Biometrika (2020), **107**, 4, *pp.* 875–889
Printed in Great Britain
doi: 10.1093/biomet/asaa031
Advance Access publication 6 July 2020

Optimal Bayesian estimation for random dot product graphs

BY FANGZHENG XIE AND YANXUN XU

Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, U.S.A.
fxie5@jhu.edu yanxun.xu@jhu.edu

SUMMARY

We propose and prove the optimality of a Bayesian approach for estimating the latent positions in random dot product graphs, which we call posterior spectral embedding. Unlike classical spectral-based adjacency, or Laplacian spectral embedding, posterior spectral embedding is a fully likelihood-based graph estimation method that takes advantage of the Bernoulli likelihood information of the observed adjacency matrix. We develop a minimax lower bound for estimating the latent positions, and show that posterior spectral embedding achieves this lower bound in the following two senses: it both results in a minimax-optimal posterior contraction rate and yields a point estimator achieving the minimax risk asymptotically. The convergence results are subsequently applied to clustering in stochastic block models with positive semidefinite block probability matrices, strengthening an existing result concerning the number of misclustered vertices. We also study a spectral-based Gaussian spectral embedding as a natural Bayesian analogue of adjacency spectral embedding, but the resulting posterior contraction rate is suboptimal by an extra logarithmic factor. The practical performance of the proposed methodology is illustrated through extensive synthetic examples and the analysis of Wikipedia graph data.

Some key words: Likelihood-based graph estimation; Minimax optimality; Positive semidefinite stochastic block model; Posterior spectral embedding.

1. INTRODUCTION

Using graphs as a data structure to represent network data, with the vertices denoting entities and the edges encoding relationships between vertices, has become increasingly important in a broad range of applications, including social networks (Young & Scheinerman, 2007), brain imaging (Priebe et al., 2017) and neuroscience (Lyzinski et al., 2017; Tang et al., 2018). For example, in a Facebook network vertices represent users, and the occurrence of an edge linking any two users indicates that they are friends on Facebook. When one collects random graph data, it may be costly or even infeasible to collect individual-specific attributes that are heterogeneous across individuals, while only the adjacency matrix of the graph is accessible. For example, in studying the structure of a Wikipedia page network, collecting the hyperlinks between articles is much more feasible than collecting the attributes associated with each individual article. To model the unobserved vertex-specific attributes that result in the observed network, Hoff et al. (2002) proposed latent positions graphs, in which each vertex is associated with an unobserved Euclidean vector called the latent position, and the edge probability between any two vertices only depends on their latent positions. Formally, each vertex i is associated with a vector x_i in some latent space $\mathcal{X}$, and there exists a symmetric function $\kappa : \mathcal{X} \times \mathcal{X} \to [0, 1]$, called a graphon (Lovász, 2012), such that an edge between vertices i and j occurs with probability $\kappa(x_i, x_j)$, and

© 2020 Biometrika Trust

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

the occurrences of these edges are independent given the latent positions. There is a vast literature addressing statistical inference on latent positions graphs; see Bickel & Chen (2009), Fortunato (2010), Goldenberg et al. (2010), Bickel et al. (2011) and Choi et al. (2012), among others.

In this paper we focus on a specific example of latent positions graphs: the random dot product graph model (Young & Scheinerman, 2007), in which the graphon function κ is simply the dot product of two latent positions: $\kappa(x_i, x_j) = x_i^{\mathrm{T}} x_j$. The random dot product graph model enjoys several nice properties. First, the well-known stochastic block model, in which the vertices are grouped into several blocks, is a special case of the random dot product graph model and can be represented with the latent positions of vertices in the same block being identical, provided that the block probability matrix is positive semidefinite. Second, the architecture of the random dot product graph is simple, as the expected value of the adjacency matrix is a symmetric low-rank matrix, motivating the use of a wide range of tractable spectral-based techniques for statistical analysis. Furthermore, the random dot product graph can provide accurate approximation to more general latent positions graphs when the dimension of the latent positions grows with the number of vertices at a certain rate (Tang et al., 2013). For a thorough review of recent advances in statistical inference on the random dot product graph model, readers are referred to Athreya et al. (2018).

The techniques for statistical analysis of the random dot product graph model have so far focused on spectral methods based on the observed adjacency matrix or its graph Laplacian matrix. For example, Sussman et al. (2014) proposed directly estimating the latent positions using adjacency spectral embedding, and proved its consistency. For the normalized graph Laplacian matrix of the adjacency matrix, Tang & Priebe (2018) found the asymptotic distribution of spectral embedding using the normalized graph Laplacian, and made a thorough comparison between adjacency spectral embedding and Laplacian spectral embedding under various contexts. The well-developed theory for spectral methods for the random dot product graph model lays a theoretical foundation for a variety of subsequent inference tasks, including spectral clustering for stochastic block models (Sussman et al., 2012; Lyzinski et al., 2014, 2017), vertex classification and nomination (Sussman et al., 2014; Lyzinski et al. 2017, 2018), nonparametric graph hypothesis testing (Tang et al., 2017a) and multiple graph inference (Tang et al., 2017b; Levin et al., 2019; Wang et al., 2019).

Despite the marvellous success of spectral methods for the random dot product graph model, it remains an open question whether these spectral estimators are minimax optimal for estimating the latent positions with respect to suitable loss functions. Taking one step back, a more fundamental question is: what is the minimax risk for estimating the latent positions, and how can one achieve it by constructing a useful estimator? In this paper we provide a detailed answer to this question. Unlike the aforementioned spectral-based approaches, we take advantage of the Bernoulli likelihood information of the observed graph adjacency matrix and design a fully likelihood-based Bayesian approach, referred to as posterior spectral embedding. Not only do we establish a minimax lower bound for estimating the latent positions, but we also show that this lower bound is achievable through the proposed Bayes procedure. Specifically, we show that posterior spectral embedding both yields the rate-optimal contraction and produces a minimax-optimal point estimator for estimating the latent positions. To the best of our knowledge, our work represents the first effort in the literature of the random dot product graph model that leverages a likelihood-based Bayesian approach with theoretical guarantee. In addition, as a sample application we improve an existing result regarding clustering in positive semidefinite stochastic block models by showing that the number of misclustered vertices can be reduced from $O(\log n)$ (Sussman et al., 2012) to $O(1)$, using the proposed posterior spectral embedding method.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

There are several results related to our method in the literature. Strong consistency for clustering in stochastic block models was achieved by Bickel & Chen (2009) and Zhao et al. (2012), but their methods are not applicable to more general random dot product graph models. In addition, their approaches are frequentist methods, whereas we develop a Bayes procedure and establish theoretical properties of the resulting full posterior distribution. A Bayesian methodology for clustering stochastic block models was used by van der Pas & van der Vaart (2018), but the consistency result was with regard to the maximum a posteriori estimator, which can be treated as a frequentist point estimator as well. The strong consistency of the full posterior distribution for clustering in stochastic block models was discussed by Zhuo & Gao (2018), but under the assumption that the stochastic block models were homogeneous. In contrast, our work includes positive semidefinite stochastic block models, and is more flexible from the perspective of the number of free parameters.

The following notation and symbols will be used in the rest of this paper. The $d \times d$ identity matrix is denoted by I_d. For an integer p, $1 \leqslant p \leqslant \infty$, and a d-dimensional Euclidean vector $x = (x_1, \ldots, x_d)^{\mathrm{T}}$, we use $\|x\|_p$ to denote its ℓ_p-norm, and when $p = \infty$, $\|x\|_\infty = \max_{k=1,\ldots,d} |x_k|$. For a vector $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$, the vector inequality $x \geqslant 0$ represents $x_k \geqslant 0$ for $k = 1, \ldots, p$. For an $n \times d$ matrix X we use $(X)_{*k}$ to denote the n-dimensional vector formed by the kth column of X. For a positive integer n, we denote by $[n]$ the set of integers $[n] = \{1, 2, \ldots, n\}$. For any two positive integers n, d with $n \geqslant d$, $\mathbb{O}(n, d)$ denotes the set of all orthogonal d-frames in $\mathbb{R}^d$, i.e., $\mathbb{O}(n, d) = \{U \in \mathbb{R}^{n \times d} : U^{\mathrm{T}} U = I_d\}$, and when $n = d$, we use the notation $\mathbb{O}(d) = \mathbb{O}(d, d)$. The symbols $\lesssim$ and $\gtrsim$ mean the corresponding inequality up to a constant, i.e., $a \lesssim b$ or $a \gtrsim b$ if $a \leqslant Cb$ or $a \geqslant Cb$ for some constant $C > 0$. We write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. For a $d \times d$ positive definite matrix Δ, we use $\lambda_k(\Delta)$ to denote its kth largest eigenvalue, and for any rectangular matrix X, we use $\sigma_k(X)$ to denote its kth largest singular value. We say that a sequence of events $(E_n)_{n=1}^{\infty}$ occurs almost always if $\mathrm{pr}(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} E_k) = 1$.

2. Preliminaries

We first give some background information on the random dot product graph model. Let the space of d-dimensional latent positions be $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leqslant 1, x \geqslant 0\}$, where $\|\cdot\|_2$ is the ℓ_2-norm of a Euclidean vector. Let $X = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^{n \times d}$ be an $n \times d$ matrix, where $x_1, \ldots, x_n \in \mathcal{X}$ represent the latent positions of n vertices in a graph. A symmetric random binary matrix $Y = (y_{ij})_{n \times n} \in \{0, 1\}^{n \times n}$ is said to be the adjacency matrix of a random dot product graph with latent position matrix X, denoted by $Y \sim \mathrm{RDPG}(X)$, if the random variables $y_{ij} \sim \mathrm{Ber}(x_i^{\mathrm{T}} x_j)$ independently, $1 \leqslant i \leqslant j \leqslant n$. Namely, $p(Y \mid X) = \prod_{i \leqslant j} (x_i^{\mathrm{T}} x_j)^{y_{ij}} (1 - x_i^{\mathrm{T}} x_j)^{1 - y_{ij}}$.

Example 1 (Positive semidefinite stochastic block model). The most popular example of the random dot product graph model is the stochastic block model with a positive semidefinite block probability matrix. Formally, given K with $K/n \to 0$, a symmetric random adjacency matrix Y is drawn from a K-block stochastic block model with a symmetric block probability matrix $B = (b_{kl})_{K \times K} \in (0, 1)^{K \times K}$ and a block assignment function $\tau : [n] \to [K]$, denoted by $Y \sim \mathrm{SBM}(B, \tau)$, if the random variables $y_{ij} \sim \mathrm{Ber}(b_{\tau(i)\tau(j)})$ independently for $1 \leqslant i \leqslant j \leqslant n$. Namely, vertices in the same block have the same connecting probability. When B is positive semidefinite with rank d, we refer to the model as a positive semidefinite stochastic block model, and there exists a matrix $L \in \mathbb{R}^{K \times d}$ such that $B = LL^{\mathrm{T}}$. By converting the block assignment function τ into an $n \times K$ matrix $Z = [\mathbb{1}\{\tau(i) = k\}]_{i \in [n], k \in [K]}$ we obtain $E_X(Y) = (ZL)(ZL)^{\mathrm{T}}$, and therefore $\mathrm{SBM}(B, \tau)$ coincides with $\mathrm{RDPG}(X)$ through the reparametrization $X = ZL$. The positive semidefinite stochastic block model will be revisited in § 4.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Remark 1. Not all stochastic block models can be represented by the random dot product graph model. Consider the following example: $Y \sim \text{SBM}(B, \tau)$ with $\tau(1) = 1, \tau(2) = \cdots = \tau(n) = 2$, where $B = (b_{k\ell})_{2\times 2}$ is indefinite, indicating that there exists some $u = (u_1, u_2)^{\mathrm{T}} \in \mathbb{R}^2$ such that $u^{\mathrm{T}}Bu < 0$. Take $v = \{u_1, u_2/(n-1), \ldots, u_2/(n-1)\}^{\mathrm{T}} \in \mathbb{R}^n$, and denote $Z = [\mathbb{1}\{\tau(i) = k\}]_{i\in[n],k\in[K]}$. It follows that $Z^{\mathrm{T}}v = u$, and hence $v^{\mathrm{T}}E(Y)v = (Z^{\mathrm{T}}v)^{\mathrm{T}}B(Z^{\mathrm{T}}v) = u^{\mathrm{T}}Bu < 0$. Since $E(Y)$ is not positive semidefinite, $\text{SBM}(B, \tau)$ cannot be represented by $\text{RDPG}(X)$ for some $X \in \mathbb{R}^{n\times 2}$.

Example 2 (Hardy–Weinberg curve example). We provide an example of the random dot product graph model that is not a stochastic block model. Let $d = 3$ and $C : (0, 1) \to \mathcal{X}^3$ be the Hardy–Weinberg curve (Athreya et al., 2020) defined by $C(t) = (t^2, 1-2t+t^2, 2t-2t^2)^{\mathrm{T}} \in \mathbb{R}^3$. Let $(t_i)_{i=1}^n$ be distinct points taking values in $(0, 1)$, and $x_i = C(t_i)$ for all $i \in [n]$. Define the latent position matrix X by $X = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^{n\times 3}$, and let $Y \sim \text{RDPG}(X)$. Then the random dot product graph model generated according to this Hardy–Weinberg curve does not fall into the category of stochastic block models.

Remark 2 (Intrinsic nonidentifiability). The latent position matrix X cannot be uniquely determined by the distribution $Y \sim \text{RDPG}(X)$, i.e., X is not identifiable. In fact, for any orthogonal matrix $W \in \mathbb{R}^{d\times d}$, the two distributions $\text{RDPG}(X)$ and $\text{RDPG}(XW)$ are identical, since for any $i, j \in [n]$, $x_i^{\mathrm{T}}x_j = (Wx_i)^{\mathrm{T}}(Wx_j)$. In addition, any d-dimensional random dot product graph model can be embedded into a d'-dimensional random dot product graph model for any $d' > d$, in the sense that there exists a d'-dimensional latent position matrix $X' \in \mathbb{R}^{n\times d'}$ such that the two distributions $\text{RDPG}(X)$ and $\text{RDPG}(X')$ are identical. The latter source of nonidentifiability, however, can be eliminated by requiring the columns of X to be linearly independent.

Remark 3 (Choice of orthogonal transformation and loss function). Since the latent position matrix X can only be identified up to an orthogonal transformation, one needs to properly rotate any estimator $\hat{X}$ to align with the underlying true X. The alignment matrix can be found by the solution to the orthogonal Procrustes problem $W^* = \arg\inf_W \|\hat{X}W - X\|_{\mathrm{F}}$, where the infimum ranges over the set of all orthogonal matrices in $\mathbb{R}^{d\times d}$ (Athreya et al., 2020). In particular, W^* has a closed-form expression. Consequently, in this work we consider the loss function

$$L_{\mathrm{F}}(\hat{X}, X) = \frac{1}{n}\inf_{W\in\mathbb{O}(d)} \|\hat{X} - XW\|_{\mathrm{F}}^2 = \inf_{W\in\mathbb{O}(d)} \frac{1}{n}\sum_{i=1}^{n} \|\hat{x}_i - W^{\mathrm{T}}x_i\|_2^2,$$

where $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$. This loss function can also be interpreted as the average error of the estimated latent positions $\hat{x}_1, \ldots, \hat{x}_n$ of all n vertices after the appropriate orthogonal alignment.

The adjacency matrix Y can be viewed as the sum of a low-rank signal matrix XX^{T} and a noise matrix $E = (e_{ij})_{n\times n}$, the elements of which are centred Bernoulli random variables $e_{ij} \sim \text{Ber}(x_i^{\mathrm{T}}x_j) - x_i^{\mathrm{T}}x_j$ independently for $1 \leqslant i \leqslant j \leqslant n$. Sussman et al. (2014) argued for embedding the adjacency matrix Y into $\mathbb{R}^{n\times d}$ by solving the least-squared problem $\hat{X} = \arg\min_{X\in\mathbb{R}^{n\times d}} \|Y - XX^{\mathrm{T}}\|_{\mathrm{F}}^2$. The resulting estimator $\hat{X}$ is referred to as the adjacency spectral embedding of Y (Sussman et al., 2012) and is denoted by $\hat{X}_{\text{ASE}}$. Theoretical properties of adjacency spectral embedding have been explored by Sussman et al. (2012) and Lyzinski et al. (2014, 2017). Notably, the following convergence result of $\hat{X}_{\text{ASE}}$ was established by Sussman et al. (2014).

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

THEOREM 1 (Sussman et al., 2014). *Suppose* $Y \sim \text{RDPG}(X)$ *for some* $X \in \mathbb{R}^{n\times d}$, *and* $(1/n)X^{\mathrm{T}}X \to \Delta$ *for some positive definite* $\Delta \in \mathbb{R}^{d\times d}$ *with distinct eigenvalues* $\lambda_1(\Delta) > \cdots > \lambda_d(\Delta) > 0$ *as* $n \to \infty$. *Assume that there exists* $\delta > 0$ *such that* $\min_{j\neq k} |\lambda_j(\Delta) - \lambda_k(\Delta)| > 2\delta$ *and* $\lambda_d(\Delta) > 2\delta$. *Then, with probability greater than* $1 - 2(d^2+1)/n^2$,

$$\frac{1}{n}\inf_{W\in\mathbb{O}(d)} \|\hat{X}_{\text{ASE}} - XW\|_{\text{F}}^2 \leqslant \frac{12d^2\log n}{\delta^3 n}. \tag{1}$$

Theorem 1 implies that after an orthogonal alignment of $\hat{X}_{\text{ASE}}$ towards X, adjacency spectral embedding yields a convergence rate $L_{\text{F}}(\hat{X}_{\text{ASE}}, X) = o_{\text{pr}}\{(M_n \log n)/n\}$ for arbitrary $M_n \to \infty$, where $(M_n)_{n=1}^{\infty}$ should be interpreted as a sequence converging to ∞ arbitrarily slowly. Nevertheless, as will be seen in § 3, this rate is suboptimal and, interestingly, can be improved by a Bayes estimator instead. Furthermore, it is unclear what the minimax risk for estimating the latent position matrix X with respect to the loss $L_{\text{F}}(\cdot,\cdot)$ is, or how to construct an estimator to achieve the minimax rate, which we will address in this paper. The distinct eigenvalues condition will also be relaxed in § 3. We begin approaching our main goal by first establishing the following minimax lower bound.

THEOREM 2. *Let* $Y \sim \text{RDPG}(X)$ *for some* $X = (x_1,\ldots,x_n)^{\mathrm{T}}$, $x_1,\ldots,x_n \in \mathcal{X}$. *Assume that* d *is fixed and does not change with* n. *Let* $\hat{X}$ *be an estimator of the latent position matrix* X *satisfying* $\|\hat{X}\|_{\text{F}} \lesssim n^{1/2}$ *with probability* 1. *Then*

$$\inf_{\hat{X}}\sup_{X\in\mathcal{X}^n} E_X\left(\frac{1}{n}\inf_{W\in\mathbb{O}(d)}\|\hat{X} - XW\|_{\text{F}}^2\right) \gtrsim \frac{1}{n}. \tag{2}$$

The above minimax lower bound does not necessarily result in a minimax rate of convergence for estimating the latent positions. Nevertheless, if we assume the existence of an estimator $\hat{X}$ with $E_X\{(1/n)\inf_W \|\hat{X} - XW\|_{\text{F}}^2\} \lesssim 1/n$, which will be rigorously proved in § 3, then simply applying Markov's inequality yields $(1/n)\inf_W \|\hat{X} - XW\|_{\text{F}}^2 = o_{\text{pr}}(M_n/n)$ for an arbitrary sequence $M_n \to \infty$. This observation suggests that the convergence rate derived in Sussman et al. (2014) for the adjacency spectral embedding might be suboptimal and motivates us to pursue an estimator achieving the minimax lower bound (2).

3. LIKELIHOOD-BASED POSTERIOR SPECTRAL EMBEDDING

Although it is intuitive and computationally convenient to directly estimate the latent position matrix X by the popular spectral-based approaches, i.e., adjacency spectral embedding, the Bernoulli likelihood information of the adjacency matrix is neglected. On the other hand, likelihood-based methods for the random dot product graph model remain underexplored. In particular, neither the existence nor the uniqueness of the maximum likelihood estimator for X has been addressed. In this section we develop a Bayesian approach for estimating the latent positions by taking advantage of the Bernoulli likelihood information.

Recall that the space of latent positions is $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leqslant 1, x \geqslant 0\}$. Let $X_0 = (x_{01},\ldots,x_{0n})^{\mathrm{T}}$ be the true latent position matrix, and $X = (x_1,\ldots,x_n)^{\mathrm{T}}$ be the latent position matrix to be assigned a prior distribution Π. Whenever we consider the distribution Π, X is treated as a random matrix taking values in the space $\mathcal{X}^n = \{X = (x_1,\ldots,x_n)^{\mathrm{T}} : x_i \in \mathcal{X}, i = 1,\ldots,n\}$. The prior distribution Π on X is constructed by assuming that $x_1,\ldots,x_n$ follow a distribution with a density function π_x supported on $\mathcal{X}$ independently, and we denote it by $X \sim \Pi$. In this

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

work we only require π_x to be bounded away from 0 and ∞ over $\mathcal{X}$, e.g., the uniform distribution on $\mathcal{X}$. It follows directly from the Bayes formula that the posterior distribution of X is

$$\Pi(X \in \mathcal{A} \mid Y) = \frac{N_n(\mathcal{A})}{D_n}, \quad N_n(\mathcal{A}) = \int_{\mathcal{A}} \prod_{i \leqslant j} \frac{p(y_{ij} \mid X)}{p(y_{ij} \mid X_0)} \Pi(\mathrm{d}X), \quad D_n = N_n(\mathcal{X}),$$

and $p(y_{ij} \mid X) = (x_i^{\mathrm{T}} x_j)^{y_{ij}} (1 - x_i^{\mathrm{T}} x_j)^{1-y_{ij}}$, for any measurable set $\mathcal{A} \subset \mathcal{X}^n$. Clearly, the posterior distribution of X incorporates the Bernoulli likelihood information through the Bayes formula, and we refer to $\Pi(X \in \cdot \mid Y)$ as posterior spectral embedding.

The following theorem, which is the key result of this work, shows that under mild regularity conditions, the posterior contraction of the latent positions is minimax optimal. The proof is deferred to the Supplementary Material.

THEOREM 3. *Let $Y \sim \mathrm{RDPG}(X_0)$ for some $X_0 = (x_{01}, \ldots, x_{0n})^{\mathrm{T}} \in \mathbb{R}^{n \times d}$, and the prior Π be as described above. Assume that $(1/n)(X_0^{\mathrm{T}} X_0) \to \Delta$ as $n \to \infty$ for some positive definite $\Delta \in \mathbb{R}^{d \times d}$. If d is fixed, and $\delta \leqslant \min_{i,j} x_{0i}^{\mathrm{T}} x_{0j} \leqslant \max_{i,j} x_{0i}^{\mathrm{T}} x_{0j} \leqslant 1 - \delta$ for some constant $\delta \in (0, 1/2)$ independent of n, then there exist some large constants $M_1, M_2 > 0$, depending on Δ and the prior π_x, such that*

$$E_0 \left\{ \Pi \left(\frac{1}{n} \|XX^{\mathrm{T}} - X_0 X_0^{\mathrm{T}}\|_{\mathrm{F}} > \frac{M_1}{\sqrt{n}} \;\middle|\; Y \right) \right\} \leqslant 8 \exp\left(-\frac{1}{2} nd \right),$$

$$E_0 \left\{ \Pi \left(\frac{1}{n} \inf_{W \in \mathbb{O}(d)} \|X - X_0 W\|_{\mathrm{F}}^2 > \frac{M_2}{n} \;\middle|\; Y \right) \right\} \leqslant 8 \exp\left(-\frac{1}{2} nd \right)$$

for sufficiently large n.

Remark 4. The assumption $(1/n)(X_0^{\mathrm{T}} X_0) \to \Delta$ as $n \to \infty$ in Theorem 3 can be equivalently written as $(1/n) \sum_{i=1}^n x_{0i} x_{0i}^{\mathrm{T}} \to \Delta$ as $n \to \infty$ for some positive definite Δ. An intuitive interpretation of this condition is that the true latent positions $x_{01}, \ldots, x_{0n}$ can be regarded as random samples drawn from some nondegenerate distribution with a positive definite second-moment matrix Δ. By the law of large numbers, the sample version of the second-moment matrix converges to the population version of the second-moment matrix. An illustrative example is the positive semidefinite stochastic block model. Suppose the distinct latent positions of $x_{01}, \ldots, x_{0n}$ are $x_{01}^*, \ldots, x_{0K}^*$, and let $n_k = \sum_{i=1}^n \mathbb{1}(x_{0i} = x_{0k}^*)$ be the number of vertices corresponding to the latent position x_{0k}^*. Assume that K is fixed, $n_k/n \to \alpha_k > 0$ as $n \to \infty$, and α_ks, x_{0k}^*s are fixed for $k = 1, \ldots, K$. Then

$$\frac{1}{n} X_0^{\mathrm{T}} X_0 = \sum_{k=1}^K \sum_{i=1}^n \mathbb{1}(x_{0i} = x_{0k}^*) x_{0i} x_{0i}^{\mathrm{T}} = \sum_{k=1}^K \frac{n_k}{n} (x_{0k}^*)(x_{0k}^*)^{\mathrm{T}} \to \sum_{k=1}^K \alpha_k (x_{0k}^*)(x_{0k}^*)^{\mathrm{T}}$$

as $n \to \infty$. Therefore, with the above assumption, the positive semidefinite stochastic block model satisfies this condition provided that $\sum_{k=1}^K \alpha_k (x_{0k}^*)(x_{0k}^*)^{\mathrm{T}}$ is positive definite.

Theorem 3 claims that, under appropriate regularity conditions, posterior spectral embedding yields a rate-optimal posterior contraction for the latent positions in the Bayesian sense. The following theorem shows that one can use posterior spectral embedding to construct a point estimator $\hat{X}$ that exactly achieves the minimax lower bound (2).

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

THEOREM 4. *Let the conditions in Theorem 3 hold, and let constant $M_1 > 0$ be given by Theorem 3. Consider the posterior mean of the edge probability matrix,*

$$\tilde{P} = \int_{X \in \mathcal{X}^n} XX^{\mathrm{T}} \Pi(\mathrm{d}X \mid Y).$$

Suppose $\tilde{P}$ yields the spectral decomposition $\tilde{P} = \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j$, where $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ are eigenvalues of $\tilde{P}$ arranged in nonincreasing order, and $\hat{u}_1, \ldots, \hat{u}_n$ are the associated eigenvectors. Let $\hat{U} = (\hat{u}_1, \ldots, \hat{u}_d)$, $\hat{S} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$, $\hat{X} = \hat{U}\hat{S}^{1/2}$ and U_0 be the left-singular vector matrix of X_0. Then, for sufficiently large n,

$$E_0 \left(\frac{1}{n} \inf_{W \in \mathbb{O}(d)} \|\hat{X} - X_0 W\|_{\mathrm{F}}^2 \right) \lesssim \frac{1}{n}. \tag{3}$$

Furthermore, for sufficiently large n,

$$\mathrm{pr}_0 \left\{ \inf_{W \in \mathbb{O}(d)} \|\hat{U} - U_0 W\|_{\mathrm{F}}^2 > \frac{128 M_1^2 d}{\lambda_d^2(\Delta) n} \right\} \leqslant 2 \exp\left(-\frac{1}{4} M_1 d \sqrt{n} \right). \tag{4}$$

We briefly compare the results of Theorem 4 with those in Sussman et al. (2014). The convergence rate (3) shows that $\hat{X}$ not only achieves the minimax lower bound (2), but also yields a convergence rate $(1/n) \inf_W \|\hat{X} - X_0 W\|_{\mathrm{F}}^2 = o_{\mathrm{pr}_0}(M_n/n)$ for any $M_n \to \infty$, improving the rate (1) obtained in Sussman et al. (2014). The convergence rate of the unscaled eigenvectors $\hat{U}$ given by (4) also improves its counterpart in Sussman et al. (2014), which is explained as follows. Denote by U the left-singular vector matrix of X, and $\hat{U}_{\mathrm{ASE}}$ that of $\hat{X}_{\mathrm{ASE}}$. Then, under the assumptions of Theorem 1, there exists a diagonal matrix W, the diagonal entries of which are either 1 or -1, such that

$$\mathrm{pr}_0 \left\{ \|(\hat{U}_{\mathrm{ASE}})_{*k} - (WU_0)_{*k}\|_2^2 > \frac{3 \log n}{\delta^2 n} \right\} \leqslant \frac{2(d^2+1)}{n^2} \tag{5}$$

for $k = 1, \ldots, d$. In contrast, the eigenvector estimate $\hat{U}$ derived using posterior spectral embedding improves the convergence rate (5). Not only do we improve the rate from $(\log n)/n$ to $1/n$, but we also sharpen the large deviation probability from $O(1/n^2)$ to $O(\mathrm{e}^{-cn^{1/2}})$ for some constant $c > 0$. The distinct eigenvalues condition for Δ required in Sussman et al. (2014) is also relaxed.

4. APPLICATION: CLUSTERING IN POSITIVE SEMIDEFINITE STOCHASTIC BLOCK MODELS

This section presents an application of posterior spectral embedding to clustering in positive semidefinite stochastic block models. In particular, we show that the result obtained in this section strengthens an existing result related to the number of misclustered vertices. We first review the K-means clustering procedure in general (Lloyd, 1982). Suppose that n data points $\hat{x}_1, \ldots, \hat{x}_n$ in $\mathbb{R}^d$ are to be assigned into K clusters, and denote $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times d}$ the corresponding data matrix. The K-means clustering centroids of $\hat{x}_1, \ldots, \hat{x}_n$, represented by an $n \times d$ matrix $C(\hat{X})$ with K distinct rows, are given by

$$C(\hat{X}) = \underset{C \in \mathcal{C}_K}{\arg\min} \|C - \hat{X}\|_{\mathrm{F}}, \quad \text{where} \quad \mathcal{C}_K = \{C \in \mathbb{R}^{n \times d} : C \text{ has } K \text{ distinct rows}\}.$$

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

The corresponding cluster assignment function is defined to be any function $\tau(\cdot;\hat{X}) : [n] \to [K]$ such that $\tau(i;\hat{X}) = \tau(j;\hat{X})$ if and only if $\{C(\hat{X})\}_{i*} = \{C(\hat{X})\}_{j*}$. Given two cluster assignment functions τ_1 and τ_2, the Hamming distance between τ_1 and τ_2 is defined by $d_{\mathrm{H}}(\tau_1,\tau_2) = \sum_{i=1}^{n} \mathbb{1}\{\tau_1(i) \neq \tau_2(i)\}$. To avoid the labelling issue, we use $\inf_{\sigma\in\mathcal{S}_K} d_{\mathrm{H}}\{\sigma \circ \tau(\cdot;X), \tau(\cdot;X_0)\}$ as the measurement for clustering performance, where $\mathcal{S}_K$ is the set of all permutations in $[K]$.

A clustering procedure for stochastic block models is called consistent if the resulting fraction of misclustered vertices is asymptotically zero. Consistent clustering procedures in stochastic block models have been investigated in earlier work, including likelihood-based methods (Choi et al., 2012), spectral clustering based on Laplacian spectral embedding (Rohe et al., 2011), K-means clustering based on adjacency spectral embedding (Sussman et al., 2012) and modularity maximization (Girvan & Newman, 2002), among others. In particular, Sussman et al. (2012) argue that by directly applying the K-means procedure to adjacency spectral embedding $\hat{X}_{\mathrm{ASE}}$, i.e., replacing the aforementioned $\hat{X}$ by $\hat{X}_{\mathrm{ASE}}$, the number of misclustered vertices can be upper bounded by $O(\log n)$. In what follows we show that this result can be strengthened by taking advantage of the $n^{1/2}$ convergence rate of posterior spectral embedding.

Our method for clustering is straightforward: similar to K-means clustering based on $\hat{X}_{\mathrm{ASE}}$, we directly apply the K-means clustering procedure to the posterior samples collected from posterior spectral embedding. Specifically, for each realization X drawn from posterior spectral embedding, we obtain a cluster assignment function $\tau(\cdot;X)$ by applying the aforementioned K-means clustering procedure to X. This results in a posterior distribution of the cluster assignment function $\Pi\{\tau(\cdot;X) \in \cdot \mid Y\}$, which is induced from the map $X \mapsto \tau(\cdot;X)$ and posterior spectral embedding $\Pi(X \in \cdot \mid Y)$. The theorem below shows that we can recover the clustering structure through the K-means procedure, even when we assume that the working model is the random dot product graph model, which is more general than the positive semidefinite stochastic block model.

THEOREM 5. *Assume the conditions in Theorem 3 hold, and let the constants $M_1, M_2 > 0$ be provided by Theorem 3. Further assume that $X_0 = (x_{01},\ldots,x_{0n})^{\mathrm{T}}$ has K distinct rows $x_{01}^*,\ldots,x_{0K}^*$ for some $K \in [n]$, they satisfy $\min_{k\neq k'} \|x_{0k}^* - x_{0k'}^*\|_2 > \xi$ for some $\xi > 0$, and $n_k := \sum_{i=1}^{n} \mathbb{1}(x_{0i} = x_{0k}^*) \to \infty$ as $n \to \infty$ for all $k \in [K]$. Then, for sufficiently large n,*

$$E_0\left[\Pi\left\{\inf_{\sigma\in\mathcal{S}_K} d_{\mathrm{H}}(\sigma\circ\tau_0,\tau_X) \geqslant \frac{16M_2^2}{\xi^2} \;\middle|\; Y\right\}\right] \leqslant 8\exp\left(-\frac{1}{2}nd\right),$$

where $\tau_0 = \tau(\cdot;X_0)$ and $\tau_X = \tau(\cdot;X)$. Let $\hat{U}$ be the left-singular vector matrix of $\hat{X}$ defined in Theorem 4, and U_0 be that of X_0. Then it almost always holds that

$$\inf_{\sigma\in\mathcal{S}_K} d_{\mathrm{H}}\{\sigma\circ\tau(\cdot;\hat{U}),\tau(\cdot;U_0)\} \leqslant \frac{16}{\xi^2}\left\{\frac{8M_1\sqrt{2d}}{\lambda_d(\Delta)}\right\}^2.$$

Remark 5. Sussman et al. (2012) directly applied the K-means clustering procedure to $\hat{X}_{\mathrm{ASE}}$, and showed that $\inf_{\sigma\in\mathcal{S}_K} d_{\mathrm{H}}\{\sigma\circ\tau(\cdot;\hat{X}_{\mathrm{ASE}}),\tau(\cdot;X_0)\} \lesssim \log n$ almost always. Namely, the number of vertices that are incorrectly clustered is $O(\log n)$ eventually. The result obtained in Theorem 5 is stronger, since it shows that this number can be further reduced to $O(1)$ in the following two senses: if the K-means clustering procedure is applied to the posterior samples drawn from posterior spectral embedding, then with posterior probability tending to 1 in pr_0-probability, the posterior number of misclustered vertices is upper bounded by a constant. If the K-means clustering procedure is directly applied to the unscaled left-singular vector $\hat{U}$ of the point estimator

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

$\hat{X}$ obtained in Theorem 4, then it almost always holds that this number can be upper bounded by a constant as well.

Remark 6. The rate $O(1)$ for the number of misclustered vertices is due to the convergence rate $E_0\{(1/n)\inf_W \|\hat{X} - X_0 W\|_{\mathrm{F}}^2\} \asymp 1/n$. This improvement is not only specific to positive semidefinite stochastic block models, but also accredited to the Bayesian approach, along with its specific proof strategy. The improvement is specific to positive semidefinite stochastic block models because the minimax lower bound provided in Theorem 2 is only valid in the context of random dot product graphs. It should also be accredited to the Bayesian approach with its corresponding proof strategy because, by doing so, we are able to achieve the desired minimax lower bound via Bayes estimates.

5. Spectral-based Gaussian spectral embedding

We have seen in § 3 and § 4 the advantages of posterior spectral embedding over adjacency spectral embedding for the random dot product graph model. The major difference is that posterior spectral embedding is a fully likelihood-based approach taking the Bernoulli likelihood information into account, while adjacency spectral embedding only leverages the low-rank structure of the expected value of the adjacency matrix $XX^{\mathrm{T}} = E_X(Y)$. Recall that adjacency spectral embedding $\hat{X}_{\mathrm{ASE}}$ is the solution to the minimization problem $\min_{X\in\mathbb{R}^{n\times d}} \|Y - XX^{\mathrm{T}}\|_{\mathrm{F}}^2$. Equivalently, we can also view $\hat{X}_{\mathrm{ASE}}$ as the maximum likelihood estimator of X using a Gaussian likelihood function,

$$\hat{X}_{\mathrm{ASE}} = \underset{X\in\mathbb{R}^{n\times d}}{\arg\min} \|Y - XX^{\mathrm{T}}\|_{\mathrm{F}}^2 = \underset{X\in\mathbb{R}^{n\times d}}{\arg\max} \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_{ij} - x_i^{\mathrm{T}}x_j)^2\right\}.$$

The above interpretation motivates us to study a Bayesian analogue of the adjacency spectral embedding, referred to as the Gaussian spectral embedding, introduced as follows.

Assume that Π_{G} is some prior distribution on the latent position matrix X supported on $\mathbb{R}^{n\times d}$. We consider the following pseudo-posterior distribution by taking the Gaussian density as the working model:

$$\begin{aligned}\Pi_{\mathrm{G}}(X\in\mathcal{A}\mid Y) &= \frac{N_n^{\mathrm{G}}(\mathcal{A})}{D_n^{\mathrm{G}}}, \quad N_n^{\mathrm{G}}(\mathcal{A}) = \int_{\mathcal{A}}\prod_{i,j\in[n]}\frac{\phi(y_{ij} - x_i^{\mathrm{T}}x_j)}{\phi(y_{ij} - x_{0i}^{\mathrm{T}}x_{0j})}\Pi_{\mathrm{G}}(\mathrm{d}X),\\ D_n^{\mathrm{G}} &= N_n(\mathbb{R}^{n\times d}),\end{aligned} \tag{6}$$

for any measurable set $\mathcal{A}\subset\mathbb{R}^{n\times d}$, where ϕ is the density function of $\mathrm{N}(0,1)$. The formulation of (6) is completely based on the spectral property of Y and $E_X(Y) = XX^{\mathrm{T}}$, and does not incorporate the Bernoulli likelihood information. We refer to the pseudo-posterior distribution (6) as the Gaussian spectral embedding of Y. Observe that when

$$\Pi_{\mathrm{G}}(\mathrm{d}X) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d \exp\left(-\frac{x_i^{\mathrm{T}}x_i}{2\sigma^2}\right)\mathrm{d}x_i \tag{7}$$

for some $\sigma^2 > 0$, the maximum a posteriori estimator of (6) is the same as the solution to the minimization problem $\min_{X\in\mathbb{R}^{n\times d}} \|Y - XX^{\mathrm{T}}\|_{\mathrm{F}}^2 + (1/2\sigma^2)\|X\|_{\mathrm{F}}^2$. In particular, when $\sigma^2\to\infty$,

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

which corresponds to a noninformative flat prior, the maximum a posteriori estimator of (6) coincides with the adjacency spectral embedding $\hat{X}_{\mathrm{ASE}}$. Therefore, one can heuristically view Gaussian spectral embedding defined through (6) as a direct Bayesian analogy of adjacency spectral embedding.

Remark 7 (Generality of Gaussian spectral embedding). Recall that the random dot product graph model can be alternatively regarded as a low-rank matrix model: $Y = XX^{\mathrm{T}} + E$ for some low-rank matrix XX^{T} and some noise matrix E. In the formulation of Gaussian spectral embedding, we do not constrain the latent positions $x_1, \ldots, x_n$ to lie in the space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leqslant 1, x \geqslant 0\}$, and do not assume a parametric form for the distribution of the entries of Y. Namely, the Gaussian spectral embedding (6) is well defined, not only for the random dot product graph model, but also for a more general class of low-rank matrix models. In the theoretical analysis below, we also assume that the sampling model for Y is a more general low-rank matrix model $Y = XX^{\mathrm{T}} + E$ for some $X \in \mathbb{R}^{n \times d}$, and the entries of E are only required to be sub-Gaussian.

THEOREM 6. *Let $Y \in \mathbb{R}^{n \times n}$ be a symmetric random matrix with $(y_{ij} : 1 \leqslant i \leqslant j \leqslant n)$ being independent, and let $E_0(Y) = X_0 X_0^{\mathrm{T}}$ for some $X_0 \in \mathbb{R}^{n \times d}$, where $d/n \to 0$. Assume that $(1/n)X_0^{\mathrm{T}} X_0 \to \Delta$ as $n \to \infty$ for some positive definite $\Delta \in \mathbb{R}^{d \times d}$, and the entries of $Y - E_0(Y)$ are sub-Gaussian, i.e., there exists some constant $\tau > 0$ such that, for all $A \in \mathbb{R}^{n \times n}$ with $\|A\|_{\mathrm{F}}^2 = 1$, and all $t > 0$, $\mathrm{pr}_0\left[|\mathrm{Tr}\left\{A^{\mathrm{T}}(Y - X_0 X_0^{\mathrm{T}})\right\}| > t\right] \leqslant \mathrm{e}^{-\tau t^2}$. Then there exist some $M > 0$ and a constant C_τ only depending on τ and Δ such that, for sufficiently large n,*

$$E\left\{\Pi_{\mathrm{G}}\left(\frac{1}{n}\inf_{W \in \mathbb{O}(d)} \|X - X_0 W\|_{\mathrm{F}}^2 > \frac{Md \log n}{n} \;\middle|\; Y\right)\right\} \leqslant 14 \exp(-C_\tau M^2 n \log n).$$

On the one hand, when the sampling model is restricted to the random dot product graph model, the posterior contraction rate for the latent positions under Gaussian spectral embedding is slower than the optimal rate $1/n$ by an extra logarithmic factor, while posterior spectral embedding yields a rate-optimal contraction. On the other hand, Gaussian spectral embedding can be applied to more general low-rank matrix models, while posterior spectral embedding is specifically designed for the random dot product graph model. In addition, posterior spectral embedding requires the latent positions $x_1, \ldots, x_n$ to lie in the space $\mathcal{X}$. Such a restriction could potentially lead to a cumbersome Markov chain Monte Carlo sampler for posterior inference. In contrast, Gaussian spectral embedding has no constraint on the latent positions, making the corresponding posterior computation relatively convenient.

6. NUMERICAL EXAMPLES

6.1. *General set-up for posterior inference*

We evaluate the performance of the proposed posterior spectral embedding in comparison with Gaussian/adjacency spectral embedding through simulated examples and the analysis of a Wikipedia graph dataset. For each of the numerical set-ups, the posterior inferences are carried out through a standard Metropolis–Hastings sampler with 15 000 iterations, where the first 5000 iterations are discarded as burn-in, and 1000 post-burn-in samples are collected every 10 iterations. The prior density for $x_1, \ldots, x_n$ is set to be the uniform distribution $\mathrm{Un}(\mathcal{X})$ for posterior spectral embedding, and the Gaussian prior in (7) with $\sigma = 10$ for Gaussian spectral embedding. Additional details of the Metropolis–Hastings sampler are provided in the Supplementary Material.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Table 1. *Simulation set-up for positive semidefinite stochastic block models*

K	$K=3$	$K=5$	$K=7$
n	$n=600$	$n=1000$	$n=1400$
$(x_{01}^*,\ldots,x_{0K}^*)$	$\begin{bmatrix}0.3 & 0.3 & 0.6\\ 0.3 & 0.6 & 0.3\end{bmatrix}$	$\begin{bmatrix}0.3 & 0.3 & 0.7 & 0.7 & 0.5\\ 0.3 & 0.7 & 0.3 & 0.7 & 0.5\end{bmatrix}$	$\begin{bmatrix}0.2 & 0.2 & 0.2 & 0.5 & 0.5 & 0.5 & 0.7\\ 0.2 & 0.5 & 0.7 & 0.2 & 0.5 & 0.7 & 0.2\end{bmatrix}$

6.2. *Simulated examples*

We first consider stochastic block models with positive semidefinite block probability matrices as our simulated examples. Three simulation set-ups are considered, and the number of communities K and the unique values of their latent positions $(x_{01}^*,\ldots,x_{0K}^*)$ are tabulated in Table 1. In each simulation set-up, the numbers of vertices in different clusters are drawn from a multinomial distribution with the probability vector $(1/K,\ldots,1/K)^{\mathrm{T}}$.

For the posterior spectral embedding, we compute the point estimator $\hat{X}$ given in Theorem 4. A point estimator for Gaussian spectral embedding is also obtained in a similar fashion. Although the data-generating models are positive semidefinite stochastic block models, the posterior inferences are performed under the more general random dot product graph models as the working models. We perform the subsequent clustering based on the K-means procedure, as described in § 4.

Rand (1971) suggested using the Rand index to evaluate the performance of clustering. Specifically, given two partitions $\mathcal{C}_1=\{c_{11},\ldots,c_{1r}\}$ and $\mathcal{C}_2=\{c_{21},\ldots,c_{2s}\}$ of $[n]$, i.e., for $i=1,2$, the c_{ij} are disjoint and their union is $[n]$, denote by a the number of pairs of elements in $[n]$ that are both in the same set in $\mathcal{C}_1$ and in the same set in $\mathcal{C}_2$, and b the number of pairs in $[n]$ that are neither in the same set in $\mathcal{C}_1$ nor in the same set in $\mathcal{C}_2$. Then the Rand index is defined as $\mathrm{RI}=2(a+b)/\{n(n-1)\}$. The Rand index is a quantity between 0 and 1, with a higher value suggesting better accordance between the two partitions. In particular, when $\mathcal{C}_1$ and $\mathcal{C}_2$ are identical up to relabelling, $\mathrm{RI}=1$.

Comparisons of the Rand indices and the embedding errors $(1/n)\inf_W\|\hat{X}-X_0W\|_{\mathrm{F}}^2$ for the three embedding approaches are tabulated in Tables 2 and 3, respectively. We see that the point estimates of posterior spectral embedding are superior to the other two competitors in terms of higher Rand indices and lower embedding errors, whereas the point estimates of Gaussian spectral embedding perform the worst in all three set-ups. All three embedding approaches perform better as the number of vertices n increases. In particular, Gaussian spectral embedding does not produce satisfactory results when $n=600$ and $n=1000$, but performs decently well when $n=1400$. These numerical results are also in accordance with the theoretical results established in § 3, § 4 and § 5, suggesting the optimality of posterior spectral embedding and the suboptimality of adjacency and Gaussian spectral embedding.

We also visualize the three embeddings of the observed adjacency matrix for the three set-ups in Figs. 1, 2 and 3, respectively. The estimation errors of the point estimates under Gaussian spectral embedding can be clearly recognized from the figures when $n=600$ and $n=1000$. We also observe that, for the underlying true latent position $(0.7,0.7)^{\mathrm{T}}$ when $K=5$, adjacency spectral embedding and the point estimator of Gaussian spectral embedding produce estimates that may stay outside the latent position space $\mathcal{X}$, whereas the point estimates of posterior spectral embedding almost stay inside the space $\mathcal{X}$. This agrees with the fact that posterior spectral embedding requires the latent positions to stay inside $\mathcal{X}$, whereas Gaussian spectral embedding and adjacency spectral embedding do not have such constraints.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Table 2. *Simulated examples: Rand indices of different clustering methods*

Method	PSE (point estimate)	ASE	GSE (point estimate)
$K = 3, n = 600$	0.9171	0.9160	0.7826
$K = 5, n = 1000$	0.9584	0.9558	0.7187
$K = 7, n = 1400$	0.9964	0.9508	0.9505

PSE, the posterior spectral embedding; ASE, the adjacency spectral embedding; GSE, the Gaussian spectral embedding.

Table 3. *Simulated examples: errors* $(1/n)\inf_W \|\hat{X} - X_0 W\|_F^2$ *of different methods*

Method	PSE (point estimate)	ASE	GSE (point estimate)
$K = 3, n = 600$	1.281×10^{-2}	1.560×10^{-2}	2.792×10^{-2}
$K = 5, n = 1000$	6.851×10^{-3}	8.548×10^{-3}	1.418×10^{-2}
$K = 7, n = 1400$	3.460×10^{-3}	3.582×10^{-3}	4.200×10^{-3}

PSE, the posterior spectral embedding; ASE, the adjacency spectral embedding; GSE, the Gaussian spectral embedding.

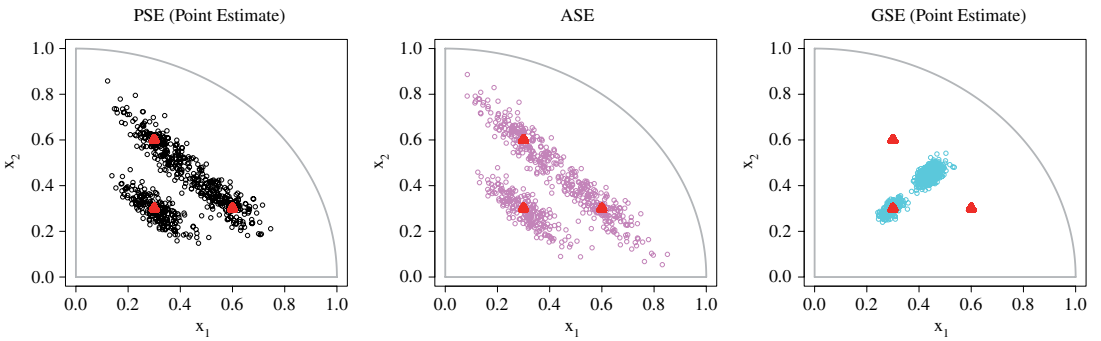

Fig. 1. Visualization of the three embedding approaches for the simulated positive semidefinite stochastic block models with $K = 3$. The red triangles are the true latent positions, and the scatter points are embedding estimates of the latent positions.

6.3. *Wikipedia graph data*

We next turn to the analysis of a real-world Wikipedia graph dataset, which is publicly available at `http://www.cis.jhu.edu/ parky/Data/data.html`. Specifically, the dataset consists of a network of articles that are within two hyperlinks of the article 'Algebraic Geometry', resulting in $n = 1382$ vertices. In addition, the articles involved are manually labelled as one of the following six classes: People, Places, Dates, Things, Math, Categories.

We first estimate the embedding dimension d by an ad hoc method. We examine the plot of the singular values of the observed adjacency matrix in Fig. 4, and directly locate an elbow that suggests a cut-off between the signal dimension and the noise dimension. For this Wikipedia dataset, the elbow is located at $\hat{d} = 3$.

We then conduct the posterior inferences under posterior spectral embedding, Gaussian spectral embedding and adjacency spectral embedding to obtain the estimates of the latent positions based on $\hat{d} = 3$. To obtain the clustering results, we further apply the `mclust` package in `R` (Fraley et al., 2012; R Development Core Team, 2020) to these embedding estimates with $K = 6$,

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

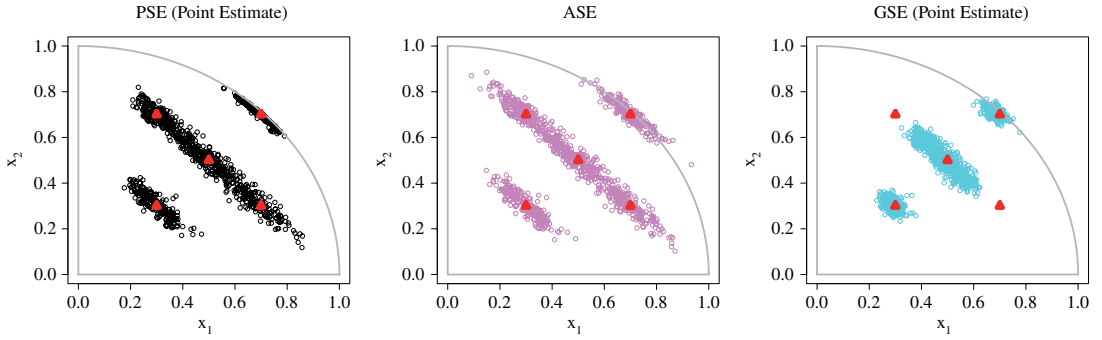

Fig. 2. Visualization of the three embedding approaches for the simulated positive semidefinite stochastic block models with $K = 5$. The red triangles are the true latent positions, and the scatter points are embedding estimates of the latent positions.

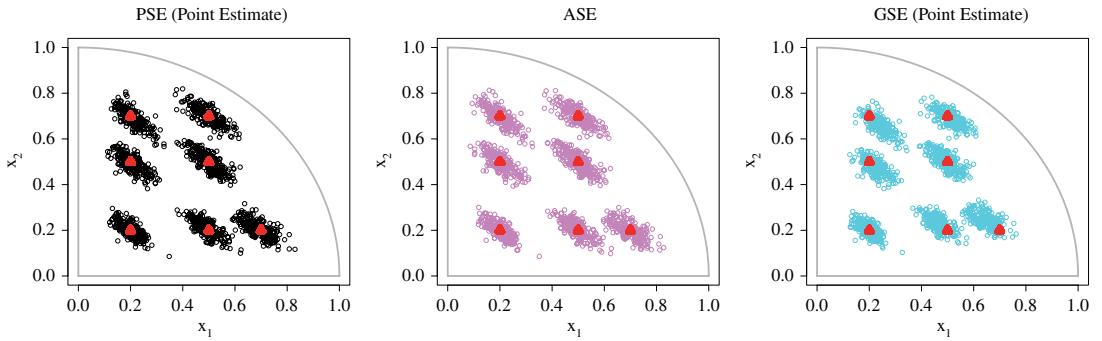

Fig. 3. Visualization of the three embedding approaches for the simulated positive semidefinite stochastic block models with $K = 7$. The red triangles are the true latent positions, and the scatter points are embedding estimates of the latent positions.

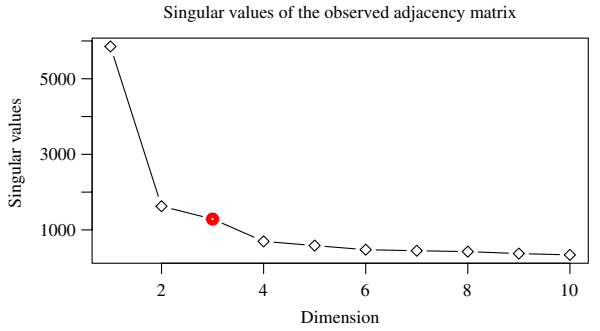

Fig. 4. Wikipedia graph data: singular values plot of the observed adjacency matrix. An elbow can be located at $\hat{d} = 3$ (the red circle).

as discussed in § 4, and compute their Rand indices with the manually labelled classes. The results are presented in Table 4, showing that the point estimate of posterior spectral embedding outperforms the other two approaches.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Table 4. *Wikipedia graph data: Rand indices of different clustering methods*

Method	PSE (point estimator)	ASE	GSE (point estimator)
Rand index	0.7451	0.7213	0.7155

PSE, the posterior spectral embedding; ASE, the adjacency spectral embedding; GSE, the Gaussian spectral embedding.

7. DISCUSSION

There are several potential extensions of the proposed methodology and the corresponding theory. First, the framework we have considered so far is based on the fact that the observed adjacency matrix of the network are Bernoulli random variables, i.e., an unweighted network. It is also common to encounter weighted network data in a wide range of applications (Schein et al., 2016; Tang et al., 2017b). Our theory and method can easily be extended to the weighted adjacency matrix, the elements of which typically follow distributions of more general forms. Alternatively, the Gaussian spectral embedding proposed in § 5 can be applied when the elements of the weighted adjacency matrix are sub-Gaussian random variables after centring. Second, we assume that the graph model is dense and undirected. Generalization of the random dot product graph model to sparse and directed networks, along with the corresponding theory, are provided in the Supplementary Material. Last, but not least, we assume that the embedding dimension d is known for ease of the mathematical analysis. When d is unknown, we can first consistently estimate d by some estimator $\hat{d}$ (Chatterjee, 2015), and then perform posterior/Gaussian spectral embedding based on $\hat{d}$. Alternatively, we can assign a prior distribution on d and let the posterior distribution adaptively select the correct dimension with moderate uncertainty. The challenge, nevertheless, is that it is nontrivial to design a reversible-jump sampler to address the cross-dimensional Monte Carlo problem for the random dot product graph model. We defer the computational issue with random d to future work. In contrast to Markov chain Monte Carlo samplers, which become computationally expensive when the number of vertices grows large, it is also worthwhile developing variational Bayes methods along with the corresponding theory for random graph models.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (1940107 and 1918854).

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains all the proofs, additional technical results, generalization to sparse and directed graphs, additional details of the implemented Metropolis–Hastings sampler for posterior inference, and an additional simulation example.

REFERENCES

ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. & SUSSMAN, D. L. (2018). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18**, 1–92.

ATHREYA, A., TANG, M., PARK, Y. & PRIEBE, C. E. (2020). On estimation and inference in latent structure random graphs. *arXiv:*1806.01401v3.

BICKEL, P. J. & CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106**, 21068–73.

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Bickel, P. J., Chen, A. & Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39**, 2280–301.
Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43**, 177–214.
Choi, D. S., Wolfe, P. J. & Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–84.
Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* **486**, 75–174.
Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report.
Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Nat. Acad. Sci.* **99**, 7821–6.
Goldenberg, A., Zheng, A. X., Fienberg, S. E. & Airoldi, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2**, 129–233.
Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97**, 1090–8.
Levin, K., Athreya, A., Tang, M., Lyzinski, V. & Priebe, C. E. (2019). A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv:*1705.09355v5.
Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–37.
Lovász, L. (2012). *Large Networks and Graph Limits*, vol. 60. Providence, RI: American Mathematical Society.
Lyzinski, V., Levin, K. & Priebe, C. E. (2018). On consistent vertex nomination schemes. *arXiv:*1711.05610v4.
Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A. & Priebe, C. E. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Statist.* **8**, 2905–22.
Lyzinski, V., Tang, M., Athreya, A., Park, Y. & Priebe, C. E. (2017). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Trans. Network Sci. Eng.* **4**, 13–26.
Priebe, C. E., Park, Y., Tang, M., Athreya, A., Lyzinski, V., Vogelstein, J.T., Qin, Y., Cocanougher, B., Eichler, K., Zlatic, M. & Cardona, A. (2017). Semiparametric spectral modeling of the *Drosophila* connectome. *arXiv:*1705.03297.
R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Assoc.* **66**, 846–50.
Rohe, K., Chatterjee, S. & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878–915.
Schein, A., Zhou, M., Blei, D. M. & Wallach, H. (2016). Bayesian Poisson Tucker decomposition for learning the structure of international relations. *Proc. Mach. Learn. Res.* **48**, 2810–19,
Sussman, D. L., Tang, M., Fishkind, D. E. & Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Am. Statist. Assoc.* **107**, 1119–28.
Sussman, D. L., Tang, M. & Priebe, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pat. Anal. Mach. Intel.* **36**, 48–57.
Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V. & Priebe, C. E. (2017a). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23**, 1599–630.
Tang, M. & Priebe, C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist.* **46**, 2360–415.
Tang, M., Sussman, D. L. & Priebe, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist.* **41**, 1406–30.
Tang, R., Ketcha, M., Badea, A., Calabrese, E. D., Margulies, D. S., Vogelstein, J. T., Priebe, C. E. & Sussman, D. L. (2018). Connectome smoothing via low-rank approximations. *IEEE Trans. Med. Imag.* **38**, 1446–56.
Tang, R., Tang, M., Vogelstein, J. T. & Priebe, C. E. (2017b). Robust estimation from multiple graphs under gross error contamination. *arXiv:*1707.03487.
van der Pas, S. L. & van der Vaart, A. W. (2018). Bayesian community detection. *Bayesian Anal.* **13**, 767–96.
Wang, S., Vogelstein, J. T. & Priebe, C. E. (2019). Joint embedding of graphs. *arXiv:*1703.03862v4.
Young, S. J. & Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *Proc. Int. Workshop on Algorithms and Models for the Web-Graph*. New York: Springer.
Zhao, Y., Levina, E. & Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40**, 2266–92.
Zhuo, B. & Gao, C. (2018). Mixing time of Metropolis–Hastings for Bayesian community detection. *arXiv:*1811.02612.

[*Received on* 25 *April* 2019. *Editorial decision on* 3 *December* 2019]

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022

Downloaded from https://academic.oup.com/biomet/article/107/4/875/5867841 by Johns Hopkins University user on 30 November 2022