first show with high probability that

(a) $\|\boldsymbol{U}_1\boldsymbol{Q}^\top - \boldsymbol{A}\boldsymbol{U}_1^*(\boldsymbol{\Lambda}_1^*)^{-1}\|_{2\to\infty} \le \frac{C(c_0)}{\sqrt{n\log n}}$ for some $\boldsymbol{Q} \in \boldsymbol{O}_{K_1}$.

(b) $\|\boldsymbol{U}_1^* - \boldsymbol{A}\boldsymbol{U}_1^*(\boldsymbol{\Lambda}_1^*)^{-1}\|_{2\to\infty} \le \frac{C(c_0)}{\sqrt{n\log n}}$.

Proof of (a) To apply Lemma C.4, we need to determine γ and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$, and then verify Assumption (A1)—(A4) as required by the theorem. Note that by Lemma C.5, we have

$$\Delta^* = \min\left\{\lambda_1^* - \lambda_2^*, |\lambda_K^*|\right\} \ge C(c_0)\theta_{\min}^2 n,$$

and $\kappa \le C(c_0)$. We choose an appropriately large $C_1(c_0)$ and let

$$\gamma = \frac{C_1(c_0)}{\sqrt{\log n}}.$$

A1 This is trivial since

$$\|\boldsymbol{M}\|_{2\to\infty} \le c_0^{-2}\theta_{\min}^2\sqrt{n} \le \Delta^*\gamma,$$

when n is sufficiently large.

A2 This obviously holds.

A3 In the proof of A4, we will show that $\varphi(\gamma) \le C(c_0)(\log n)^{-3/2}$, hence, $32\kappa\max\{\gamma, \varphi(\gamma)\} \le 1$ when n is sufficiently large.

Next, we show the spectral norm perturbation bound by applying the subexponential case of matrix Bernstein inequality (Theorem 6.2 of Tropp (2012), restated as Lemma C.3 in this paper). Let $\boldsymbol{X}^{ij} = (A_{ij} - M_{ij})(\boldsymbol{E}^{ij} + \boldsymbol{E}^{ji})$ for $i < j$, and $\boldsymbol{X}^{ii} = (A_{ii} - M_{ii})\boldsymbol{E}^{ii}$ for $i = 1, \ldots, n$, where $\boldsymbol{E}^{ij}$ is a $n \times n$ matrix with 1 on the (i, j)-th entry and 0 elsewhere. Obviously, $\mathbb{E}[\boldsymbol{X}^{ij}] = \boldsymbol{0}$. Notice that

$$\mathbb{E}[(\boldsymbol{X}^{ij})^p] = \begin{cases} \mathbb{E}[(A_{ij} - M_{ij})^p](\boldsymbol{E}^{ij} + \boldsymbol{E}^{ji}) & \text{when } p \text{ is odd;} \\ \mathbb{E}[(A_{ij} - M_{ij})^p](\boldsymbol{E}^{ij} + \boldsymbol{E}^{ji}) & \text{when } p \text{ is even.} \end{cases}$$

Also, note that $-(\boldsymbol{E}^{ii} + \boldsymbol{E}^{jj}) \preceq \boldsymbol{E}^{ij} + \boldsymbol{E}^{ji} \preceq \boldsymbol{E}^{ii} + \boldsymbol{E}^{jj}$. Then, by (15) in Assumption 1, for any integer $p \ge 2$, we have

$$\left|\mathbb{E}[(A_{ij} - M_{ij})^p]\right| \le \mathbb{E}[|A_{ij} - M_{ij}|^p] \le C'\left(\frac{p!}{2}\right)R^{p-2}M_{ij},$$

where C' and R only depend on c_0. Then,

$$\mathbb{E}[(\boldsymbol{X}^{ij})^p] \preceq C'\frac{p!}{2}R^{p-2}M_{ij}(\boldsymbol{E}^{ii} + \boldsymbol{E}^{jj}), \quad p = 2, 3, 4, \ldots.$$

Thus, the conditions of Lemma C.3 are verified. Notice that $\sum_{i\le j}\boldsymbol{X}^{ij} = \boldsymbol{A} - \boldsymbol{M}$. Denote

$$\sigma^2 = \left\|\sum_{i\le j} C'M_{ij}(\boldsymbol{E}^{ii} + \boldsymbol{E}^{jj})\right\| = C'\left(\max_i \sum_{j=1}^n M_{ij}\right) \lesssim \theta_{\min}^2 n,$$

where $\lesssim$ only hides a constant depending on c_0. Then, for all $t \geq 0$,

$$\mathbb{P}\left(\|\boldsymbol{A}-\boldsymbol{M}\| \geq t\right) \leq n \exp\left(-\frac{t^2}{2(\sigma^2+Rt)}\right) \leq n \exp\left(-\frac{1}{4}\left(\frac{t^2}{\sigma^2} \wedge \frac{t}{R}\right)\right).$$

By (12) in Assumption 1, when n is sufficiently large, with probability $1-O(n^{-3})$, for a sufficiently large $C(c_0)$,

$$\|\boldsymbol{A}-\boldsymbol{M}\| \leq C(c_0)\theta_{\min}\sqrt{n\log n}.$$

A4 By Lemma C.2, for any $1 \leq i \leq n$ and $\boldsymbol{W} \in \mathbb{R}^{n\times r}$, with probability $1-O(n^{-4})$, there holds

$$\|(\boldsymbol{A}-\boldsymbol{M})_{i\cdot}\boldsymbol{W}\|_2 \leq C_2(c_0)\max\left(\theta_{\min}\|\boldsymbol{W}\|_F\sqrt{\log n}, \|\boldsymbol{W}\|_{2\to\infty}(\log n)\right), \tag{26}$$

for a sufficiently large $C_2(c_0)$. Since $\Delta^* \geq C(c_0)\theta_{\min}^2 n$, we have

$$\|(\boldsymbol{A}-\boldsymbol{M})_{i\cdot}\boldsymbol{W}\|_2 \leq C_2(c_0)\Delta^*\|\boldsymbol{W}\|_{2\to\infty}\max\left(\frac{\sqrt{n\log n}}{\theta_{\min}n}\frac{\|\boldsymbol{W}\|_F}{\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty}}, \frac{\log n}{\theta_{\min}^2 n}\right).$$

Now, we follow similar arguments as in the proof of Lemma 2.1 of Jin et al. (2017). Define the quantities $t_1 = C_2(c_0)\left(\theta_{\min}n\right)^{-1}\sqrt{n\log n}$ and $t_2 = C_2(c_0)\left(\theta_{\min}^2 n\right)^{-1}\log n$. Define the function

$$\widetilde{\varphi}(x) = \max\left(t_1 x, t_2\right).$$

Then, we have for any $1 \leq i \leq n$, with probability $1-O(n^{-4})$,

$$\|(\boldsymbol{A}-\boldsymbol{M})_{i\cdot}\boldsymbol{W}\|_2 \leq \Delta^*\|\boldsymbol{W}\|_{2\to\infty}\widetilde{\varphi}\left(\frac{\|\boldsymbol{W}\|_F}{\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty}}\right). \tag{27}$$

First, notice that $(\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty})^{-1}\|\boldsymbol{W}\|_F \in \left[n^{-1/2}, 1\right]$. Next, observe that by (12), when n is sufficiently large, we have $t_2/t_1 = (\theta_{\min}\sqrt{n})^{-1}\sqrt{\log n} > n^{-1/2}$. Therefore, when $x \in \left[n^{-1/2}, t_2/t_1\right]$, $t_1 x \leq t_2$, i.e., $\widetilde{\varphi}(x) = t_2$; when $x \in (t_2/t_1, 1]$, $t_1 x > t_2$, i.e., $\widetilde{\varphi}(x) = t_1 x$. As a result, we can construct $\varphi(\cdot)$ as:

$$\varphi(x) = \begin{cases} \sqrt{n}t_2 x & \text{for } 0 \leq x \leq n^{-1/2}; \\ t_2 & \text{for } n^{-1/2} < x \leq t_2/t_1; \\ t_1 x & \text{for } t_2/t_1 < x \leq 1; \\ t_1 & \text{for } x > 1. \end{cases}$$

Obviously, $\varphi(x)$ is continuous and non-decreasing in $\mathbb{R}_+$, with $\varphi(0) = 0$ and $\varphi(x)/x$ being non-increasing in $\mathbb{R}_+$. By (27) and $0 \leq \widetilde{\varphi}(x) \leq \varphi(x)$, we have with probability $1-O(n^{-4})$,

$$\|(\boldsymbol{A}-\boldsymbol{M})_{i\cdot}\boldsymbol{W}\|_2 \leq \Delta^*\|\boldsymbol{W}\|_{2\to\infty}\varphi\left(\frac{\|\boldsymbol{W}\|_F}{\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty}}\right).$$

By the definition of t_1 and (12), we have $t_1 \leq (\log n)^{-5/2}$. Furthermore, since $\varphi(x) \leq t_1$, we have $\varphi(\gamma) \leq (\log n)^{-5/2}$.

After verifying (A1)—(A4), we obtain the bound (24). Note that based on the definition of $\varphi(x)$, we have $\kappa(\kappa+\varphi(1)) \leq C(c_0)$ and $\gamma+\varphi(\gamma) \leq \frac{C(c_0)}{\sqrt{\log n}}$. Also, by Lemma C.6, $\|\boldsymbol{U}_1^*\|_{2\to\infty} \leq \frac{C(c_0)}{\sqrt{n}}$. Since $\|\boldsymbol{M}\|_{2\to\infty} \leq c_0^{-2}\theta_{\min}^2\sqrt{n}$ and $\Delta^* \geq C(c_0)\theta_{\min}^2 n$, we have $\|\boldsymbol{M}\|_{2\to\infty}/\Delta^* \leq \frac{C(c_0)}{\sqrt{n}}$. Above all, we obtain that, with probability $1-O(n^{-3})$,

$$\begin{aligned}\left\|\boldsymbol{U}_1\boldsymbol{Q}-\boldsymbol{A}\boldsymbol{U}_1^*(\boldsymbol{\Lambda}_1^*)^{-1}\right\|_{2\to\infty} &\lesssim \kappa(\kappa+\varphi(1))(\gamma+\varphi(\gamma))\left\|\boldsymbol{U}_1^*\right\|_{2\to\infty} + \gamma\,\|\boldsymbol{M}\|_{2\to\infty}/\Delta^* \\ &\lesssim \frac{1}{\sqrt{n\log n}},\end{aligned}$$

where $\lesssim$ only hides a constant depending on c_0 in Assumption 1.

Proof of (b) Based on the fact $\boldsymbol{U}_1^* = \boldsymbol{M}\boldsymbol{U}_1^*(\boldsymbol{\Lambda}_1^*)^{-1}$, we have

$$\left\|\boldsymbol{U}_1^* - \boldsymbol{A}\boldsymbol{U}_1^*\left(\boldsymbol{\Lambda}_1^*\right)^{-1}\right\|_{2\to\infty} = \left\|(\boldsymbol{M}-\boldsymbol{A})\boldsymbol{U}_1^*\left(\boldsymbol{\Lambda}_1^*\right)^{-1}\right\|_{2\to\infty}.$$

By Lemma C.6 and assumption (12), applying (26) with $\boldsymbol{W} = \boldsymbol{U}_1^*$ yields that with probability $1 - O\left(n^{-3}\right)$,

$$\begin{aligned}
\left\|(\boldsymbol{A}-\boldsymbol{M})\boldsymbol{U}_1^*(\boldsymbol{\Lambda}_1^*)^{-1}\right\|_{2\to\infty} &\le \left(\max_{1\le i\le n}\left\|(\boldsymbol{A}-\boldsymbol{M})_{i\cdot}\boldsymbol{U}_1^*\right\|_2\right)\cdot\left\|\left(\boldsymbol{\Lambda}_1^*\right)^{-1}\right\| \\
&\le C(c_0)\max\left(\theta_{\min}\left\|\boldsymbol{U}_1^*\right\|_F\sqrt{\log n},\ \left\|\boldsymbol{U}_1^*\right\|_{2\to\infty}(\log n)\right)\cdot\left(\theta_{\min}^2 n\right)^{-1} \\
&\le C(c_0)\max\left(\left(\theta_{\min}\sqrt{n}\right)^{-1}\left\|\boldsymbol{U}_1^*\right\|_F\sqrt{\frac{\log n}{n}},\ \left(\theta_{\min}\sqrt{n}\right)^{-2}\left\|\boldsymbol{U}_1^*\right\|_{2\to\infty}(\log n)\right) \\
&\lesssim \frac{1}{\sqrt{n\log^5 n}},
\end{aligned}$$

where $\lesssim$ only hides a constant depending on c_0 in Assumption 1.

With (a) and (b), we have

$$\left\|\boldsymbol{U}_1\boldsymbol{Q}^\top - \boldsymbol{U}_1^*\right\|_{2\to\infty} \lesssim \frac{1}{\sqrt{n\log n}} \quad \text{for some } \boldsymbol{Q}\in\mathcal{O}_{K_1}.$$

For eigenvalues in group (iii), we similarly have

$$\left\|\boldsymbol{U}_2\boldsymbol{Q}^\top - \boldsymbol{U}_2^*\right\|_{2\to\infty} \lesssim \frac{1}{\sqrt{n\log n}} \quad \text{for some } \boldsymbol{Q}\in\mathcal{O}_{K_2}.$$

Combining these results yields the second statement in this lemma: With probability $1 - O(n^{-3})$, we have

$$\left\|\boldsymbol{U}\boldsymbol{Q}^\top - \boldsymbol{U}^*\right\|_{2\to\infty} \lesssim \frac{1}{\sqrt{n\log n}} \quad \text{for some } \boldsymbol{Q}\in\mathcal{O}_{K-1}.$$

To show the first statement, we apply Lemma C.4 to group (i) with $s = 0$ and $r = 1$. Note that by Lemma C.5, we have

$$\Delta^* = \min\{\lambda_1^*, \lambda_1^* - \lambda_2^*\} \ge C(c_0)\theta_{\min}^2 n$$

and $\kappa \le C(c_0)$. Following similar procedures, we can select $\boldsymbol{u}_1$ such that

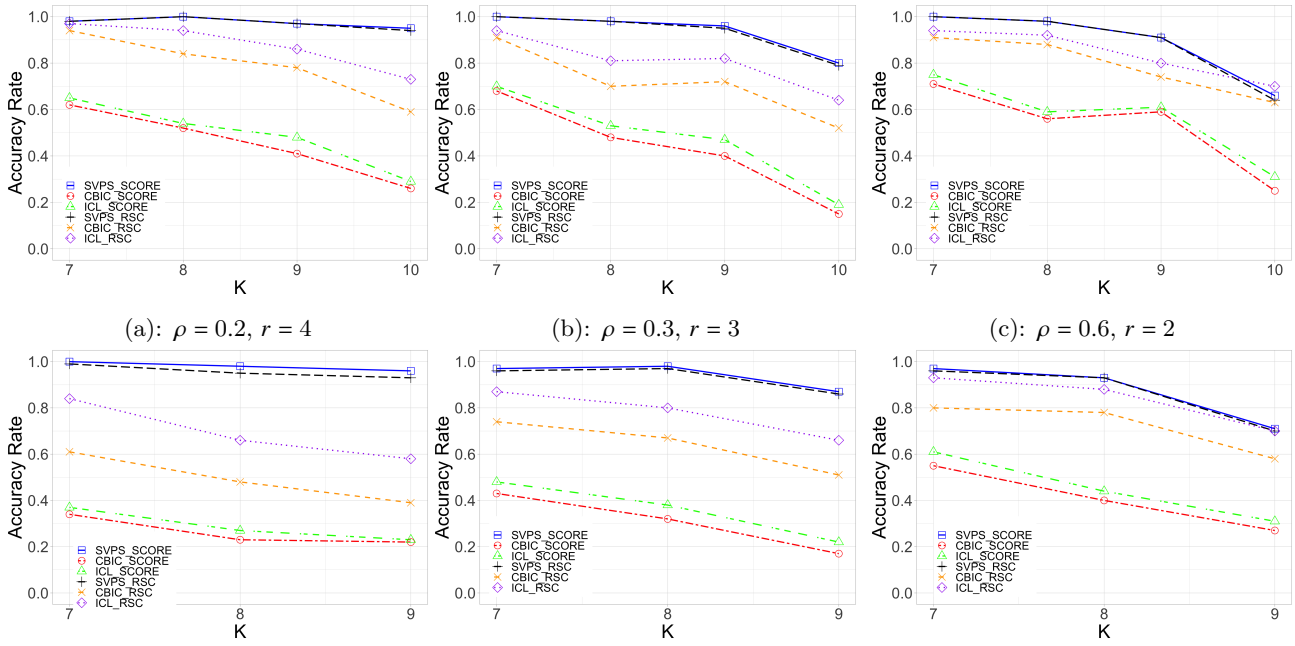
(a) $\left\|\boldsymbol{u}_1 - \boldsymbol{A}\boldsymbol{u}_1^*/\lambda_1^*\right\|_\infty \le \frac{C(c_0)}{\sqrt{n\log n}}$.

(b) $\left\|\boldsymbol{u}_1^* - \boldsymbol{A}\boldsymbol{u}_1^*/\lambda_1^*\right\|_\infty \le \frac{C(c_0)}{\sqrt{n\log n}}$.

Therefore, with probability $1 - O(n^{-3})$, we have

$$\left\|\boldsymbol{u}_1 - \boldsymbol{u}_1^*\right\|_\infty \le \frac{C(c_0)}{\sqrt{n\log n}},$$

which proves the first statement. □



(a): $\rho = 0.2$, $r = 4$ (b): $\rho = 0.3$, $r = 3$ (c): $\rho = 0.6$, $r = 2$

(d): $\rho = 0.09$, $r = 4$ (e): $\rho = 0.15$, $r = 3$ (f): $\rho = 0.3$, $r = 2$

Figure 4: Accuracy rate plots of experiments in Appendix D. The top and bottom rows correspond to Setting 1 and 2, respectively.

D ADDITIONAL EXPERIMENTS

This section continues from Section 5.1, presenting additional results of synthetic simulations with larger values of K. The generating mechanism of the weighted DCSBM and experiment setup follow from Section 5.1. We consider the following two settings under the Poisson DCSBM:

- **Setting 1:** $\boldsymbol{n}_{all} = (20, 60, 20, 60, 20, 60, 20, 60, 20, 60)$; $K = 7, 8, 9, 10$
- **Setting 2:** $\boldsymbol{n}_{all} = (30, 60, 90, 30, 60, 90, 30, 60, 90)$; $K = 7, 8, 9$

The threshold used in SVPS is 2.02. The plots of the accuracy rates of the compared methods are shown in Figure 4, with figure captions indicating the choice of (ρ, r). As we can see, SVPS still achieves higher accuracy rate, especially when using SCORE for spectral clustering, the performance gap is very significant.