Bernoulli **28**(2), 2022, 1049–1073
https://doi.org/10.3150/21-BEJ1376

Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral estimator

MINH TANG[1], JOSHUA CAPE[2] and CAREY E. PRIEBE[3]

[1]*Department of Statistics, North Carolina State University, Raleigh, NC, USA . E-mail: mtang8@ncsu.edu*
[2]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA . E-mail: joshua.cape@pitt.edu*
[3]*Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA . E-mail: cep@jhu.edu*

We establish asymptotic normality results for estimation of the block probability matrix **B** in stochastic blockmodel graphs using spectral embedding when the average degrees grows at the rate of $\omega(\sqrt{n})$ in n, the number of vertices. As a corollary, we show that when **B** is of full-rank, estimates of **B** obtained from spectral embedding are asymptotically efficient. When **B** is singular the estimates obtained from spectral embedding can have smaller mean square error than those obtained from maximizing the log-likelihood under no rank assumption, and furthermore, can be almost as efficient as the true MLE that assumes the rank of **B** is known. Our results indicate, in the context of stochastic blockmodel graphs, that spectral embedding is not just computationally tractable, but that the resulting estimates are also admissible, even when compared to the purportedly optimal but computationally intractable maximum likelihood estimation under no rank assumption.

Keywords: Asymptotic efficiency; random dot product graph; stochastic blockmodels; asymptotic normality; spectral embedding

1. Introduction

Statistical inference on graphs is a burgeoning field of research in machine learning and statistics, with numerous applications to social network, neuroscience, etc. Many of the graphs in application domains are large and complex but nevertheless are believed to be composed of multiple smaller-scale communities. Thus, an essential task in graph inference is detecting/identifying local (sub)communities.

The resulting problem of community detection on graphs is well-studied (see the surveys [1,23]), with many available techniques including those based on maximizing modularity and likelihood [8, 12,43,50], random walks [45,47], spectral clustering [35,42,46,51,55], and semidefinite programming [2,26]. It is widely known that under suitable models — such as the popular stochastic blockmodel and its variants [29,31,38] — one can consistently recover the underlying communities as the number of observed nodes increases, and furthermore, there exist deep and beautiful phase transitions phenomena with respect to the statistical and computational limits for recovery.

Another important question in graph inference is, subsequent to community detection, that of characterizing the nature and/or structure of these communities. One of the simplest and possibly most essential examples is determining the probabilities for nodes to form link within and between communities. Consistent recovery of the underlying communities yields a straightforward and universally employed procedure for consistent estimation of these probabilities, namely averaging the number of edges within a community and/or between communities. This procedure, when interpreted in the context of estimating the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ of a collection of independent binomially

distributed random variables $\{X_1, X_2, \ldots, X_r\}$ with $X_i \sim \text{Bin}(n_i, \theta_i)$, corresponds to maximum likelihood estimation with no restrictive assumption on the parameters $\{\theta_i\}$. However, in the context of graphs and communities, there are often natural relationships among the communities; hence a graph with K communities need not require $K(K+1)/2$ parameters to describe the within and between community connection probabilities. The above procedure is therefore potentially sub-optimal.

Motivated by the above observation, our paper studies the asymptotic properties of three different estimators for $\mathbf{B}$, the matrix of edge probabilities among communities, of a stochastic blockmodel graph. Two estimators are based on maximum likelihood methods and the remaining estimator is based on spectral embedding. We show that, given an observed graph with adjacency matrix $\mathbf{A}$, the most commonly used estimator — the MLE under no rank assumption on $\mathbf{B}$ — is sub-optimal when $\mathbf{B}$ is not invertible. Moreover, when $\mathbf{B}$ is singular, the estimator based on spectral embedding $\mathbf{A}$ is often times better (smaller mean squared error) than the MLE under no rank assumption, and is almost as efficient as the asymptotically (first-order) efficient MLE whose parametrization depend on $\text{rk}(\mathbf{B})$. Finally, when $\mathbf{B}$ is invertible, the three estimators are asymptotically first-order efficient.

1.1. Background

We now formalize the setting considered in this paper. We begin by recalling the notion of stochastic blockmodel graphs due to [29]. Stochastic blockmodel graphs and its variants, such as degree-corrected blockmodels and mixed membership models [4,31] are the most popular models for graphs with intrinsic community structure. In addition, they are widely used as building blocks for constructing approximations (see e.g., [25,32,58]) of the more general latent position graphs or graphons models [28,36].

Definition 1. Let $K \geq 1$ be a positive integer and let $\boldsymbol{\pi} \in \mathcal{S}_{K-1}$ be a non-negative vector in $\mathbb{R}^K$ with $\sum_k \pi_k = 1$; here $\mathcal{S}_{K-1}$ denote the $K-1$ dimensional simplex in $\mathbb{R}^K$. Let $\mathbf{B} \in [0,1]^{K \times K}$ be symmetric. We say that $(\mathbf{A}, \boldsymbol{\tau}) \sim \text{SBM}(\mathbf{B}, \boldsymbol{\pi})$ with sparsity factor ρ if the following hold. First $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ where τ_i are i.i.d. with $\mathbb{P}[\tau_i = k] = \pi_k$. Then $\mathbf{A} \in \{0,1\}^{n \times n}$ is a symmetric matrix such that, conditioned on $\boldsymbol{\tau}$, for all $i \leq j$ the $\mathbf{A}_{ij}$ are independent Bernoulli random variables with $\mathbb{E}[\mathbf{A}_{ij}] = \rho \mathbf{B}_{\tau_i, \tau_j}$. We write $\mathbf{A} \sim \text{SBM}(\mathbf{B}, \boldsymbol{\pi})$ when only $\mathbf{A}$ is observed, i.e., $\boldsymbol{\tau}$ is integrated out from $(\mathbf{A}, \boldsymbol{\tau})$.

For $(\mathbf{A}, \boldsymbol{\tau}) \sim \text{SBM}(\mathbf{B}, \boldsymbol{\pi})$ or $\mathbf{A} \sim \text{SBM}(\mathbf{B}, \boldsymbol{\pi})$ with $\mathbf{A}$ having n vertices and (known) sparsity factor ρ, the likelihood of $(\mathbf{A}, \tau)$ and $\mathbf{A}$ are, respectively

$$L(\mathbf{A}, \tau; \mathbf{B}, \boldsymbol{\pi}) = \Big(\prod_{i=1}^{n} \pi_{\tau_i} \Big) \Big(\prod_{i \leq j} (\rho \mathbf{B}_{\tau_i, \tau_j})^{\mathbf{A}_{ij}} (1 - \rho \mathbf{B}_{\tau_i, \tau_j})^{1 - \mathbf{A}_{ij}} \Big), \tag{1.1}$$

$$L(\mathbf{A}; \mathbf{B}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\tau} \in [K]^n} L(\mathbf{A}, \boldsymbol{\tau}; \mathbf{B}, \boldsymbol{\pi}). \tag{1.2}$$

When we observed $\mathbf{A} \sim \text{SBM}(\mathbf{B}, \boldsymbol{\pi})$ with $\mathbf{B} \in [0,1]^{K \times K}$ where ρ and K are assumed known, a maximum likelihood estimate of $\boldsymbol{\pi}$ and $\mathbf{B}$ is given by

$$(\hat{\mathbf{B}}^{(N)}, \hat{\boldsymbol{\pi}}) = \underset{\mathbf{B} \in [0,1]^{K \times K},\ \boldsymbol{\pi} \in \mathcal{S}_{K-1}}{\text{argmax}} L(\mathbf{A}; \mathbf{B}, \boldsymbol{\pi}). \tag{1.3}$$

The maximum likelihood estimate (MLE) $\hat{\mathbf{B}}^{(N)}$ in Eq. (1.3) requires estimation of the $K(K+1)/2$ entries of $\mathbf{B}$ and is generally intractable as it requires marginalization over the latent vertex-to-block

assignment vector $\boldsymbol{\tau}$. Another parametrization of $\mathbf{B}$ is via the eigendecomposition $\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ where $d = \mathrm{rk}(\mathbf{B})$, $\mathbf{V}$ is a $K \times d$ matrix with $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ and $\mathbf{D}$ is diagonal. This parametrization results in the estimation of $d(2K - d + 1)/2 \leq K(K+1)/2$ parameters, with $Kd - d(d+1)/2$ parameters being estimated for $\mathbf{V}$ (as an element of the Stiefel manifold of orthonormal d frames in $\mathbb{R}^K$) and d parameters estimated for $\mathbf{D}$. Therefore, when $d = \mathrm{rk}(\mathbf{B})$ is also assumed known, another MLE of $\mathbf{B}$ and $\boldsymbol{\pi}$ is given by

$$(\hat{\mathbf{B}}^{(M)}, \hat{\boldsymbol{\pi}}) = \underset{\mathbf{B}\in[0,1]^{K\times K},\ \mathrm{rk}(\mathbf{B})=d,\ \boldsymbol{\pi}\in\mathcal{S}_{K-1}}{\operatorname{argmax}} L(\mathbf{A}; \mathbf{B}, \boldsymbol{\pi}). \tag{1.4}$$

The MLE parametrization in Eq. (1.3) is the one that is universally used, see e.g., [7,12,14,50]; variants of MLE estimation such as maximization of the profile likelihood or variational inference [8,17] are also based on approximating the MLE in Eq. (1.3). In contrast, the MLE parametrization used in Eq. (1.4) has, to the best of our knowledge, never been considered heretofore in the literature. We refer to $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(M)}$ as the "naive" MLE and the "true" or rank-constrained MLE, respectively.

The estimator $\hat{\mathbf{B}}^{(N)}$ is asymptotically normal around $\mathbf{B}$; in particular Lemma 1 (and its proof) in [7] reveals that

Theorem 1 ([7]). *Let $\mathbf{A}_n \sim \mathrm{SBM}(\mathbf{B}, \boldsymbol{\pi})$ for $n \geq 1$ be a sequence of stochastic blockmodel graphs with sparsity factors ρ_n. Let $\hat{\mathbf{B}}^{(N)}$ be the MLE of $\mathbf{B}$ obtained from $\mathbf{A}_n$ with ρ_n assumed known. If $\rho_n \equiv 1$ for all n, then*

$$n(\hat{\mathbf{B}}^{(N)}_{kk} - \mathbf{B}_{kk}) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{2\mathbf{B}_{kk}(1-\mathbf{B}_{kk})}{\pi_k^2}\Big), \quad \text{for } k \in [K] \tag{1.5}$$

$$n(\hat{\mathbf{B}}^{(N)}_{k\ell} - \mathbf{B}_{k\ell}) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{\mathbf{B}_{k\ell}(1-\mathbf{B}_{k\ell})}{\pi_k\pi_\ell}\Big), \quad \text{for } k \in [K], \ell \in [K], k \neq \ell \tag{1.6}$$

as $n \to \infty$, and the $K(K+1)/2$ random variables $\{n(\hat{\mathbf{B}}^{(N)}_{k\ell} - \mathbf{B}_{k\ell})\}_{k\leq\ell}$ are asymptotically independent. If, however, $\rho_n \to 0$ with $n\rho_n = \omega(\log n)$, then

$$n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(N)}_{kk} - \mathbf{B}_{kk}) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{2\mathbf{B}_{kk}}{\pi_k^2}\Big), \quad \text{for } k \in [K] \tag{1.7}$$

$$n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(N)}_{k\ell} - \mathbf{B}_{k\ell}) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{\mathbf{B}_{k\ell}}{\pi_k\pi_\ell}\Big), \quad \text{for } k \in [K], \ell \in [K], k \neq \ell \tag{1.8}$$

as $n \to \infty$, and the $K(K+1)/2$ random variables $\{n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(N)}_{k\ell} - \mathbf{B}_{k\ell}\}_{k\leq\ell}$ are asymptotically independent.

Furthermore, $\hat{\mathbf{B}}^{(N)}$ is also purported to be optimal; Section 5 of [7] states that

> These results easily imply that classical optimality properties of these procedures such as achievement of the information bound, hold.

We present Theorem 2 and the examples in Section 2 to illustrate that $\hat{\mathbf{B}}^{(N)}$ is optimal only if $\mathbf{B}$ is invertible, and that for singular $\mathbf{B}$, the covariance matrix of $\hat{\mathbf{B}}^{(N)}$ is larger in the positive semidefinite ordering than that of the rank-aware MLE $\hat{\mathbf{B}}^{(M)}$.

1.2. Central limit theorem for $\hat{\mathbf{B}}^{(M)}$

We now state a central limit theorem for the entries of $\hat{\mathbf{B}}^{(M)}$. We first recall the notion of the duplication, elimination, and commutation matrices [39,40]. Let $\mathbf{M}$ be a $n \times m$ matrix. The *vectorization* of $\mathbf{M}$, denoted as $\text{vec}(\mathbf{M})$, is a column vector in $\mathbb{R}^{nm}$ obtained by stacking the columns of $\mathbf{M}$. When $\mathbf{M}$ is symmetric then the *half vectorization* of $\mathbf{M}$, denoted as $\text{vech}(\mathbf{M})$, is the column vector obtained by vectorizing only the lower triangular entries of $\mathbf{M}$. For integer $n \geq 1$, the duplication matrix $\mathcal{D}_n$ is the unique $n^2 \times n(n+1)/2$ matrix such that, for any $n \times n$ symmetric matrix $\mathbf{M}$

$$\text{vec}(\mathbf{M}) = \mathcal{D}_n \text{vech}(\mathbf{M}). \tag{1.9}$$

Similarly, the elimination matrix $\mathcal{L}_n$ is the unique $n(n+1)/2 \times n^2$ matrix such that

$$\text{vech}(\mathbf{M}) = \mathcal{L}_n \text{vec}(\mathbf{M}) \tag{1.10}$$

for any $n \times n$ symmetric matrix $\mathbf{M}$. Finally, for any integers $n, m \geq 1$, the commutation matrix $\mathcal{T}_{mn}$ is the unique $mn \times mn$ matrix such that, for any $n \times m$ matrix $\mathbf{M}$,

$$\text{vec}(\mathbf{A}^\top) = \mathcal{T}_{mn} \text{vec}(\mathbf{A}). \tag{1.11}$$

The commutation matrix swaps the ordering of matrices in a Kronecker product. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be matrices of dimensions $p \times q$ and $m \times n$, respectively. Then

$$\mathcal{T}_{pm}(\mathbf{M}_1 \otimes \mathbf{M}_2) = (\mathbf{M}_2 \otimes \mathbf{M}_1)\mathcal{T}_{qn}. \tag{1.12}$$

Theorem 2. *Assume the setting in Theorem 1. Let $d = \text{rk}(\mathbf{B})$ and let $\mathbf{B}_{11}$ denote the top-left $d \times d$ block of $\mathbf{B}$. Assume, without loss of generality, that $\mathbf{B}_{11}$ is invertible. Also let $\mathbf{B}_{12} = \mathbf{B}_{21}^\top$ denote the $d \times (K-d)$ top right block of $\mathbf{B}$. Let c_1, c_2 and c_3 be defined as*

$$c_1 = \binom{d+1}{2}, \quad c_2 = d(K-d), \quad c_3 = \binom{K+1}{2}.$$

Let $\mathcal{J}$ be the $K(K+1)/2 \times (Kd - d(d-1)/2)$ matrix with blocks structure of the form

$$\mathcal{J} = \begin{bmatrix} \mathbf{I}_{c_1} & \mathbf{0}_{c_1 \times c_2} \\ (\mathbf{B}_{21}\mathbf{B}_{11}^{-1} \otimes \mathbf{I}_d)\mathcal{D}_d & \mathbf{I}_{K-d} \otimes \mathbf{B}_{11} \\ \mathcal{L}_{K-d}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1} \otimes \mathbf{B}_{21}\mathbf{B}_{11}^{-1})\mathcal{D}_d & \mathcal{L}_{K-d}(\mathbf{I}_{(K-d)^2} + \mathcal{T}_{(K-d)^2})(\mathbf{I}_{K-d} \otimes \mathbf{B}_{21}) \end{bmatrix} \tag{1.13}$$

Let $\mathbf{D}$ be the $c_3 \times c_3$ diagonal matrix with entries indexed by $1 \leq k \leq \ell \leq K$ such that

$$\mathbf{D}_{(k,\ell),(k,\ell)} = \frac{\pi_k \pi_\ell}{\mathbf{B}_{k\ell}(1 - \mathbf{B}_{k\ell})(1 + \delta_{k\ell})} \quad \textit{when } \rho_n \equiv 1,$$

$$\mathbf{D}_{(k,\ell),(k,\ell)} = \frac{\pi_k \pi_\ell}{\mathbf{B}_{k\ell}(1 + \delta_{k\ell})}, \quad \textit{when } \rho_n \to 0.$$

Here $\delta_{k\ell} = 1$ if $k = \ell$ and $\delta_{k\ell} = 0$ otherwise. We then have

$$n\rho_n^{1/2}(\text{vech}(\hat{\mathbf{B}}^{(M)}) - \text{vech}(\mathbf{B})) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{J}(\mathcal{J}^\top \mathbf{D} \mathcal{J})^{-1} \mathcal{J}^\top)$$

as $n \to \infty$.

We emphasize that the diagonal entries of $\mathbf{D}^{-1}$ in the statement of the above theorem are the variances of the $\hat{\mathbf{B}}^{(N)}_{k\ell}$ as given in Theorem 1. Furthermore, with $\mathbf{Z} = \mathbf{D}^{1/2}\mathcal{J}$,

$$\mathcal{J}(\mathcal{J}^\top \mathbf{D}\mathcal{J})^{-1}\mathcal{J}^\top = \mathbf{D}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{D}^{-1/2} \preceq \mathbf{D}^{-1}.$$

The covariance matrix of $\hat{\mathbf{B}}^{(M)}$ is thus smaller, with respect to the positive semidefinite ordering, than that of $\hat{\mathbf{B}}^{(N)}$, with equality if and only if $\mathbf{B}$ is of full rank.

1.3. Low-rank approximation for $\hat{\mathbf{B}}^{(N)}$

Another possible estimator for $\mathbf{B}$ is obtained by considering, when appropriate, the low-rank approximation of the naive estimator $\hat{\mathbf{B}}^{(N)}$. More specifically, given the estimator $\hat{\mathbf{B}}^{(N)}$ described in Section 1.1 and suppose that the rank of $\mathbf{B}$ is known. If d is unknown then there exists an estimate $\hat{d}$ satisfying $\hat{d} = d$ almost surely as $n \to \infty$; see the discussion in Section 1.4 for more details. Letting $d = \mathrm{rk}(\mathbf{B})$, we now consider the best rank-d approximation of $\hat{\mathbf{B}}^{(N)}$ in Frobenius norm, i.e.,

$$\hat{\mathbf{B}}^{(N,d)} = \operatorname*{argmin}_{\mathbf{M}:\ \mathrm{rk}(\mathbf{M})=d} \|\mathbf{M} - \hat{\mathbf{B}}^{(N)}\|_F.$$

The estimator $\hat{\mathbf{B}}^{(N,d)}$ is obtained by truncating the singular value decomposition of $\hat{\mathbf{B}}^{(N)}$ and keeping the d largest singular values and the corresponding left and right singular vectors. As $\mathbf{B}$ is rank d, Theorem 1 in [56] yields the following perturbation expansion

$$\hat{\mathbf{B}}^{(N,d)} = \mathbf{B} + (\hat{\mathbf{B}}^{(N)} - \mathbf{B}) - (\mathbf{I} - \mathbf{V}\mathbf{V}^\top)(\hat{\mathbf{B}}^{(N)} - \mathbf{B})(\mathbf{I} - \mathbf{V}\mathbf{V}^\top) + \mathcal{O}(\|\hat{\mathbf{B}}^{(N)} - \mathbf{B}\|_F^2).$$

Here $\mathbf{V}\mathbf{V}^\top$ is the orthogonal projection onto the column space of $\mathbf{B}$. Let $\mathbf{\Pi}_{\mathbf{V}}^{\perp} = \mathbf{I} - \mathbf{V}\mathbf{V}^\top$. Then by Theorem 1, the term $\mathcal{O}(\|\hat{\mathbf{B}}^{(N)} - \mathbf{B}\|_F^2)$ is negligible in the limit, i.e.,

$$\begin{aligned} n\rho_n^{1/2}\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N,d)} - \mathbf{B}\Big) &= n\rho_n^{1/2}\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N)} - \mathbf{B}\Big) \\ &\quad - n\rho_n^{1/2}\Big(\mathcal{L}_K\big(\mathbf{\Pi}_{\mathbf{V}}^{\perp} \otimes \mathbf{\Pi}_{\mathbf{V}}^{\perp}\big)\mathcal{D}_K\Big)\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N)} - \mathbf{B}\Big) + o_{\mathbb{P}}(1). \end{aligned} \tag{1.14}$$

Now from Lemma 3.5 in [40], we have $\mathcal{L}_K\mathcal{D}_K = \mathbf{I}_{K(K+1)/2}$ and hence

$$n\rho_n^{1/2}\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N,d)} - \mathbf{B}\Big) = n\rho_n^{1/2}\Big(\mathcal{L}_K\big(\mathbf{I} - \mathbf{\Pi}_{\mathbf{V}}^{\perp} \otimes \mathbf{\Pi}_{\mathbf{V}}^{\perp}\big)\mathcal{D}_K\Big)\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N)} - \mathbf{B}\Big) + o_{\mathbb{P}}(1).$$

Another application of Theorem 1 yields the following limiting distribution for $\hat{\mathbf{B}}^{(N,d)}$. We emphasize that while Corollary 1 is a straightforward result to derive, it is, to the best of our knowledge, a new result that had not appeared previously in the literature.

Corollary 1. *Assume the setting in Theorem 1. Suppose $d = \mathrm{rk}(\mathbf{B})$ is known. Let $\mathbf{\Pi}_{\mathbf{V}}^{\perp} = \mathbf{I} - \mathbf{V}\mathbf{V}^\top$ where $\mathbf{V}$ is the $K \times d$ orthonormal matrix whose columns are the eigenvectors corresponding to the non-zero eigenvalues of $\mathbf{B}$. Then the truncated low-rank estimator $\hat{\mathbf{B}}^{(N,d)}$ satisfies*

$$n\sqrt{\rho_n}\mathrm{vech}\Big(\hat{\mathbf{B}}^{(N,d)} - \mathbf{B}\Big) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(\mathbf{0}, \mathcal{L}_K\big(\mathbf{I} - \mathbf{\Pi}_{\mathbf{V}}^{\perp} \otimes \mathbf{\Pi}_{\mathbf{V}}^{\perp}\big)\mathcal{D}_K\, \mathbf{D}^{-1}\, \mathcal{D}_K^\top\big(\mathbf{I} - \mathbf{\Pi}_{\mathbf{V}}^{\perp} \otimes \mathbf{\Pi}_{\mathbf{V}}^{\perp}\big)\mathcal{L}_K^\top\Big) \tag{1.15}$$

Here $\mathbf{D}$ is the $\binom{K+1}{2} \times \binom{K+1}{2}$ diagonal matrix defined in Theorem 2, i.e., the diagonal entries of $\mathbf{D}^{-1}$ are the limiting variances for $\hat{\mathbf{B}}^{(N)}$.

We note that if $\mathbf{B}$ is full-rank then $\mathbf{\Pi}_{\mathbf{V}}^{\perp} = \mathbf{0}$ and the limiting distribution of the truncated $\hat{\mathbf{B}}^{(N,d)}$ is identical to that of the naive MLE $\hat{\mathbf{B}}^{(N)}$. Furthermore, from Eq. (1.14), the entries of $\hat{\mathbf{B}}^{(N,d)}$ can be written as linear combinations of the entries of $\hat{\mathbf{B}}^{(N)}$ and that the coefficients for these linear combinations depend only on the eigenvectors of $\mathbf{B}$ and not on the block assignment probabilities $\boldsymbol{\pi}$. Meanwhile, the variances of $\hat{\mathbf{B}}^{(N)}$ do depend on $\boldsymbol{\pi}$. Now suppose $\mathbf{B}$ is such that $\mathrm{rk}(\mathbf{B}) = d < K$. Then there exists a choice of $\boldsymbol{\pi}$ such that for some pair (k, ℓ), we have $\mathrm{Var}[\hat{\mathbf{B}}_{k\ell}^{(N,d)}] > \mathrm{Var}[\hat{\mathbf{B}}_{k\ell}^{(N)}]$. The covariance matrix for $\hat{\mathbf{B}}^{(N,d)}$ when $d < K$ is therefore not guaranteed to be smaller, with respect to the ordering of positive semidefinite matrices, than the covariance matrix for $\hat{\mathbf{B}}^{(N)}$.

1.4. Spectral embedding estimate for B

Another alternative to $\hat{\mathbf{B}}^{(N)}$ is based on spectral embedding of $\mathbf{A}$. More specifically, given $\mathbf{A} \sim \mathrm{SBM}(\mathbf{B}, \boldsymbol{\pi})$ with known sparsity factor ρ, we consider the following procedure for estimating $\mathbf{B}$:

1. Assuming $d = \mathrm{rk}(\mathbf{B})$ is known, let $\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^{\top} + \hat{\mathbf{U}}_{\perp}\hat{\mathbf{\Lambda}}_{\perp}\hat{\mathbf{U}}_{\perp}^{\top}$ be the eigendecomposition of $\mathbf{A}$ where $\hat{\mathbf{\Lambda}}$ is the diagonal matrix containing the d largest eigenvalues of $\mathbf{A}$ in modulus and $\hat{\mathbf{U}}$ is the $n \times d$ matrix whose columns are the corresponding eigenvectors of $\mathbf{A}$.
2. Assuming K is known, cluster the rows of $\hat{\mathbf{U}}$ into K clusters using K-means, obtaining an "estimate" $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$.
3. For $k \in [K]$, let $\hat{s}_k$ be the vector in $\mathbb{R}^n$ where the i-th entry of $\hat{s}_k$ is 1 if $\hat{\tau}_i = k$ and 0 otherwise and let $\hat{n}_k = |\{i : \hat{\tau}_i = k\}|$ be the number of vertices in the kth cluster.
4. Estimate $\mathbf{B}_{k\ell}$ by $\hat{\mathbf{B}}_{k\ell}^{(S)} = \frac{1}{\hat{n}_k \hat{n}_\ell \rho} \hat{s}_k^{\top} \hat{\mathbf{U}} \hat{\mathbf{\Lambda}} \hat{\mathbf{U}}^{\top} \hat{s}_\ell$.

The above procedure assumes d and K are known. When d is unknown, it can be consistently estimated using the following approach: let $\hat{d}$ be the number of eigenvalues of $\mathbf{A}$ exceeding $4\sqrt{\delta(\mathbf{A})}$ in modulus; here $\delta(\mathbf{A})$ denotes the max degree of $\mathbf{A}$. Then $\hat{d}$ is a consistent estimate of d as $n \to \infty$. This follows directly from tail bounds for $\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|$ (see e.g., [35,44]) and Weyl's inequality. The estimation of K is considered in [10,34,57] among others. The results in [10,34] are applicable for dense graphs where $\rho_n \equiv 1$ and are based on showing that if $\mathbf{A}$ is a K-block SBM, then there exists a consistent estimate $(\hat{\mathbf{A}}_{ij})$ of $\mathbb{E}[\mathbf{A}]$ such that the matrix $\widetilde{\mathbf{A}}$ with entries $\widetilde{\mathbf{A}}_{ij} = (\mathbf{A}_{ij} - \hat{\mathbf{A}}_{ij})/\sqrt{(n-1)\hat{\mathbf{A}}_{ij}(1 - \hat{\mathbf{A}}_{ij})}$ has a limiting Tracy-Widom distribution, i.e., $n^{2/3}(\lambda_1(\widetilde{\mathbf{A}}) - 2)$ converges to Tracy-Widom. The results in [57] are applicable for sparse graphs where $\rho_n \to 0$; see in particular Corollary 2.8 of [57]. However since the estimation of K generally requires fitting SBM graphs with K' blocks for various candidate choices of K', for ease of exposition we shall assume throughout this paper that K is known so that we only need to estimate d when it is unknown.

The above spectral embedding procedure (and related procedures based on eigendecomposition of other matrices such as the normalized Laplacian) is also well-studied, see e.g., [5,15,22,30,35,37,42,46,49,51,53] among others. The available results, however, focused only on showing that $\hat{\boldsymbol{\tau}}$ consistently recovers $\boldsymbol{\tau}$. Since $\hat{\boldsymbol{\tau}}$ is almost surely an exact recovery of $\boldsymbol{\tau}$ in the limit, i.e., $\hat{\tau}_i = \tau_i$ for all i as $n \to \infty$ (see e.g., [37,42]), the estimate of $\mathbf{B}_{k\ell}$ given by $\frac{1}{\hat{n}_k \hat{n}_\ell \rho} \hat{s}_k^{\top} \mathbf{A} \hat{s}_\ell$ is a consistent estimate of $\mathbf{B}$, and furthermore, coincides with $\hat{\mathbf{B}}^{(N)}$ as $n \to \infty$.

The quantity $\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^{\top} - \mathbb{E}[\mathbf{A}]$ is also widely analyzed in the context of matrix and graphon estimation using universal singular values thresholding [13,25,32,59]; the focus there had been in showing the minimax rates of convergence of $\frac{1}{n^2}\|\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^{\top} - \mathbb{E}[\mathbf{A}]\|_F^2$ to 0 as $n \to \infty$. In particular, in the context of stochastic blockmodel graphs with sparsity parameter ρ and letting $\mathbf{P} = \mathbb{E}[\mathbf{A}]$ be the matrix of edge

probablities and $\hat{\mathbf{P}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top$ be the rank-d approximation of $\mathbf{A}$, we have

$$\|\hat{\mathbf{P}} - \mathbf{P}\|_F = O_{\mathbb{P}}(\sqrt{n\rho}), \quad \max_{ij} |\hat{\mathbf{P}}_{ij} - \mathbf{P}_{ij}| = O_{\mathbb{P}}(\rho^{1/2} n^{-1/2}).$$

See e.g., Theorem 1 in [59] and Theorem 5 in [48] for these results. In comparison,

$$\|\mathbf{A} - \mathbf{P}\|_F = \Omega_{\mathbb{P}}(n\rho), \quad \max_{ij} |\mathbf{A}_{ij} - \mathbf{P}_{ij}| = \Omega_{\mathbb{P}}(1).$$

The low-rank approximation $\hat{\mathbf{P}}$ of $\mathbf{A}$ is therefore a much better estimate of $\mathbf{P}$, in both Frobenius norm and uniform entrywise norm, than $\mathbf{A}$ itself. These different rates of convergence, however, do not necessary translate to different rate of convergence for $\hat{\mathbf{B}}^{(S)}_{k\ell} - \mathbf{B}_{k\ell}$ compared to that of $\hat{\mathbf{B}}^{(N)}_{k\ell} - \mathbf{B}_{k\ell}$. Indeed, both $\hat{\mathbf{B}}^{(S)}_{k\ell}$ and $\hat{\mathbf{B}}^{(N)}_{k\ell}$ require averaging over a subset of the entries of $\hat{\mathbf{P}}$ and $\mathbf{A}$, respectively. Note that the entries of $\mathbf{A}$ are independent but the entries of $\hat{\mathbf{P}}$ are *not* independent, and hence the effects of averaging will be quite different for $\hat{\mathbf{P}}$ compared to $\mathbf{A}$. See Corollary 3 and its accompanying discussion.

In summary, formal comparisons between $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(S)}$ are severely lacking. Our paper addresses this important void in the literature. The contributions of our paper are as follows. For stochastic blockmodel graphs with sparisty factors ρ_n satisfying $n\rho_n = \omega(\sqrt{n})$, we establish asymptotic normality of $n\rho_n^{1/2}(\hat{\mathbf{B}}^{(S)} - \mathbf{B})$ in Theorem 3 and Theorem 4. As a corollary of this result, we show that when $\mathbf{B}$ is of full-rank, $n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(S)} - \mathbf{B})$ has the same limiting distribution as $n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(N)} - \mathbf{B})$ given in Eq. (1.5) and Eq. (1.6) and that both estimators are asymptotically efficient; the two estimators $\hat{\mathbf{B}}^{(M)}$ and $\hat{\mathbf{B}}^{(N)}$ are identical in this setting. When $\mathbf{B}$ is singular, we show that $n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(S)} - \mathbf{B})$ can have smaller variances than $n\sqrt{\rho_n}(\hat{\mathbf{B}}^{(N)} - \mathbf{B})$, and thus a bias-corrected $\hat{\mathbf{B}}^{(S)}$ can have smaller mean square error than $\hat{\mathbf{B}}^{(N)}$. Furthermore, the bias-corrected $\hat{\mathbf{B}}^{(S)}$ can be almost as efficient as the asymptotically first-order efficient estimator $\hat{\mathbf{B}}^{(M)}$. Finally, we also provide some justification of the potential necessity of the condition that the average degree satisfies $n\rho_n = \omega(\sqrt{n})$; in essence, as $\rho_n \to 0$, the bias incurred by the low-rank representation $\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top$ of $\mathbf{A}$ overwhelms the reduction in variance resulting from the low-rank representation.

2. Central limit theorem for $\hat{\mathbf{B}}^{(S)}_{k\ell}$

Let $(\mathbf{A}, \boldsymbol{\tau}) \sim \mathrm{SBM}(\mathbf{B}, \boldsymbol{\pi}, \rho)$ be a stochastic blockmodel graph on n vertices with sparsity factor ρ and $\mathrm{rk}(\mathbf{B}) = d$. We first consider the setting wherein both $\boldsymbol{\tau}$ and d are assumed known. The setting wherein $\boldsymbol{\tau}$ is unknown and needs to be recovered will be addressed subsequently in Corollary 4, while the setting when d is unknown was previously addressed following the introduction of the estimator $\hat{\mathbf{B}}^{(S)}$ in Section 1. When $\boldsymbol{\tau}$ is known, the spectral embedding estimate of $\mathbf{B}_{k\ell}$ (with ρ assumed known) is $\hat{\mathbf{B}}^{(S)}_{k\ell} = \frac{1}{\rho n_k n_\ell} \boldsymbol{s}_k^\top \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top \boldsymbol{s}_l$ where $\boldsymbol{s}_k$ is the vector whose elements $\{s_{ki}\}$ are such that $s_{ki} = 1$ if vertex i is assigned to block k and $s_{ki} = 0$ otherwise; here n_k denote the number of vertices v_i assigned to block k.

We then have the following non-degenerate limiting distribution of $\hat{\mathbf{B}}^{(S)} - \mathbf{B}$. We shall present two variants of this limiting distribution. The first variant, Theorem 3, applies to the setting where the average degree grows linearly with n, i.e., the sparsity factor $\rho_n \to c > 0$; without loss of generality, we can assume $\rho_n \equiv c = 1$. The second variant applies to the setting where the average degree grows sub-linearly in n, i.e., $\rho_n \to 0$ with $n\rho_n = \omega(\sqrt{n})$. For ease of exposition, these variants will be presented using the following parametrization of stochastic blockmodel graphs as a sub-class of the more general random dot product graphs model [48,60].

Definition 2 (Generalized random dot product graph). Let d be a positive integer and $p \geq 1$ and $q \geq 0$ be such that $p + q = d$. Let $\mathbf{I}_{p,q}$ denote the diagonal matrix whose diagonal elements contains p entries equaling 1 and q entries equaling -1. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ such $x^\top \mathbf{I}_{p,q} y \in [0, 1]$ for all $x, y \in \mathcal{X}$. Let F be a distribution taking values in $\mathcal{X}$. We say $(\mathbf{X}, \mathbf{A}) \sim \mathrm{GRDPG}_{p,q}(F)$ with sparsity factor $\rho \in (0, 1]$ if the following holds. First let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d}}{\sim} F$ and set $\mathbf{X} = [X_1 \mid \cdots \mid X_n]^\top \in \mathbb{R}^{n \times d}$. Then $\mathbf{A} \in \{0, 1\}^{n \times n}$ is a symmetric matrix such that, conditioned on $\mathbf{X}$, for all $i \geq j$ the A_{ij} are independent and

$$A_{ij} \sim \mathrm{Bernoulli}(\rho X_i^\top \mathbf{I}_{p,q} X_j). \tag{2.1}$$

We therefore have

$$\mathbb{P}[\mathbf{A} \mid \mathbf{X}] = \prod_{i \leq j} (\rho X_i^\top \mathbf{I}_{p,q} X_j)^{A_{ij}} (1 - \rho X_i^\top \mathbf{I}_{p,q} X_j)^{(1 - A_{ij})}. \tag{2.2}$$

It is straightforward to show that any stochastic blockmodel graph $(\mathbf{A}, \boldsymbol{\tau}) \sim \mathrm{SBM}(\boldsymbol{\pi}, \mathbf{B})$ is also a generalized random dot product graph $(\mathbf{X}, \mathbf{A}) \sim \mathrm{GRDPG}_{p,q}(F)$ where F is a mixture of point masses. Indeed, suppose $\mathbf{B}$ is a $K \times K$ matrix and let $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$ be the eigendecomposition of $\mathbf{B}$. Then, denoting by $\nu_1, \nu_2, \ldots, \nu_K$ the rows of $\mathbf{U}|\boldsymbol{\Sigma}|^{1/2}$, we can define $F = \sum_{k=1}^K \pi_k \delta_{\nu_k}$ where δ_{ν_k} is the Dirac delta function at $\nu_k \in \mathbb{R}^d$. The values for p and q are the number of positive and negative eigenvalues of $\mathbf{B}$, respectively.

Theorem 3. *Let $\mathbf{A} \sim \mathrm{SBM}(\boldsymbol{\pi}, \mathbf{B})$ be a K-block stochastic blockmodel graph on n vertices with sparsity factor $\rho_n \equiv 1$. Let $\nu_1, \ldots, \nu_K$ be point masses in $\mathbb{R}^d$ such that $\mathbf{B}_{k\ell} = \nu_k^\top \mathbf{I}_{p,q} \nu_\ell$ and let $\Delta = \sum_k \pi_k \nu_k \nu_k^\top$. For $k \in [K]$ and $\ell \in [K]$, let $\theta_{k\ell}$ be given by*

$$\begin{aligned} \theta_{k\ell} = &\sum_{r=1}^K \pi_r \big(\mathbf{B}_{kr}(1 - \mathbf{B}_{kr}) + \mathbf{B}_{\ell r}(1 - \mathbf{B}_{\ell r})\big) \nu_k^\top \Delta^{-1} \mathbf{I}_{p,q} \Delta^{-1} \nu_\ell \\ &- \sum_{r=1}^K \sum_{s=1}^K \pi_r \pi_s \mathbf{B}_{sr}(1 - \mathbf{B}_{sr}) \nu_s^\top \Delta^{-1} \mathbf{I}_{p,q} \Delta^{-1} (\nu_\ell \nu_k^\top + \nu_k \nu_\ell^\top) \Delta^{-1} \nu_s. \end{aligned} \tag{2.3}$$

Now let $\zeta_{k\ell} = \nu_k^\top \Delta^{-1} \nu_\ell$. Define σ_{kk}^2 for $k \in [K]$ to be

$$\begin{aligned} \sigma_{kk}^2 = &4\mathbf{B}_{kk}(1 - \mathbf{B}_{kk}) \zeta_{kk}^2 + 4 \sum_r \pi_r \mathbf{B}_{kr}(1 - \mathbf{B}_{kr}) \zeta_{kr}^2 \big(\tfrac{1}{\pi_k} - 2\zeta_{kk}\big) \\ &+ 2 \sum_r \sum_s \pi_r \pi_s \mathbf{B}_{rs}(1 - \mathbf{B}_{rs}) \zeta_{kr}^2 \zeta_{ks}^2 \end{aligned} \tag{2.4}$$

and define $\sigma_{k\ell}^2$ *for* $1 \le k < \ell \le K$ *to be*

$$\begin{aligned}\sigma_{k\ell}^2 &= \big(\mathbf{B}_{kk}(1-\mathbf{B}_{kk}) + \mathbf{B}_{\ell\ell}(1-\mathbf{B}_{\ell\ell})\big)\zeta_{k\ell}^2 + 2\mathbf{B}_{k\ell}(1-\mathbf{B}_{k\ell})\zeta_{kk}\zeta_{\ell\ell} \\ &\quad + \sum_r \pi_r \mathbf{B}_{kr}(1-\mathbf{B}_{kr})\zeta_{\ell r}^2\big(\tfrac{1}{\pi_k} - 2\zeta_{kk}\big) \\ &\quad + \sum_r \pi_r \mathbf{B}_{\ell r}(1-\mathbf{B}_{\ell r})\zeta_{kr}^2\big(\tfrac{1}{\pi_\ell} - 2\zeta_{\ell\ell}\big) \\ &\quad - 2\sum_r \pi_r \big(\mathbf{B}_{kr}(1-\mathbf{B}_{kr}) + \mathbf{B}_{\ell r}(1-\mathbf{B}_{\ell r})\big)\zeta_{kr}\zeta_{r\ell}\zeta_{k\ell} \\ &\quad + \frac{1}{2}\sum_r\sum_s \pi_r\pi_s \mathbf{B}_{rs}(1-\mathbf{B}_{rs})(\zeta_{kr}\zeta_{\ell s} + \zeta_{\ell r}\zeta_{ks})^2.\end{aligned} \tag{2.5}$$

Then for any $k \in [K]$ *and* $\ell \in [K]$,

$$n\big(\hat{\mathbf{B}}_{k\ell}^{(S)} - \mathbf{B}_{k\ell} - \frac{\theta_{k\ell}}{n}\big) \xrightarrow{\mathrm{d}} N(0, \sigma_{k\ell}^2) \tag{2.6}$$

as $n \to \infty$.

Theorem 4. *Let* $\mathbf{A} \sim \mathrm{SBM}(\boldsymbol{\pi}, \mathbf{B}, \rho_n)$ *be a* K*-block stochastic blockmodel graph on* n *vertices with sparsity factor* ρ_n. *Let* $\nu_1, \ldots, \nu_K$ *be point masses in* $\mathbb{R}^d$ *such that* $\mathbf{B}_{k\ell} = \nu_k^\top \mathbf{I}_{p,q} \nu_\ell$ *and let* $\Delta = \sum_k \pi_k \nu_k \nu_k^\top$. *For* $k \in [K]$ *and* $\ell \in [K]$, *let* $\widetilde{\theta}_{k\ell}$ *be given by*

$$\begin{aligned}\widetilde{\theta}_{k\ell} &= \sum_{r=1}^K \pi_r \big(\mathbf{B}_{kr} + \mathbf{B}_{\ell r}\big)\nu_k^\top \Delta^{-1}\mathbf{I}_{p,q}\Delta^{-1}\nu_\ell \\ &\quad - \sum_{r=1}^K\sum_{s=1}^K \pi_r\pi_s \mathbf{B}_{sr}\nu_s^\top \Delta^{-1}\mathbf{I}_{p,q}\Delta^{-1}(\nu_\ell\nu_k^\top + \nu_k\nu_\ell^\top)\Delta^{-1}\nu_s,\end{aligned} \tag{2.7}$$

let $\widetilde{\sigma}_{kk}^2$ *for* $k \in [K]$ *be*

$$\begin{aligned}\widetilde{\sigma}_{kk}^2 &= 4\mathbf{B}_{kk}\zeta_{kk}^2 + 4\sum_r \pi_r \mathbf{B}_{kr}\zeta_{kr}^2\big(\tfrac{1}{\pi_k} - 2\zeta_{kk}\big)^2 \\ &\quad + 2\sum_r\sum_s \pi_r\pi_s \mathbf{B}_{rs}\zeta_{kr}^2\zeta_{ks}^2,\end{aligned} \tag{2.8}$$

and let $\widetilde{\sigma}_{k\ell}^2$ *for* $k \in [K]$, $\ell \in [K]$, $k \ne \ell$ *be*

$$\begin{aligned}\widetilde{\sigma}_{k\ell}^2 &= \big(\mathbf{B}_{kk} + \mathbf{B}_{\ell\ell}\big)\zeta_{k\ell}^2 + 2\mathbf{B}_{k\ell}\zeta_{kk}\zeta_{\ell\ell} - 2\sum_r \pi_r\big(\mathbf{B}_{kr} + \mathbf{B}_{\ell r}\big)\zeta_{kr}\zeta_{r\ell}\zeta_{k\ell} \\ &\quad + \sum_r \pi_r \mathbf{B}_{kr}\zeta_{\ell r}^2\big(\tfrac{1}{\pi_k} - 2\zeta_{kk}\big)^2 + \sum_r \pi_r \mathbf{B}_{\ell r}\zeta_{kr}^2\big(\tfrac{1}{\pi_\ell} - 2\zeta_{\ell\ell}\big)^2 \\ &\quad + \frac{1}{2}\sum_r\sum_s \pi_r\pi_s \mathbf{B}_{rs}(\zeta_{kr}\zeta_{\ell s} + \zeta_{\ell r}\zeta_{ks})^2.\end{aligned} \tag{2.9}$$

If $\rho_n \to 0$ *and* $n\rho_n = \omega(\sqrt{n})$, *then for any* $k \in [K]$ *and* $\ell \in [K]$,

$$n\rho_n^{1/2}\big(\hat{\mathbf{B}}^{(S)}_{k\ell} - \mathbf{B}_{k\ell} - \frac{\widetilde{\theta}_{k\ell}}{n\rho_n}\big) \xrightarrow{\mathrm{d}} N(0, \widetilde{\sigma}^2_{k\ell}) \tag{2.10}$$

as $n \to \infty$.

Remark 1. For ease of exposition, we only presented expressions for the variances $\sigma^2_{k\ell}$ and $\widetilde{\sigma}^{(2)}_{k\ell}$ in Theorem 3 and Theorem 4. Expressions for the covariances $\sigma_{k\ell,k'\ell'}$ or $\widetilde{\sigma}_{k\ell,k'\ell'}$ between $\hat{\mathbf{B}}^{(S)}_{k\ell}$ and $\hat{\mathbf{B}}^{(S)}_{k'\ell'}$ are given in Section C of the supplementary material [52]. These expressions, together with the Cramer-Wold device, yield the convergence of $n\rho_n^{1/2}\mathrm{vech}\Big(\hat{\mathbf{B}}^{(S)} - \mathbf{B}^{(S)} - \frac{\boldsymbol{\theta}}{n\rho_n}\Big)$ to a multivariate normal distribution. Finally we emphasize that the limiting distribution of $\hat{\mathbf{B}}^{(S)}$ also depends on the asymptotic biases $\{\theta_{k\ell}\}$ which are unknown and need to be estimated; consistent estimators $\hat{\theta}_{k\ell}$, with $\rho_n^{-1/2}(\hat{\theta}_{k\ell} - \theta_{k\ell}) \longrightarrow 0$, are given in Corollary 4.

The proofs of Theorem 3 and Theorem 4 are given in Section B of the supplementary material [52]. These two results differ mainly due to how the quantities $\theta_{k\ell}$ and $\sigma_{k\ell}$ are defined in Theorem 3 versus how the quantities $\widetilde{\theta}_{k\ell}$ and $\widetilde{\sigma}_{k\ell}$ are defined in Theorem 4. As a corollary of Theorem 3 and Theorem 4, we have the following result for the asymptotic efficiency of $\hat{\mathbf{B}}^{(S)}$ whenever $\mathbf{B}$ is invertible (see also Theorem 1). We emphasize that a much simpler and more direct proof of this result is also provided in Remark B.1 of the supplementary material [52]. In essence, the technical challenges and complexity in Theorem 3 and Theorem 4 arise only when $\mathbf{B}$ is singular.

Corollary 2. *Let* $\mathbf{A} \sim \mathrm{SBM}(\boldsymbol{\pi}, \mathbf{B}, \rho_n)$ *be a* K*-block stochastic blockmodel graph on* n *vertices with sparsity factor* ρ_n. *Suppose* $\mathbf{B}$ *is invertible. Then* $\theta_{k\ell} = \widetilde{\theta}_{k\ell} = 0$ *for all* $k, \ell \in [K]$. *Furthermore,* $\sigma_{k\ell}$ *and* $\widetilde{\sigma}_{k\ell}$ *as defined in Theorem 3 and Theorem 4 satisfy*

$$\sigma^2_{kk} = \frac{2\mathbf{B}_{kk}(1 - \mathbf{B}_{kk})}{\pi_k^2}; \quad \sigma^2_{k\ell} = \frac{\mathbf{B}_{k\ell}(1 - \mathbf{B}_{k\ell})}{\pi_k \pi_\ell} \text{ if } k \neq \ell, \tag{2.11}$$

$$\widetilde{\sigma}^2_{kk} = \frac{2\mathbf{B}_{kk}}{\pi_k^2}; \quad \widetilde{\sigma}^2_{k\ell} = \frac{\mathbf{B}_{k\ell}}{\pi_k \pi_\ell} \text{ if } k \neq \ell. \tag{2.12}$$

Therefore, for all $k \in [K]$, $\ell \in [K]$, *if* $\rho_n \equiv 1$, *then*

$$n(\hat{\mathbf{B}}^{(S)}_{k\ell} - \mathbf{B}_{k\ell}) \xrightarrow{\mathrm{d}} N(0, \sigma^2_{k\ell}). \tag{2.13}$$

as $n \to \infty$. *If* $\rho_n \to 0$ *and* $n\rho_n = \omega(\sqrt{n})$, *then*

$$n\rho_n^{1/2}(\hat{\mathbf{B}}^{(S)}_{k\ell} - \mathbf{B}_{k\ell}) \xrightarrow{\mathrm{d}} N(0, \widetilde{\sigma}^2_{k\ell}). \tag{2.14}$$

as $n \to \infty$. $\hat{\mathbf{B}}^{(S)}$ *is thus asymptotically efficient.*

Proof of Corollary 2. The proof follows from the observation that $\zeta_{rs} = \frac{1}{\pi_r}$ for $r = s$ and $\zeta_{rs} = 0$ otherwise. Indeed, $\Delta = \sum_k \pi_k \nu_k \nu_k^\top = \boldsymbol{\nu}^\top \mathrm{diag}(\boldsymbol{\pi})\boldsymbol{\nu}$ where $\boldsymbol{\nu}$ is a $K \times K$ *invertible* matrix with $\boldsymbol{\nu}\mathbf{I}_{p,q}\boldsymbol{\nu}^\top = \mathbf{B}$ and $\mathrm{diag}(\boldsymbol{\pi})$ is the diagonal matrix with diagonal entries $\pi_1, \ldots, \pi_K$. Hence

$$\zeta_{rs} = \nu_r^\top \Delta^{-1} \nu_s = \nu_r^\top \boldsymbol{\nu}^{-1} \mathrm{diag}(\boldsymbol{\pi})^{-1} (\boldsymbol{\nu}^{-1})^\top \nu_s = \frac{1}{\pi_r} \mathbb{1}\{r = s\}. \tag{2.15}$$

As an example, the expression for $\theta_{k\ell}$ in Eq. (2.3) reduces to

$$\begin{aligned}\theta_{k\ell} &= \sum_{r=1}^{K} \pi_r \big(\mathbf{B}_{kr}(1-\mathbf{B}_{kr}) + \mathbf{B}_{\ell r}(1-\mathbf{B}_{\ell r})\big) \nu_k^\top \Delta^{-1} \mathbf{I}_{p,q} \Delta^{-1} \nu_\ell \\ &\quad - \sum_{r=1}^{K}\sum_{s=1}^{K} \pi_r \pi_s \mathbf{B}_{sr}(1-\mathbf{B}_{sr}) \nu_s^\top \Delta^{-1} \mathbf{I}_{p,q} \Delta^{-1} \Big(\tfrac{\mathbb{1}\{s=k\}}{\pi_k} \nu_\ell + \tfrac{\mathbb{1}\{s=\ell\}}{\pi_\ell} \nu_k \Big) \\ &= 0\end{aligned}$$

for all $k \in [K]$, $\ell \in [K]$. The expression for σ^2_{kk}, $\widetilde{\sigma}^2_{kk}$, $\sigma^2_{k\ell}$ and $\widetilde{\sigma}^2_{k\ell}$ follows similarly. □

The expressions for the variances $\sigma^2_{k\ell}$ and $\widetilde{\sigma}^2_{k\ell}$ of the spectral embedding estimator $\hat{\mathbf{B}}^{(S)}$ in Theorem 3 and Theorem 4 are sufficiently complicated and it is not clear, at first blush, how these variances relate to that of the estimator $\hat{\mathbf{B}}^{(N)}$ given in Theorem 1. The following corollary provides a more succinct expression for the covariance matrix of $\hat{\mathbf{B}}^{(S)}$; see Remark B.2 in the supplementary material [52] for a proof.

Corollary 3. *Assume the setting in either Theorem 3 or Theorem 4. Let* $\mathbf{\Theta}$ *be the* $K \times K$ *matrix whose elements are* $\theta_{k\ell}$ *or* $\widetilde{\theta}_{k\ell}$*, depending on whether* $\rho_n \equiv 1$ *or* $\rho_n \to 0$*. Let* $\operatorname{diag}(\boldsymbol{\pi})$ *be the diagonal matrix whose diagonal elements are the* $\pi_1, \ldots, \pi_K$ *and let* $\mathbf{V}$ *be a* $K \times d$ *orthonormal matrix whose columns are the eigenvectors of* $\mathbf{B}$ *corresponding to the non-zero eigenvalues. Define the idempotent matrix*

$$\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}} = \mathbf{I} - \mathbf{V}(\mathbf{V}^\top \operatorname{diag}(\boldsymbol{\pi}) \mathbf{V})^{-1} \mathbf{V}^\top \operatorname{diag}(\boldsymbol{\pi}).$$

We then have

$$n\rho_n^{1/2}\Big(\hat{\mathbf{B}}^{(S)} - \mathbf{B} - \frac{\mathbf{\Theta}}{n\rho_n}\Big) = n\rho_n^{1/2}(\hat{\mathbf{B}}^{(N)} - \mathbf{B}) - n\rho_n^{1/2}\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}}(\hat{\mathbf{B}}^{(N)} - \mathbf{B})(\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}})^\top + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\rho_n^{-1}).$$

Therefore, as $n \to \infty$ *with* $n\rho_n = \omega(\sqrt{n})$*, we have*

$$n\rho_n^{1/2}\operatorname{vech}\Big(\hat{\mathbf{B}}^{(S)} - \mathbf{B} - \frac{\mathbf{\Theta}}{n\rho_n}\Big) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(\mathbf{0}, \mathcal{L}_K(\mathbf{I} - \widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}} \otimes \widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}})\mathcal{D}_K \mathbf{D}^{-1} \mathcal{D}_K^\top (\mathbf{I} - \widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}} \otimes \widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}})^\top \mathcal{L}_K^\top\Big).$$

Here $\mathbf{D}$ *is the* $\binom{K+1}{2} \times \binom{K+1}{2}$ *diagonal matrix defined in Theorem 2, i.e., the diagonal entries of* $\mathbf{D}^{-1}$ *are the variances for* $\hat{\mathbf{B}}^{(N)}$*.*

The discussion in Section 1.3 and Section 1.4 specified that the estimator $\hat{\mathbf{B}}^{(N,d)}$ is obtained by truncating the SVD of $\hat{\mathbf{B}}^{(N)}$ while the spectral estimator $\hat{\mathbf{B}}^{(S)}$ is obtained by averaging the entries within and between the blocks of the truncated SVD of $\mathbf{A}$. In other words, $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(S)}$ differs only in the ordering of the averaging and truncated SVD operations. From Corollary 1 and Corollary 3, we see that this difference in the order of operations manifest itself in the definition of the projection matrices $\mathbf{\Pi}^{\perp}_{\mathbf{V}}$ and $\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}}$; we note that $\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}}$ is idempotent but not necessarily symmetric. More specifically, since the spectral estimator $\hat{\mathbf{B}}^{(S)}$ is based on the low-rank approximation of $\mathbf{A}$ which depends on both $\mathbf{B}$ and $\boldsymbol{\pi}$, the projection matrix $\widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}}$ depends on both $\mathbf{B}$ and $\boldsymbol{\pi}$. In contrast, $\hat{\mathbf{B}}^{(N,d)}$ is a function of $\hat{\mathbf{B}}^{(N)}$ and hence the projection matrix $\mathbf{\Pi}^{\perp}_{\mathbf{V}}$ depends only on $\mathbf{B}$. The limiting distributions for these two estimators are thus different, unless $\boldsymbol{\pi} = (1/K, 1/K, \ldots, 1/K)$ in which case $\operatorname{diag}(\boldsymbol{\pi}) = \frac{1}{K}\mathbf{I}$ and $\mathbf{\Pi}^{\perp}_{\mathbf{V}} = \widetilde{\mathbf{\Pi}}^{\perp}_{\mathbf{V}}$. The

two examples in Section 2.3 suggests that neither of the estimators dominate the other with respect to MSE. We leave the theoretical comparisons of these two estimators for future work.

Remark 2. The following argument indicates that the variances of $\hat{\mathbf{B}}^{(S)}$ are, under certain mild conditions, smaller than that of $\hat{\mathbf{B}}^{(N)}$ provided that K is reasonably large compared to $d = \mathrm{rk}(\mathbf{B})$. More specifically, letting $\mathrm{diag}(\boldsymbol{\pi})$ be the diagonal matrix whose diagonal elements are $\pi_1, \ldots, \pi_K$, we have

$$\zeta_{k\ell} = \nu_k^\top \Delta^{-1} \nu_\ell = (\boldsymbol{\nu}^\top (\boldsymbol{\nu}\mathrm{diag}(\boldsymbol{\pi})\boldsymbol{\nu}^\top)^{-1}\boldsymbol{\nu})_{k\ell} = \frac{h_{k\ell}}{\sqrt{\pi_k \pi_\ell}}$$

where $\boldsymbol{\nu}$ is the $d \times K$ matrix whose columns represent the point masses $\nu_1, \nu_2, \ldots, \nu_K$ and $h_{k\ell}$ are the $k\ell$-th entry of the idempotent matrix $\mathrm{diag}(\boldsymbol{\pi})^{1/2}\boldsymbol{\nu}^\top(\boldsymbol{\nu}\mathrm{diag}(\boldsymbol{\pi})\boldsymbol{\nu}^\top)^{-1}\boldsymbol{\nu}\mathrm{diag}(\boldsymbol{\pi})^{1/2}$. We note that $0 \leq |h_{k\ell}| \leq \max\{h_{kk}, h_{\ell\ell}\} \leq 1$ for all k, ℓ. As $\sum_k h_{kk} = d$, if we assume that $\max_k h_{kk} \leq cd/K$ for some fixed constant c not depending on d and K, then the variances σ_{kk}^2 are of the form

$$\begin{aligned}\sigma_{kk}^2 &= \frac{4\mathbf{B}_{kk}(1-\mathbf{B}_{kk})h_{kk}^2}{\pi_k^2} + \sum_r \frac{4\mathbf{B}_{kr}(1-\mathbf{B}_{kr})h_{kr}^2(1-2h_{kk})}{\pi_k^2} + \sum_r \sum_s \frac{2\mathbf{B}_{rs}(1-\mathbf{B}_{rs})h_{kr}^2 h_{ks}^2}{\pi_k^2} \\ &= \mathcal{O}(d^2 K^{-2}\pi_k^{-2}\mathbf{B}_{kk}(1-\mathbf{B}_{kk}))\end{aligned}$$

Thus, for a fixed d, as K increases the variances σ_{kk}^2 for $\hat{\mathbf{B}}^{(S)}$ is smaller than that of $\hat{\mathbf{B}}^{(N)}$ by a factor proportional to d^2/K^2. Similar phenomenon holds for the other variances $\sigma_{k\ell}^2$. In summary, with d fixed and K increasing, if the $\{\pi_k\}$ and the $\{\mathbf{B}_{k\ell}\}$ are sufficiently homogeneous so that $\max_k h_{kk} \leq cd/K$ for some fixed constant c, then the variances of $\hat{\mathbf{B}}^{(S)}$ is smaller than that of $\hat{\mathbf{B}}^{(N)}$ by a factor proportional to d^2/K^2.

2.1. One-step MLE update

Theorem 1 and Corollary 2 imply that the naive MLE estimate $\hat{\mathbf{B}}^{(N)}$ and the spectral embedding estimate $\hat{\mathbf{B}}^{(S)}$ are asymptotically efficient when $\mathbf{B}$ is of full-rank. These estimators are no longer asymptotically efficient if $\mathbf{B}$ is singular. Nevertheless, as $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(S)}$ both have the same rate of convergence as the true MLE estimate $\hat{\mathbf{B}}^{(M)}$, we can consider estimators obtained from one-step updates of $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(S)}$. These estimators are then asymptotically efficient even when $\mathbf{B}$ is singular.

More specifically, let $\widetilde{\mathbf{B}}$ be any estimator of $\mathbf{B}$ for which $n\rho_n^{1/2}(\mathrm{vech}(\widetilde{\mathbf{B}} - \mathbf{B}))$ converges to multivariate normal. The following argument adapts Section 5.7 of [54] for the one-step MLE in classical statistics, where the rate of convergence is $n^{1/2}$, to the SBMs setting. Assume that $d = \mathrm{rk}(\mathbf{B})$ is known and that, without loss of generality, the top left $d \times d$ block of $\mathbf{B}$ is invertible. Using the same notations as in the statement of Theorem 2, let

$$\widetilde{\eta} = \mathrm{vech}(\mathbf{B}_{11}), \quad \widetilde{\nu} = \mathrm{vec}(\mathbf{B}_{11}^{-1}\mathbf{B}_{12}), \quad \widetilde{\theta} = (\widetilde{\nu}, \widetilde{\eta}) \in \mathbb{R}^{d(d+1)/2 + d(K-d)}.$$

Now for any $\theta = (\nu, \eta) \in \mathbb{R}^{d(d+1)/2+d(K-d)}$, denote by $g^{-1}(\theta)$ the $K \times K$ matrix whose blocks structure satisfies

$$g^{-1}(\theta) = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}^\top & \mathbf{M}_{12}^\top \mathbf{M}_{11}^{-1}\mathbf{M}_{12} \end{bmatrix}, \quad \mathrm{vech}(\mathbf{M}_{11}) = \nu, \quad \mathrm{vec}(\mathbf{M}_{11}^{-1}\mathbf{M}_{12}) = \eta.$$

Here $\mathbf{M}_{11}$ and $\mathbf{M}_{12}$ are matrices of dimensions $d \times d$ and $d \times (K-d)$ respectively. Note that $g^{-1}(\widetilde{\theta}) = \widetilde{\mathbf{B}}$. Now let $\widetilde{\mathcal{J}}$ and $\widetilde{\mathbf{D}}$ be the matrices in the statement of Theorem 2 but with the entries of $\widetilde{\mathbf{B}}$ replacing

the entries of $\mathbf{B}$ and n_k/n replacing π_k; here n_k is the number of nodes assigned to block k in $\mathbf{A}$. Also let $\nabla\ell(\widetilde{\mathbf{B}}) \in \mathbb{R}^{K(K+1)/2}$ be the gradient of the log-likelihood with respect to the upper triangular entries of $\mathbf{B}$, evaluated at $\widetilde{\mathbf{B}}$, i.e., the entries of $\nabla\ell(\widetilde{\mathbf{B}})$ are indexed by the pairs (k, ℓ) with $1 \le k \le \ell \le K$ and

$$\big(\nabla\ell(\widetilde{\mathbf{B}})\big)_{k,\ell} = \frac{m_{k\ell}}{\widetilde{\mathbf{B}}_{k,\ell}} - \frac{n_{k\ell} - m_{k\ell}}{1 - \widetilde{\mathbf{B}}_{k\ell}}, \quad n_{k\ell} = \begin{cases} n_k n_\ell, & k \neq \ell \\ \binom{n_k+1}{2}, & k = \ell \end{cases}.$$

Here $m_{k\ell}$ denote the number of *observed* edges between community k and ℓ if $k \neq \ell$ and denote the number of observed edges within community k if $k = \ell$. With the above notations, define

$$\nabla\ell(\widetilde{\theta}) = \widetilde{\mathcal{J}}^\top \nabla\ell(\widetilde{\mathbf{B}}), \quad \hat{\theta} = \widetilde{\theta} - (\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \nabla\ell(\widetilde{\theta}), \quad \hat{\mathbf{B}} = g^{-1}(\hat{\theta}). \tag{2.16}$$

The estimator $\hat{\theta}$ is the one-step MLE update of $\widetilde{\theta}$ and $\hat{\mathbf{B}}$ is the corresponding estimate for $\mathbf{B}$. As $\widetilde{\mathbf{B}}$ is a $n\rho_n^{1/2}$-consistent estimator for $\mathbf{B}$, $\widetilde{\theta}$ is a $n\rho_n^{1/2}$-consistent estimator for θ. The matrix $\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}}$ then converges in probability to the Fisher information matrix $\mathcal{J}^\top \mathbf{D} \mathcal{J}$ in the statement of Theorem 2. Letting $\theta_0 = (\mathrm{vech}(\mathbf{B}_{11}), \mathrm{vec}(\mathbf{B}_{11}^{-1}\mathbf{B}_{12}))$, we have, for any fixed constant M

$$\sup_{n\rho_n^{1/2}\|\theta_* - \theta_0\| \le M} \|n\rho_n^{1/2}(\nabla\ell(\theta_*) - \nabla\ell(\theta_0)) - (\mathcal{J}^\top \mathbf{D} \mathcal{J})(n\rho_n^{1/2}(\theta_* - \theta_0)\| \xrightarrow{p} 0 \tag{2.17}$$

as $n \to \infty$. This is the analogue of Eq. (5.44) in [54]. We thus have

$$\begin{aligned} n\rho_n^{1/2}(\hat{\theta} - \theta_0) &= n\rho_n^{1/2}(\widetilde{\theta} - \theta_0) - n\rho_n^{1/2}(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \nabla\ell(\widetilde{\theta}) \\ &= n\rho_n^{1/2}(\widetilde{\theta} - \theta_0) - n\rho_n^{1/2}(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1}(\nabla\ell(\widetilde{\theta}) - \nabla\ell(\theta_0)) + n\rho_n^{1/2}(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \nabla\ell(\theta_0) \\ &= n\rho_n^{1/2}(\mathbf{I} - (\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \mathcal{J}^\top \mathbf{D} \mathcal{J})(\widetilde{\theta} - \theta_0) + n\rho_n^{1/2}(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \nabla\ell(\theta_0) + o_{\mathbb{P}}(1) \\ &= n\rho_n^{1/2}(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1} \nabla\ell(\theta_0) + o_{\mathbb{P}}(1) \\ &\xrightarrow{d} \mathcal{N}(0, (\mathcal{J}^\top \mathbf{D} \mathcal{J})^{-1}). \end{aligned}$$

In the above display, the second to last equality follows from Eq. (2.17), the last equality follows from the convergence in probability of $(\widetilde{\mathcal{J}}^\top \widetilde{\mathbf{D}} \widetilde{\mathcal{J}})^{-1}$ to $\mathcal{J}^\top \mathbf{D} \mathcal{J}$ and Slutsky's theorem, and the convergence to multivariate normal follows from the proof of Theorem 2 and Slutsky's theorem. In summary, we have the following result.

Theorem 5. *Let $\mathbf{A} \sim \mathrm{SBM}(\boldsymbol{\pi}, \mathbf{B}, \rho_n)$ be a K-block stochastic blockmodel graph on n vertices with sparsity factor ρ_n. Let $\hat{\mathbf{B}}$ be the one-step MLE update of $\widetilde{\mathbf{B}}$ where $\widetilde{\mathbf{B}}$ is either the naive MLE estimate $\hat{\mathbf{B}}^{(N)}$ or the (de-biased) spectral embedding estimate $\hat{\mathbf{B}}^{(S)}$. Then $\hat{\mathbf{B}}$ is* asymptotically efficient *whenever $\widetilde{\mathbf{B}}$ is a $n\rho_n^{1/2}$-consistent estimator for $\mathbf{B}$.*

Note that we can assume $d = \mathrm{rk}(\mathbf{B})$ is known in the statement of Theorem 5. When d is unknown then there is an estimator $\hat{d}$ such that $\hat{d}$ converges to d almost surely as $n \to \infty$. Given $\hat{d}$, we can then find, asymptotically almost surely, a $\hat{d} \times K$ sub-matrix of $\widetilde{\mathbf{B}}$ with full-row rank and use this sub-matrix to construct the one-step MLE update as given in Eq. (2.16). Finally we emphasize that in the case of the spectral estimator $\hat{\mathbf{B}}^{(S)}$, we first need to de-bias $\hat{\mathbf{B}}^{(S)}$ before applying the one-step MLE update, i.e., the one-step MLE update is applied to the estimate $\hat{\mathbf{B}}^{(S)} - \hat{\boldsymbol{\Theta}}$ where $\hat{\boldsymbol{\Theta}}$ is the $K \times K$ matrix whose entries are the estimates $\hat{\theta}_{k\ell}$ for $\theta_{k\ell}$ (when $\rho_n \equiv 1$) or $\widetilde{\theta}_{k\ell}$ (when $\rho_n \to 0$). See Corollary 4 for more details on computing $\hat{\theta}_{k\ell}$.

2.2. On superefficiency and regularity of $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$

The spectral estimate $\hat{\mathbf{B}}^{(S)}$ and the true maximum likelihood estimate $\hat{\mathbf{B}}^{(M)}$ depend on knowing or being able to consistently estimate the rank of $\mathbf{B}$. Theorem 2 and Remark 2 indicates that these estimators can out-perform the "naive" MLE $\hat{\mathbf{B}}^{(N)}$, in terms of mean square error for $\hat{\mathbf{B}}^{(S)}$ and in terms of domination of the covariance matrix for $\hat{\mathbf{B}}^{(M)}$, when $\mathbf{B}$ is singular. It is thus natural to inquire whether this improvement resembles that of super-efficient estimators such as Hodges' estimator; more specifically, whether or not the improvement in mean square error for a given singular $\mathbf{B}$ is associated with inflated (possibly infinite) asymptotic mean square error for some $\mathbf{B}'$ in a neighborhood of $\mathbf{B}$. The answer turns out to be quite interesting.

We first recall the notion of regular estimators as developed in the pioneering work of Hájek and Le Cam [27]. In classical parametric statistics, an estimator $T = \{T_n\}$ of a paremeter θ is *regular* at θ_0 if, for any sequence $\{\theta_n\}$ converging to θ_0 with $n^{1/2}(\theta_n - \theta_0)$ remaining bounded, we have $n^{1/2}(T_n - \theta_n)$ converging to some non-degenerate distribution $\mathcal{L}_{\theta_0}$ where $\mathcal{L}_{\theta_0}$ depends only on θ_0 and not on the particular sequence $\{\theta_n\}$; see e.g., Section 2.2 of [9] for a more formal definition. Note that T_n in $T_n - \theta_n$ denote the estimate given n i.i.d data points generated from a distribution with parameter θ_n. An estimator T is regular on Θ if it is regular at all $\theta_0 \in \Theta$.

We now adapt this definition to our stochastic blockmodels setting. Assume, for ease of exposition, that the sparsity factor $\rho_n \equiv 1$. We say that an estimator $\hat{\mathbf{B}}$ is regular at $\mathbf{B}$ if, for any sequence $\{\mathbf{B}_n\}$ converging to $\mathbf{B}$ with $n\|\mathbf{B}_n - \mathbf{B}\|_F$ remaining bounded, we have $n(\text{vech}(\hat{\mathbf{B}}_n - \mathbf{B}_n))$ converging to some non-degenerate distribution $\mathcal{L}_{\mathbf{B}}$ not depending on the particular sequence $\{\mathbf{B}_n\}$. Note that we assume the $\mathbf{B}_n$ to have the same number of blocks as $\mathbf{B}$, treating the block assignment probablities $\boldsymbol{\pi}$ as nuisance parameters.

With the above notions of regularity in place, we see that $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$ are regular at all full-rank $\mathbf{B}$. Indeed, let $\mathbf{B}$ be a $K \times K$ matrix of block probabilities, with K fixed but arbitrary. As $\mathbf{B}$ is full-rank, there exists a constant $\epsilon > 0$ depending on $\mathbf{B}$ but not on n such that all eigenvalues of $\mathbf{B}$ exceed ϵ. Now for any sequence $\{\mathbf{B}_n\}$ converging to $\mathbf{B}$, the $\mathbf{B}_n$ are also full-rank (for sufficiently large n). We can then, almost surely, recover the rank of the $\mathbf{B}_n$ by an eigenvalue thresholding procedure, i.e., estimate the rank of $\mathbf{B}_n$ as $\hat{d}_n$ where $\hat{d}_n$ is the number of eigenvalues of $\mathbf{A}_n$ exceeding $4\sqrt{\delta(\mathbf{A}_n)}$ in modulus where $\delta(\mathbf{A}_n)$ denote the max degree of $\mathbf{A}_n$. Then, almost surely, $\hat{d}_n = K$ for *all but a finite number* of n; indeed, for $\rho_n \equiv 1$ we have $\lambda_K(\mathbf{A}_n) = \Omega(n\epsilon) \gg 4\sqrt{\delta(\mathbf{A}_n)} \geq \lambda_{K+1}(\mathbf{A}_n)$ for all but a finite number of n. The limiting distributions of either $n(\text{vech}(\hat{\mathbf{B}}_n^{(S)} - \mathbf{B}_n))$ or $n(\text{vech}(\hat{\mathbf{B}}_n^{(M)} - \mathbf{B}_n))$ will, therefore, not depend on the particular sequence $\mathbf{B}_n$ as estimation uses the correct rank K for all but a finite number of $\mathbf{B}_n$.

The case when $\mathbf{B}$ is singular is considerably more subtle. Suppose $\mathbf{B}$ has rank $d < K$. Once again, there exists a constant $\epsilon > 0$ such that all d *non-zero* eigenvalues of $\mathbf{B}$ exceed ϵ. Then for any sequence $\mathbf{B}_n$ converging to $\mathbf{B}$, the number of non-zero eigenvalues of $\mathbf{B}_n$ exceeding ϵ is also exactly d (for sufficiently large n). Once again, letting $\hat{d}_n$ be the number of eigenvalues of $\mathbf{A}_n$ exceeding $4\sqrt{\delta(\mathbf{A}_n)}$, we have $\hat{d}_n = d$ asymptotically almost surely. However, since we only require that the sequence $\{\mathbf{B}_n\}$ converges to $\mathbf{B}$ in Frobenius norm, the ranks of the $\mathbf{B}_n$ need not converge, let alone to d. Indeed, consider the case when $\mathbf{B}_n$ has rank d for even n and rank $d+1$ for odd n. The limiting distribution of both $n(\text{vech}(\hat{\mathbf{B}}_n^{(S)} - \mathbf{B}_n))$ and $n(\text{vech}(\hat{\mathbf{B}}_n^{(M)} - \mathbf{B}_n))$ thus depend on the particular sequence $\{\mathbf{B}_n\}$ as it is possible that estimation uses mis-specified models for all n. More specifically, the limiting distribution of $n(\text{vech}(\hat{\mathbf{B}}_n^{(S)} - \mathbf{B}_n))$ and $n(\text{vech}(\hat{\mathbf{B}}_n^{(M)} - \mathbf{B}_n))$ will both have a bias term depending on $n\|\mathbf{B}'_n - \mathbf{B}_n\|_F$ where, for any n, $\mathbf{B}'_n$ is the best rank d approximation of $\mathbf{B}_n$. As the sequence $\{\mathbf{B}_n\}$ is arbitrary, the mean square error of these estimators could be arbitrarily large for all n with $d < \text{rk}(\mathbf{B}_n)$.

The estimators $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$ are thus not regular for singular $\mathbf{B}$ if our only restriction is that the sequence $n\|\mathbf{B}_n - \mathbf{B}\|_F$ remains bounded. However, if we also require that $\text{rk}(\mathbf{B}_n) \to \text{rk}(\mathbf{B})$ then $\hat{\mathbf{B}}^{(S)}$

and $\hat{\mathbf{B}}^{(M)}$ will be regular for all $\mathbf{B}$. Indeed, the estimation now uses the correct model for all but a finite number of n, and the best rank-d approximation of $\mathbf{B}_n$ is $\mathbf{B}_n$ itself. The regularity of $\hat{\mathbf{B}}^{(M)}$ and $\hat{\mathbf{B}}^{(S)}$ imply that their mean square error will not be inflated. Therefore, if $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$ improve over $\hat{\mathbf{B}}^{(N)}$ at some fixed $\mathbf{B}$, then $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(M)}$ also improve over $\hat{\mathbf{B}}^{(N)}$ for all $\mathbf{B}_n$ in a neighborhood of $\mathbf{B}$ and having the same rank as $\mathbf{B}$. In summary, we have the following result.

Proposition 1. *Let* $\mathbf{B}$ *be a* $K \times K$ *matrix with* $\mathrm{rk}(\mathbf{B}) = d \leq K$. *Assume* $\rho_n \equiv 1$. *Then for all sequence* $\{\mathbf{B}_n\}$ *such that* $\limsup n\|\mathbf{B}_n - \mathbf{B}\|_F < \infty$ *and* $\mathrm{rk}(\mathbf{B}_n) \to \mathbf{B}$, *we have*

$$\limsup n^2 \mathbb{E}[\|\hat{\mathbf{B}}_n^{(S)} - \mathbf{B}_n\|_F^2] < \infty, \quad \textit{and} \quad \limsup n^2 \mathbb{E}[\|\hat{\mathbf{B}}_n^{(M)} - \mathbf{B}_n\|_F^2] < \infty.$$

That is, the estimators $\hat{\mathbf{B}}^{(S)}$ *and* $\hat{\mathbf{B}}^{(M)}$ *are regular at* $\mathbf{B}$ *for any sequence* $\mathbf{B}_n \to \mathbf{B}$ *that also satisfies* $\mathrm{rk}(\mathbf{B}_n) \to \mathrm{rk}(\mathbf{B})$.

We emphasize that the possibility of having $\{\mathbf{B}_n\}$ converging to $\mathbf{B}$ along a sequence of matrices with the same ranks as $\mathbf{B}$ is the main distinguishing feature between our stochastic blocksmodel setting and Hodges' phenomenon. More specifically, since Hodges' estimator is superefficient only at some (collection of) discrete point(s) $\theta_0 \in \mathbb{R}^p$, one cannot guarantee that, given a sequence $\{\theta_n\}$ converging to θ_0, the θ_n will not lie in the set with inflated mean square error, i.e., the set of superefficient points is not connected. In contrast, for a given d, the subset of $K \times K$ matrices with rank d is *connected* and hence we can have regularity along this connected subset.

Finally we note that the use of low-rank estimation can also be justified under the umbrella of oracle procedures, i.e., for a *fixed* $\mathbf{B}$, the performance of the estimators $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$ when the rank of $\mathbf{B}$ is unknown is, asymptotically, equivalent to the performance of an *oracle* estimator which knows the rank of $\mathbf{B}$. Oracle properties are desirable in sparse and low-rank estimation, see e.g., the motivation in [21]. However, it is also well-known that estimators with oracle properties are not regular in the classical sense of Hájek and Le Cam and will thus have inflated errors that can not be bounded uniformly in the vicinity of any non-regular point [16,33]. Our notion of regularity is therefore slightly more restrictive but more suitable for low-rank estimation.

In summary, if we allow for the possibility that $\mathbf{B}$ is singular and wish to leverage this restriction (when appropriate) for improved performance, then the estimators $\hat{\mathbf{B}}^{(S)}$ and $\hat{\mathbf{B}}^{(M)}$ have desirable pointwise asymptotic behavior, have oracle properties, and are regular under the slightly more restrictive condition that the sequence $\{\hat{\mathbf{B}}_n\}$ converging to some singular $\mathbf{B}$ have the same ranks as $\mathbf{B}$.

2.3. Examples

Remark 2 indicates that when d is small enough compared to K, spectral estimator $\hat{\mathbf{B}}^{(S)}$ could have smaller mean square error than that of $\hat{\mathbf{B}}^{(N)}$. The following examples illustrate that the improvement in mean square error can occur even when $d = K - 1$.

Remark 3. As a special case of Theorem 3, we consider the two blocks stochastic blockmodel with block probability matrix $\mathbf{B} = \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix}$ and block assignment probabilities $\boldsymbol{\pi} = (\pi_p, \pi_q)$, $\pi_p + \pi_q = 1$. Then $\Delta = \pi_p p^2 + \pi_q q^2$ and Eq. (2.3) reduces to

$$\theta_{11} = \tfrac{2\pi_q p^2 q^2}{\Delta^3}\big(\pi_p p^2(1-p^2) + (\pi_q - \pi_p)pq(1-pq) - \pi_q q^2(1-q^2)\big),$$
$$\theta_{12} = \tfrac{pq}{\Delta^3}\big(\pi_p p^2(1-p^2)(\pi_q q^2 - \pi_p p^2) + (\pi_p - \pi_q)pq(1-pq)(\pi_p q^2 - \pi_q p^2)$$

$$+\pi_q q^2(1-q^2)(\pi_p p^2 - \pi_q q^2)\big),$$

$$\theta_{22} = \tfrac{2\pi_p p^2 q^2}{\Delta^3}\big(\pi_q q^2(1-q^2) + (\pi_p - \pi_q)pq(1-pq) - \pi_p p^2(1-p^2)\big).$$

Meanwhile, we also have

$$\sigma_{11}^2 = \tfrac{8p^6(1-p^2)}{\Delta^2}\big(1 - \tfrac{\pi_p p^2}{2\Delta}\big)^2 + \tfrac{4\pi_q p^3 q^3(1-pq)}{\pi_p \Delta^2}\big(1 - \tfrac{\pi_p p^2}{\Delta}\big)^2 + \tfrac{2\pi_q^2 p^4 q^6(1-q^2)}{\Delta^4}$$

$$\sigma_{12}^2 = \tfrac{2\pi_q^2 p^4 q^6(1-p^2)}{\Delta^4} + \tfrac{\pi_p \pi_q pq(1-pq)}{\Delta^4}\big(\tfrac{\pi_q q^4}{\pi_p} + \tfrac{\pi_p p^4}{\pi_q}\big)^2 + \tfrac{2\pi_p^2 p^6 q^4(1-q^2)}{\Delta^4}$$

$$\sigma_{22}^2 = \tfrac{8q^6(1-q^2)}{\Delta^2}\big(1 - \tfrac{\pi_q q^2}{2\Delta}\big)^2 + \tfrac{4\pi_p p^3 q^3(1-pq)}{\pi_q \Delta^2}\big(1 - \tfrac{\pi_q q^2}{\Delta}\big)^2 + \tfrac{2\pi_p^2 q^4 p^6(1-p^2)}{\Delta^4}.$$

The naive (MLE) estimator $\hat{\mathbf{B}}^{(N)}$ has asymptotic variances

$$\mathrm{Var}[\hat{\mathbf{B}}_{11}^{(N)}] = \tfrac{2p^2(1-p^2)}{\pi_p^2}; \quad \mathrm{Var}[\hat{\mathbf{B}}_{12}^{(N)}] = \tfrac{pq(1-pq)}{\pi_p \pi_q}; \quad \mathrm{Var}[\hat{\mathbf{B}}_{22}^{(N)}] = \tfrac{2q^2(1-q^2)}{\pi_q^2}.$$

We now evaluate the asymptotic variances for the rank-constrained MLE $\hat{\mathbf{B}}^{(M)}$. Suppose for simplicity that $n\pi_p$ vertices are assigned to block 1 and $n\pi_q$ vertices are assigned to block 2. Let $n_{11} = \binom{(n\pi_p+1)}{2}$, $n_{12} = n^2\pi_p\pi_q$ and $n_{22} = \binom{(n\pi_p+1)}{2}$. Let $\mathbf{A}$ be given and assume for the moment that $\boldsymbol{\tau}$ is observed. Then the log-likelihood for $\mathbf{A}$ is equivalent to the log-likelihood for observing $m_{11} \sim \mathrm{Bin}(n_{11}, p^2)$, $m_{12} \sim \mathrm{Bin}(n_{12}, pq)$ and $m_{22} \sim \mathrm{Bin}(n_{22}, q^2)$ with m_{11}, m_{12}, and m_{22} mutually independent. More specifically, ignoring terms of the form $\binom{n_{ij}}{m_{ij}}$, we have

$$\begin{aligned}\ell(\mathbf{A} \mid p, q) = {} & m_{11}\log p^2 + (n_{11} - m_{11})\log(1-p^2) + m_{12}\log pq \\ & + (n_{12} - m_{12})\log(1-pq) + m_{22}\log q^2 + (n_{22} - m_{22})\log(1-q^2)\end{aligned}$$

We therefore have

$$\mathrm{Var}\Big(\tfrac{\partial \ell}{\partial p}\Big) = \mathrm{Var}\Big(\tfrac{2m_{11}p}{p^2} - \tfrac{2(n_{11}-m_{11})p}{1-p^2} + \tfrac{m_{12}q}{pq} - \tfrac{(n_{12}-m_{12})q}{1-pq}\Big) = \tfrac{4n_{11}}{1-p^2} + \tfrac{n_{12}pq}{p^2(1-pq)},$$

$$\mathrm{Var}\Big(\tfrac{\partial \ell}{\partial q}\Big) = \tfrac{n_{12}pq}{q^2(1-pq)} + \tfrac{4n_{22}}{1-q^2}, \quad \mathrm{Cov}\Big(\tfrac{\partial \ell}{\partial p}, \tfrac{\partial \ell}{\partial q}\Big) = \tfrac{n_{12}}{1-pq}.$$

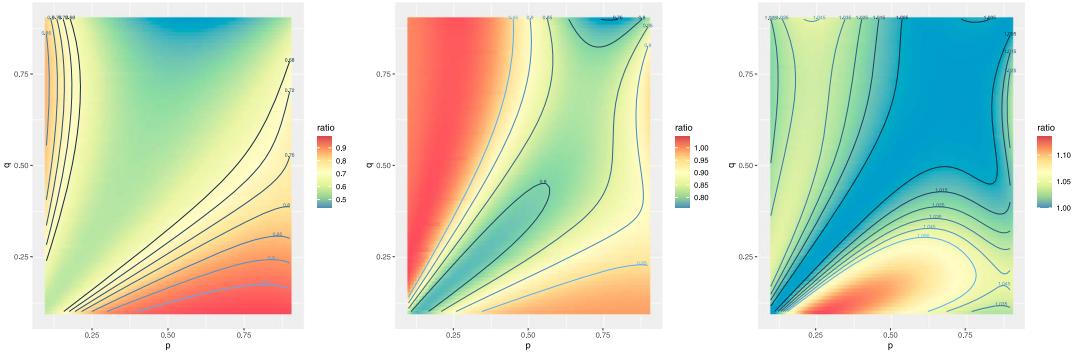
Now $n_{11}/n^2 \to \pi_p^2/2$, $n_{12}/n^2 \to \pi_p\pi_q$ and $n_{22}/n^2 \to \pi_q^2/2$. We therefore have

$$\frac{1}{n^2}\mathrm{Var}\Big[\Big(\tfrac{\partial \ell}{\partial(p,q)}, \tfrac{\partial \ell}{\partial(p,q)}\Big)\Big] \xrightarrow{\text{a.s.}} \mathcal{I} := \begin{bmatrix} \dfrac{2\pi_p^2}{1-p^2} + \dfrac{\pi_p\pi_q q}{p(1-pq)} & \dfrac{\pi_p\pi_q}{1-pq} \\ \dfrac{\pi_p\pi_q}{1-pq} & \dfrac{2\pi_q^2}{1-q^2} + \dfrac{\pi_p\pi_q p}{q(1-pq)} \end{bmatrix}$$

as $n \to \infty$. Let $(\hat{p}, \hat{q})$ be the MLE of (p, q) (there is no close form expression for $\hat{p}$ and $\hat{q}$ in this setting) and let $\mathcal{J}^\top = \begin{bmatrix} 2p & q & 0 \\ 0 & p & 2q \end{bmatrix}$ be the Jacobian of $\mathbf{B}$ with respect to p and q. Theorem 2 then implies, as $n \to \infty$,

$$n\big(\mathrm{vech}(\hat{\mathbf{B}}^{(M)} - \mathbf{B})\big) \xrightarrow{\text{d}} \mathrm{MVN}\big(\mathbf{0}, \mathcal{J}\mathcal{I}^{-1}\mathcal{J}^\top\big).$$

We now compare the naive estimator $\hat{\mathbf{B}}^{(N)}$ and its truncated rank-1 approximation (see Section 1.3), the adjacency spectral embedding estimator $\hat{\mathbf{B}}^{(S)}$, and the true maximum likelihood estimator $\hat{\mathbf{B}}^{(M)}$.



(a) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N)})$ (b) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N,1)})$ (c) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(M)})$

Figure 1. The ratios $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N)})$, $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N,1)})$, and $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(M)})$ for values of $p \in [0.1, 0.9]$ and $q \in [0.1, 0.9]$ in a 2-blocks rank 1 SBM with $\boldsymbol{\pi} = (0.7, 0.3)$. The labeled lines in each plot are the contour lines for the MSE ratios. The $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})$ are computed using the bias-adjusted estimates $\{\hat{\mathbf{B}}^{(S)}_{ij} - \hat{\theta}_{ij}\}_{i \le j}$ (see Corollary 4). The $\mathrm{MSE}(\hat{\mathbf{B}}^{(N,1)})$ are for the truncated rank-1 approximations of $\hat{\mathbf{B}}^{(N)}$ (see Corollary 1).

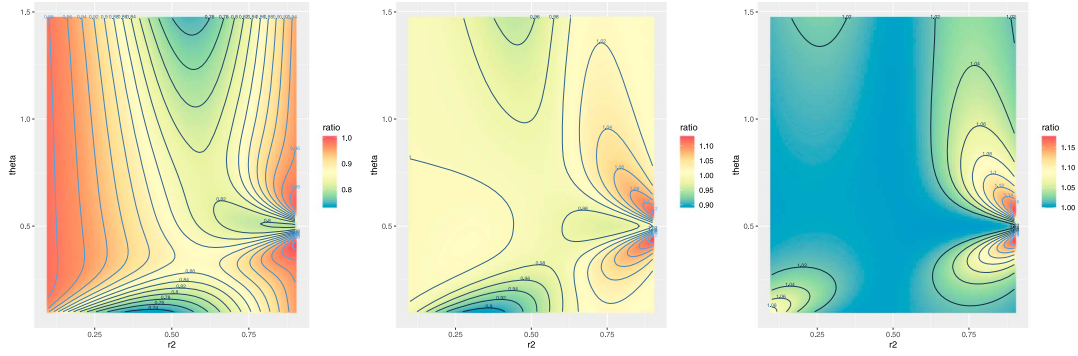
We fixed $(\pi_p, \pi_q) = (0.7, 0.3)$ and $\rho_n \equiv 1$ and let p and q varies in the interval $[0.1, 0.9]$. Plots of the (ratio) of the MSE for $\hat{\mathbf{B}}^{(S)}$ (adjusted for the bias terms $\theta_{k\ell}$; see Corollary 4) against the MSE of $\hat{\mathbf{B}}^{(N)}$ and its truncated rank-1 approximation $\hat{\mathbf{B}}^{(N,1)}$, and the MSE of $\hat{\mathbf{B}}^{(M)}$, are given in Figure 1. For this example, $\hat{\mathbf{B}}^{(S)}$ have smaller mean squared error than both $\hat{\mathbf{B}}^{(N)}$ and $\hat{\mathbf{B}}^{(N,1)}$ over the whole range of p and q and furthermore, $\hat{\mathbf{B}}^{(S)}$ has mean squared error almost as small as that of $\hat{\mathbf{B}}^{(M)}$ for a large range of p and q. We emphasize that the MSE as presented are exact theoretical quantities, e.g., they are computed using Theorem 2, Theorem 3, or Corollary 1

Remark 4. As another example, we compare the estimators $\hat{\mathbf{B}}^{(S)}$, $\hat{\mathbf{B}}^{(M)}$, and $\hat{\mathbf{B}}^{(N)}$ and its truncated low-rank approximation in the setting of a 3-blocks SBM $\mathbf{B}$ with $\mathrm{rk}(\mathbf{B}) = 2$. A minimal parametrization of $\mathbf{B}$ requires 5 parameters $(r_1, r_2, r_3, \theta, \gamma)$, i.e.,

$$\mathbf{B}_{11} = r_1^2; \quad \mathbf{B}_{22} = r_2^2; \quad \mathbf{B}_{33} = r_3^2;$$
$$\mathbf{B}_{12} = r_1 r_2 \cos\theta; \quad \mathbf{B}_{13} = r_1 r_3 \cos\gamma; \quad \mathbf{B}_{23} = r_2 r_3 \cos(\theta - \gamma).$$

Now let $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ be the block assignment probability vector. Let $\mathbf{A} \sim \mathrm{SBM}(\mathbf{B}, \boldsymbol{\pi})$ be a graph on n vertices and suppose for simplicity that the number of vertices in block i is $n_i = n\pi_i$. Let $n_{ii} = n_i^2/2$ for $i = 1, 2, 3$ and $n_{ij} = n_i n_j$ if $i \neq j$. Let m_{ij} for $i \le j$ be independent random variables with $m_{ij} \sim \mathrm{Bin}(n_{ij}, \mathbf{B}_{ij})$. Then, assuming $\boldsymbol{\tau}$ is known, the log-likelihood for $\mathbf{A}$ is

$$\begin{aligned}
\ell(\mathbf{A}) = {} & m_{11}\log(r_1^2) + (n_{11} - m_{11})\log(1 - r_1^2) + m_{22}\log(r_2^2) + (n_{22} - m_{22})\log(1 - r_2^2) \\
& + m_{33}\log(r_3^2) + (n_{33} - m_{33})\log(1 - r_3^2) + m_{12}\log(r_1 r_2 \cos\theta) \\
& + (n_{12} - m_{12})\log(1 - r_1 r_2 \cos\theta) + m_{13}\log(r_1 r_3 \cos\gamma) \\
& + (n_{13} - m_{13})\log(1 - r_1 r_3 \cos\gamma) + m_{23}\log(r_2 r_3 \cos(\theta - \gamma)) \\
& + (n_{23} - m_{23})\cos(1 - r_2 r_3 \cos(\theta - \gamma))
\end{aligned}$$



(a) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N)})$ (b) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N,2)})$ (c) $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(M)})$

Figure 2. The ratios $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(N)})$, $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)}/\mathrm{MSE}(\hat{\mathbf{B}}^{(N,2)}$ and $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})/\mathrm{MSE}(\hat{\mathbf{B}}^{(M)})$ for values of $r_2 \in [0.1, 0.9]$ and $\theta \in [0.1, \frac{\pi}{2} - 0.1]$ in a 3-block, rank 2 SBM. The labeled lines in each plot are the contour lines for the MSE ratios. The $\mathrm{MSE}(\hat{\mathbf{B}}^{(S)})$ are computed using the bias-adjusted estimates $\{\hat{\mathbf{B}}_{ij}^{(S)} - \hat{\theta}_{ij}\}_{i \le j}$ (see Corollary 4). The $\mathrm{MSE}(\hat{\mathbf{B}}^{(N,2)})$ are for the truncated rank-2 approximations of $\hat{\mathbf{B}}^{(N)}$ (see Corollary 1).

The Fisher information matrix $\mathcal{I}$ for $(r_1, r_2, r_3, \theta, \gamma)$ in this setting is straightforward to derive but tedious to present and we omit it here. Plots of the MSE for the bias-adjusted estimates $\hat{\mathbf{B}}^{(S)}$ against the MSE of the naive estimates $\hat{\mathbf{B}}^{(N)}$ and their truncated rank-2 approximation $\hat{\mathbf{B}}^{(N,2)}$, and the true MLE estimates $\hat{\mathbf{B}}^{(M)}$ for some subset of $\mathbf{B}$ are given in Figure 2. In Figure 2, we fix $\boldsymbol{\pi} = (0.2, 0.3, 0.5)$, $r_3 = 0.7$, $\gamma = 0.5$ and $r_1 = 1 - r_2$. We then let r_2 and θ vary in the intervals $[0.1, 0.9]$ and $[0.1, 1.48]$, respectively. The MSE presented here are the exact limiting MSE as computed using Theorem 2, Theorem 3 and Corollary 1. Once again $\hat{\mathbf{B}}^{(S)}$ has smaller mean squared error than $\hat{\mathbf{B}}^{(N)}$ over the whole range of r_2 and θ, and has mean squared error almost as small as that of $\hat{\mathbf{B}}^{(M)}$ for a large range of r_2 and θ. Finally neither $\hat{\mathbf{B}}^{(S)}$ nor $\hat{\mathbf{B}}^{(N,2)}$ dominates the other with respect to MSE over the whole parameter space.

We emphasize that the rank assumptions placed on $\mathbf{B}$ in the previous two examples are natural assumptions, i.e., *apriori* there is no reason why $\mathbf{B}$ needs to be invertible, and hence procedures that can both estimate $\mathrm{rk}(\mathbf{B})$ and incorporate it in the subsequent estimation of $\mathbf{B}$ are equally flexible and generally more efficient. Indeed, the assumption that $\mathrm{rk}(\mathbf{B}) < K$ is natural whenever the communities can be described by latent factors or when the dimension used for the spectral embedding is different from the number of clusters. As a motivating example, given Euclidean data in $\mathbb{R}^d$, it is quite common to cluster the data using k-means or hierarchical clustering into $K \neq d$ clusters Now since stochastic blockmodel graphs are widely used as a generative model for spectral clustering problems, it is therefore also natural to have $K \neq d$. However, in contrast to Euclidean data, any stochastic blockmodel graph with K communities/cluster is necessarily of rank $d \le K$, and thus the rank assumption of this paper is both natural and sufficient for all applications of stochastic blockmodels to model networks with community structure. This is in contrast to other potentially more restrictive assumptions such as assuming that $\mathbf{B}$ is of the form $q\mathbf{1}\mathbf{1}^\top + (p-q)\mathbf{I}$ for $p > q$ (i.e., the planted-partition model). Indeed, a K-block SBM from the planted-partition model is parametrized by two parameters, irrespective of K and as such the three estimators considered in this paper are provably sub-optimal for estimating the parameters of the planted partition model. Note, however, that this sub-optimality is only pointwise, i.e., estimators for the planted-partition model might not be regular (c.f. Section 2.2)

3. Discussions

Theorem 3 and Theorem 4 were presented in the context wherein the vertices to block assignments $\boldsymbol{\tau}$ are assumed known. For unknown $\boldsymbol{\tau}$, Lemma 1 (presented below) implies that $\hat{\boldsymbol{\tau}}$ obtained using K-means (or Gaussian mixture modeling) on the rows of $\hat{\mathbf{U}}$ is an exact recovery of $\boldsymbol{\tau}$, provided that $n\rho_n = \omega(\log^2 n)$. The lemma implies Corollary 4 showing that we can replace the quantities $\theta_{k\ell}$ and $\widetilde{\theta}_{k\ell}$ in Eq. (2.6) and Eq. (2.10) of Theorem 3 and Theorem 4 by consistent estimates $\hat{\theta}_{k\ell}$ without changing the resulting limiting distribution. We emphasize that it is essential for Corollary 4 that $\hat{\boldsymbol{\tau}}$ is an exact recovery of $\boldsymbol{\tau}$ in order for the limiting distributions in Eq. (2.6) and Eq. (2.10) to remain valid when $\hat{\theta}_{k\ell}$ is substituted for $\theta_{k\ell}$ and $\widetilde{\theta}_{k\ell}$. Indeed, if there is even a single vertex that is mis-clustered by $\hat{\boldsymbol{\tau}}$, then $\hat{\theta}_{k\ell}$ as defined will introduce an additional (random) bias term in the limiting distribution of Eq. (3.3).

Remark 5. For ease of exposition, bounds in this paper are often written as holding "with high probability". A random variable $\xi \in \mathbb{R}$ is $O_{\mathbb{P}}(f(n))$ if, for any positive constant $c > 0$ there exists a $n_0 \in \mathbb{N}$ and a constant $C > 0$ (both of which possibly depend on c) such that for all $n \geq n_0$, $|\xi| \leq Cf(n)$ with probability at least $1 - n^{-c}$; moreover, a random variable $\xi \in \mathbb{R}$ is $o_{\mathbb{P}}(f(n))$ if for any positive constant $c > 0$ and any $\epsilon > 0$ there exists a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $|\xi| \leq \epsilon f(n)$ with probability at least $1 - n^{-c}$. Similarly, when ξ is a random vector in $\mathbb{R}^d$ or a random matrix in $\mathbb{R}^{d_1 \times d_2}$, $\xi = O_{\mathbb{P}}(f(n))$ or $\xi = o_{\mathbb{P}}(f(n))$ if $\|\xi\| = O_{\mathbb{P}}(f(n))$ or $\|\xi\| = o_{\mathbb{P}}(f(n))$, respectively. Here $\|x\|$ denotes the Euclidean norm of x when x is a vector and the spectral norm of x when x is a matrix. We write $\xi = \zeta + O_{\mathbb{P}}(f(n))$ or $\xi = \zeta + o_{\mathbb{P}}(f(n))$ if $\xi - \zeta = O_{\mathbb{P}}(f(n))$ or $\xi - \zeta = o_{\mathbb{P}}(f(n))$, respectively.

Lemma 1. *Let* $(\mathbf{A}_n, \mathbf{X}_n) \sim \mathrm{GRDPG}_{p,q}(F)$ *be a generalized random dot product graph on* n *vertices with sparsity factor* ρ_n. *Let* $\hat{\mathbf{U}}_n(i)$ *and* $\mathbf{U}_n(i)$ *be the* i*-th row of* $\hat{\mathbf{U}}_n$ *and* $\mathbf{U}_n$, *respectively. Here* $\hat{\mathbf{U}}_n$ *and* $\mathbf{U}_n$ *are the eigenvectors of* $\mathbf{A}_n$ *and* $\mathbf{X}_n\mathbf{X}_n^\top$ *corresponding to the* $p+q$ *largest eigenvalues* (*in modulus*) *of* $\mathbf{A}_n$ *and* $\mathbf{X}_n\mathbf{X}_n^\top$. *Then there exist a* $d \times d$ *orthogonal matrix* $\mathbf{W}_n$ *such that, for* $n\rho_n = \omega(\log^2(n))$,

$$\max_{i \in [n]} \|\mathbf{W}_n \hat{\mathbf{U}}_n(i) - \mathbf{U}_n(i)\| = O_{\mathbb{P}}\Big(\frac{\log n}{n\sqrt{\rho_n}}\Big). \tag{3.1}$$

The proof of Lemma 1 is given in Section B of the supplementary material [52]. If $\mathbf{A}_n \sim \mathrm{SBM}(\mathbf{B}, \boldsymbol{\pi})$ with sparsity factor ρ_n and $\mathbf{B}$ is $K \times K$, then the rows of $\mathbf{U}_n$ take on at most K possible distinct values. Moreover, for any vertices i and j with $\tau_i \neq \tau_j$, $\|\mathbf{U}_n(i) - \mathbf{U}_n(j)\| \geq Cn^{-1/2}$ for some constant C depending only on $\mathbf{B}$. Now if $n\rho_n = \omega(\log^2 n)$, then Lemma 1 implies, for sufficiently large n,

$$\|\mathbf{W}_n \hat{\mathbf{U}}_n(i) - \mathbf{U}_n(i)\| < \min_{j\,:\,\tau_j \neq i} \|\mathbf{W}_n \hat{\mathbf{U}}_n(i) - \mathbf{U}_n(j)\|, \quad \text{for all } i \in [n].$$

Hence, since $\mathbf{W}_n$ is an orthogonal matrix, K-means clustering of the rows of $\hat{\mathbf{U}}_n$ yield an assignment $\hat{\boldsymbol{\tau}}$ that is indeed, up to a permutation of the block labels, an exact recovery of $\boldsymbol{\tau}$ as $n \to \infty$. We note that Lemma 1 is an extension of our earlier results on bounding the perturbation $\hat{\mathbf{U}}_n - \mathbf{U}_n\mathbf{W}$ using the $2 \to \infty$ matrix norm [11,37,38]. Lemma 1 is very similar in flavor to other recent results [3,18, 41] where eigenvector perturbations of $\mathbf{A}_n$ (compared to the eigenvectors of $\mathbf{X}_n\mathbf{X}_n^\top$) in the ℓ_∞ norm is established in the regime where $n\rho_n = \omega(\log^c n)$ for some constant $c > 0$. Finally, we emphasize that Lemma 1 allows for exact recovery in stochastic blockmodel graphs where the number of blocks $K = K(n)$ increases with n. Indeed, as long as the *minimum* distance $\min_{ij\,:\,\tau_i \neq \tau_j} \|\mathbf{U}_n(i) - \mathbf{U}_n(j)\|$ between the cluster centroids exceeds the upperbound in Eq. (3.1), then vertices in the same blocks will always be closer, in the spectral embedding space for $\mathbf{A}$, than vertices in different blocks. Using a

packing number argument, we can show that there are stochastic blockmodels with $K(n) = n/(\log^2 n)$ blocks that satisfy this minimum distance requirement.

Corollary 4. *Assume the setting and notations of Theorem 3. Assume K known, let $\hat{\boldsymbol{\tau}}: [n] \mapsto [K]$ be the vertex to cluster assignments when the rows of $\hat{\mathbf{U}}$ are clustered into K clusters. For $k \in [K]$, let $\hat{\mathbf{s}}_k \in \{0,1\}^n$ where the i-th entry of $\hat{\mathbf{s}}_k$ is 1 if $\hat{\tau}_i = k$ and 0 otherwise. Let $\hat{n}_k = |\{i : \hat{\tau}_i = k\}|$ and let $\hat{\pi}_k = \frac{\hat{n}_k}{n}$. For $k \in [K]$, let $\hat{\nu}_k = \frac{1}{\hat{n}_k}\hat{\mathbf{s}}_k^\top \hat{\mathbf{U}}|\hat{\mathbf{\Lambda}}|^{1/2}$, let $\hat{\mathbf{B}}_{k\ell} = \hat{\mathbf{B}}^{(S)}_{k\ell} = \hat{\nu}_k^\top \mathbf{I}_{p,q} \hat{\nu}_\ell$, and let $\hat{\Delta} = \sum_k \hat{\pi}_k \hat{\nu}_k \hat{\nu}_k^\top$. For $k \in [K]$ and $\ell \in [K]$, let $\hat{\theta}_{k\ell}$ be given by*

$$\begin{aligned}\hat{\theta}_{k\ell} = &\sum_{r=1}^{K} \hat{\pi}_r \big(\hat{\mathbf{B}}_{kr}(1-\hat{\mathbf{B}}_{kr}) + \hat{\mathbf{B}}_{\ell r}(1-\hat{\mathbf{B}}_{\ell r})\big)\hat{\nu}_k^\top \hat{\Delta}^{-1}\mathbf{I}_{p,q}\hat{\Delta}^{-1}\nu_\ell \\ &- \sum_{r=1}^{K}\sum_{s=1}^{K} \hat{\pi}_r\hat{\pi}_s\hat{\mathbf{B}}_{rs}(1-\hat{\mathbf{B}}_{sr})\hat{\nu}_s^\top \hat{\Delta}^{-1}\mathbf{I}_{p,q}\hat{\Delta}^{-1}(\hat{\nu}_\ell\hat{\nu}_k^\top + \hat{\nu}_k\hat{\nu}_\ell^\top)\hat{\Delta}^{-1}\hat{\nu}_s.\end{aligned} \tag{3.2}$$

Then there exists a (sequence of) permutation(s) $\psi \equiv \psi_n$ on $[K]$ such that for any $k \in [K]$ and $\ell \in [K]$,

$$n(\hat{\mathbf{B}}^{(S)}_{\psi(k),\psi(\ell)} - \mathbf{B}_{k\ell} - \tfrac{\hat{\theta}_{k\ell}}{n}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_{k\ell}) \tag{3.3}$$

as $n \to \infty$.

An almost identical result holds in the setting when $\rho_n \to 0$. More specifically, assume the setting and notations of Theorem 4 and let $\hat{\nu}_k$, $\hat{\Delta}$ and $\hat{\mathbf{B}} = \hat{\mathbf{B}}^{(S)}$ be as defined in Corllary 4. Now let $\hat{\theta}_{k\ell}$ be given by

$$\begin{aligned}\hat{\theta}_{k\ell} = &\sum_{r=1}^{K} \hat{\pi}_r \big(\hat{\mathbf{B}}_{kr} + \hat{\mathbf{B}}_{\ell r}\big)\hat{\nu}_k^\top \hat{\Delta}^{-1}\mathbf{I}_{p,q}\hat{\Delta}^{-1}\nu_\ell \\ &- \sum_{r=1}^{K}\sum_{s=1}^{K} \hat{\pi}_r\hat{\pi}_s\hat{\mathbf{B}}_{rs}\hat{\nu}_s^\top \hat{\Delta}^{-1}\mathbf{I}_{p,q}\hat{\Delta}^{-1}(\hat{\nu}_\ell\hat{\nu}_k^\top + \hat{\nu}_k\hat{\nu}_\ell^\top)\hat{\Delta}^{-1}\hat{\nu}_s.\end{aligned} \tag{3.4}$$

Then there exists a (sequence of) permutation(s) $\psi \equiv \psi_n$ on $[K]$ such that for any $k \in [K]$ and $\ell \in [K]$,

$$n\rho_n^{1/2}(\hat{\mathbf{B}}_{\psi(k),\psi(\ell)} - \mathbf{B}_{k\ell} - \tfrac{\hat{\theta}_{k\ell}}{n\rho_n}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \widetilde{\sigma}^2_{k\ell}) \tag{3.5}$$

as $n \to \infty$, $\rho_n \to 0$ and $n\rho_n = \omega(\sqrt{n})$.

Finally, we provide some justification on the necessity of the assumption $n\rho_n = \omega(\sqrt{n})$ in the statement of Theorem 4, even though Lemma 1 implies that $\hat{\boldsymbol{\tau}}$ is an exact recovery of $\boldsymbol{\tau}$ for $n\rho_n = \omega(\log^2(n))$. Consider the case of $\mathbf{A}$ being an Erdős-Rényi graph on n vertices with edge probability p. The estimate $\hat{p}$ obtained from the spectral embedding in this setting is $\frac{1}{n^2}\hat{\lambda}(\mathbf{1}^\top\hat{\boldsymbol{u}})^2$ where $\mathbf{1}$ is the all ones vector, $\hat{\lambda}$ is the largest eigenvalue of $\mathbf{A}$, and $\hat{\boldsymbol{u}}$ is the associated (unit-norm) eigenvector of $\mathbf{A}$. Let $\boldsymbol{e} = n^{-1/2}\mathbf{1}$. We then have

$$n(\hat{p} - p) = \tfrac{1}{n}\hat{\lambda}(\mathbf{1}^\top\hat{\boldsymbol{u}})^2 - np = \hat{\lambda}\big((\boldsymbol{e}^\top\hat{\boldsymbol{u}})^2 - 1\big) + \hat{\lambda} - np.$$

When p remains a constant as n changes, then $\boldsymbol{e}^\top \hat{\boldsymbol{u}} = 1 - \frac{1-p}{2np} + O_{\mathbb{P}}(n^{-3/2})$ and $\hat{\lambda} - np = \boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])\boldsymbol{e} + (1-p) + O_{\mathbb{P}}(n^{-1/2})$ [24]. We thus infer

$$\begin{aligned} n(\hat{p} - p) &= -(1-p)\tfrac{\hat{\lambda}}{np} + \boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])\boldsymbol{e} + (1-p) + O_{\mathbb{P}}(n^{-1/2}) \\ &= \boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])\boldsymbol{e} + O_{\mathbb{P}}(n^{-1/2}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2p(1-p)) \end{aligned}$$

as $\boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])\boldsymbol{e}$ is a sum of $n(n+1)/2$ independent mean 0 random variables with variance $n^{-1}p(1-p)$. On the other hand, if $p \to 0$ as n increases, then Theorem 6.2 of [19] (more specifically Eq. (6.9) and Eq. (6.26) of [19]) implies

$$\boldsymbol{e}^\top \hat{\boldsymbol{u}} = 1 - \tfrac{1-p}{2np} + O_{\mathbb{P}}\big((np)^{-3/2} + \tfrac{\log^c n}{n\sqrt{p}}\big) \tag{3.6}$$

and

$$\begin{aligned} \hat{\lambda} - np &= \boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e} + \tfrac{\boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])^2\boldsymbol{e}}{np} + O_{\mathbb{P}}\big((np)^{-1} + \tfrac{\log^c n}{n\sqrt{p}}\big) \\ &= \boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e} + (1-p) + O_{\mathbb{P}}\big((np)^{-1} + \tfrac{\log^c n}{n\sqrt{p}}\big). \end{aligned} \tag{3.7}$$

The second equality in Eq. (3.7) follows from Lemma 6.5 of [19] which states that $\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])^k\boldsymbol{e} = \boldsymbol{e}^\top\mathbb{E}[(\mathbf{A} - \mathbb{E}[\mathbf{A}])^k]\boldsymbol{e} + O_{\mathbb{P}}(\frac{(np)^{k/2}\log^{kc}(n)}{\sqrt{n}})$ for some universal constant $c > 0$ provided that $np = \omega(\log n)$. Hence

$$n(\hat{p} - p) = -(1-p)\tfrac{\hat{\lambda}}{np} + \boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e} + (1-p) + O_{\mathbb{P}}((np)^{-1/2}). \tag{3.8}$$

Once again $\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e}$ is a sum of $n(n+1)/2$ independent mean 0 random variables with variance $n^{-1}p(1-p)$, but since $p \to 0$, the individual variance vanishes at a faster rate than $O(n^{-1})$ as $n \to \infty$. In order to obtain a non-degenerate limiting distribution for $\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e}$, it is necessary that we consider $p^{-1/2}\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e}$. This, however, lead to non-trivial technical difficulties. In particular,

$$\begin{aligned} np^{-1/2}(\hat{p} - p) &= p^{-1/2}\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e} + (1-p)(\tfrac{\hat{\lambda} - np}{np^{3/2}}) + O_{\mathbb{P}}(n^{-1/2}p^{-1}) \\ &= p^{-1/2}\boldsymbol{e}^\top(\mathbf{A} - \mathbf{E}[\mathbf{A}])\boldsymbol{e}\big(1 + \tfrac{1-p}{np}\big) + O_{\mathbb{P}}(n^{-1/2}p^{-1}) \end{aligned}$$

upon iterating the term $(\hat{\lambda} - np)$. To guarantee that $O_{\mathbb{P}}(n^{-1/2}p^{-1})$ vanishes in the above expression, it might be necessary to require $np = \omega(\sqrt{n})$. That is to say, the expansions for $\boldsymbol{e}^\top\hat{\boldsymbol{u}}$ and $\hat{\lambda} - np$ in Eq. (3.6) and Eq. (3.7) is not sufficiently refined.

We surmise that to extend Theorem 4, even in the context of Erdős-Rényi graphs, to the setting wherein $np = o(\sqrt{n})$, it is necessary to consider higher order expansion for $\boldsymbol{e}^\top\hat{\boldsymbol{u}}$ and $\hat{\lambda} - np$. But this necessitates evaluating $\boldsymbol{e}^\top\mathbb{E}[(\mathbf{A} - \mathbb{E}[\mathbf{A}])^k]\boldsymbol{e}$ for $k \geq 3$, a highly non-trivial task; in particular $np = \omega(\log n)$ potentially require evaluating $\mathbb{E}[\boldsymbol{e}^\top(\mathbf{A} - \mathbb{E}[\mathbf{A}])^k\boldsymbol{e}]$ for $k = O(\log n)$. See also the conditions in Theorem 1 and Corollary 1 of [20]. In a slightly related vein, [6] evaluates $\mathrm{tr}[\mathbb{E}[(\mathbf{A} - \mathbb{E}[\mathbf{A}])^k$ in the case of Erdős-Rényi graphs and two-blocks planted partition SBM graphs.

In conclusion, *exact recovery* of $\boldsymbol{\tau}$ via $\hat{\boldsymbol{\tau}}$ is *not* sufficient to guarantee control of $\hat{\mathbf{B}}_{k\ell}^{(S)} - \mathbf{B}_{k\ell} = \boldsymbol{s}_k^\top(\hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top - \mathbb{E}[\mathbf{A}])\boldsymbol{s}_\ell$. In essence, as $\rho_n \to 0$, the bias incurred by the low-rank approximation $\hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top$ of $\mathbf{A}$ overwhelms the reduction in variance resulting from the low-rank approximation.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that considerably improved the quality of this paper.

Funding

The authors were supported in part by Johns Hopkins University Human Language Technology Center of Excellence and the XDATA and D3M programs of the Defense Advanced Research Projects Agency as administered through contract FA8750-12-2-0303 and contract FA8750-17-2-0112.

Supplementary Material

Proofs of stated results (DOI: 10.3150/21-BEJ1376SUPP; .pdf). The supplementary document consists of three sections. Section A provides a proof of Theorem 2. Section B provides the proofs for Theorem 3, Theorem 4, and Lemma 1. Derivations of the covariance terms in Theorem 3 and Theorem 4 are given in Section C.

References

[1] Abbe, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

[2] Abbe, E., Bandeira, A.S. and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[3] Abbe, E., Fan, J., Wang, K. and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[4] Airoldi, E.M., Blei, D.M., Fienberg, S.E. and Xing, E.P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[5] Athreya, A., Priebe, C.E., Tang, M., Lyzinski, V., Marchette, D.J. and Sussman, D.L. (2016). A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* **78** 1–18. MR3494576 https://doi.org/10.1007/s13171-015-0071-x

[6] Banerjee, D. and Ma, Z. (2017). Optimal hypothesis testing for stochastic blockmodels with growing degrees. Preprint. Available at https://arxiv.org/abs/1705.05305.

[7] Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

[8] Bickel, P.J. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–73.

[9] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[10] Bickel, P.J. and Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. MR3453655 https://doi.org/10.1111/rssb.12117

[11] Cape, J., Tang, M. and Priebe, C.E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Ann. Statist.* **47** 2405–2439. MR3988761 https://doi.org/10.1214/18-AOS1752

[12] Celisse, A., Daudin, J.-J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467 https://doi.org/10.1214/12-EJS729

[13] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

[14] Choi, D.S., Wolfe, P.J. and Airoldi, E.M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253 https://doi.org/10.1093/biomet/asr053

[15] Coja-Oghlan, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622 https://doi.org/10.1017/S0963548309990514

[16] DasGupta, A. and Johnstone, I.M. (2014). Asymptotic risk and Bayes risk of thresholding and superefficient estimates and optimal thresholding. In *Contemporary Developments in Statistical Theory. Springer Proc. Math. Stat.* **68** 41–67. Cham: Springer. MR3149915 https://doi.org/10.1007/978-3-319-02651-0_4

[17] Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817 https://doi.org/10.1007/s11222-007-9046-7

[18] Eldridge, J., Belkin, M. and Wang, Y. (2018). Unperturbed: Spectral analysis beyond Davis-Kahan. In *Proceedings of Algorithmic Learning Theory* 321–358.

[19] Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2013). Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.* **41** 2279–2375. MR3098073 https://doi.org/10.1214/11-AOP734

[20] Fan, J., Fan, Y., Han, X. and Lv, J. (2021). Asymptotic theory of eigenvectors for large random matrices. *J. Amer. Statist. Assoc.*

[21] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[22] Fishkind, D.E., Sussman, D.L., Tang, M., Vogelstein, J.T. and Priebe, C.E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990 https://doi.org/10.1137/120875600

[23] Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* **486** 75–174. MR2580414 https://doi.org/10.1016/j.physrep.2009.11.002

[24] Füredi, Z. and Komlós, J. (1981). The eigenvalues of random symmetric matrices. *Combinatorica* **1** 233–241. MR0637828 https://doi.org/10.1007/BF02579329

[25] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[26] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62** 2788–2797. MR3493879 https://doi.org/10.1109/TIT.2016.2546280

[27] Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 175–194. MR0400513

[28] Hoff, P.D., Raftery, A.E. and Handcock, M.S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

[29] Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[30] Joseph, A. and Yu, B. (2016). Impact of regularization on spectral clustering. *Ann. Statist.* **44** 1765–1791. MR3519940 https://doi.org/10.1214/16-AOS1447

[31] Karrer, B. and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[32] Klopp, O., Tsybakov, A.B. and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454

[33] Leeb, H. and Pötscher, B.M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142** 201–211. MR2394290 https://doi.org/10.1016/j.jeconom.2007.05.017

[34] Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370

[35] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[36] Lovász, L. (2012). *Large Networks and Graph Limits. American Mathematical Society Colloquium Publications* **60**. Providence, RI: Amer. Math. Soc. MR3012035 https://doi.org/10.1090/coll/060

[37] Lyzinski, V., Sussman, D.L., Tang, M., Athreya, A. and Priebe, C.E. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.* **8** 2905–2922. MR3299126 https://doi.org/10.1214/14-EJS978

[38] Lyzinski, V., Tang, M., Athreya, A., Park, Y. and Priebe, C.E. (2017). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Trans. Netw. Sci. Eng.* **4** 13–26. MR3625952 https://doi.org/10.1109/TNSE.2016.2634322

[39] Magnus, J.R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381–394. MR0520247

[40] Magnus, J.R. and Neudecker, H. (1980). The elimination matrix: Some lemmas and applications. *SIAM J. Algebr. Discrete Methods* **1** 422–449. MR0593853 https://doi.org/10.1137/0601049

[41] Mao, X., Sarkar, P. and Chakrabarti, D. (2021). Estimating mixed memberships with sharp eigenvector deviations. *J. Amer. Statist. Assoc.*

[42] McSherry, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. Los Alamitos, CA: IEEE Computer Soc. MR1948742

[43] Newman, M.E.J.and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 1–15.

[44] Oliveira, R.I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. http://arxiv.org/abs/0911.0600.

[45] Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *Proceedings of the* 20*th International Conference on Computer and Information Sciences* 284–293.

[46] Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[47] Rosvall, M. and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**.

[48] Rubin-Delanchy, P., Cape, J., Tang, M. and Priebe, C.E. (2017). A statistical interpretation of spectral embedding: The generalised random dot product graph. Preprint. Available at https://arxiv.org/abs/1709.05506.

[49] Sarkar, P. and Bickel, P.J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.* **43** 962–990. MR3346694 https://doi.org/10.1214/14-AOS1285

[50] Snijders, T.A.B. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742 https://doi.org/10.1007/s003579900004

[51] Sussman, D.L., Tang, M., Fishkind, D.E. and Priebe, C.E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899 https://doi.org/10.1080/01621459.2012.699795

[52] Tang, M., Cape, J. and Priebe, C.E. (2022). Supplement to "Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral estimator." https://doi.org/10.3150/21-BEJ1376SUPP

[53] Tang, M. and Priebe, C.E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist.* **46** 2360–2415. MR3845021 https://doi.org/10.1214/17-AOS1623

[54] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

[55] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z

[56] Vu, T., Chunikhina, E. and Raich, R. (2021). Perturbation expansions and error bounds for the truncated singular value decomposition. *Linear Algebra Appl.* **627** 94–139. MR4275028 https://doi.org/10.1016/j.laa.2021.05.020

[57] Wang, Y.X.R. and Bickel, P.J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528. MR3650391 https://doi.org/10.1214/16-AOS1457

[58] Wolfe, D.A.and Olhede, S.C. (2013). Nonparametric graphon estimation. Preprint. Available at http://arxiv.org/abs/1309/5936.

[59] Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *Proceedings of the* 35*th International Conference on Machine Learning* 5433–5442.

[60] Young, S.J. and Scheinerman, E.R. (2007). Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph*. *Lecture Notes in Computer Science* **4863** 138–149. Berlin: Springer. MR2504912 https://doi.org/10.1007/978-3-540-77004-6_11

Received July 2020 and revised May 2021