Principal Subspace Estimation Under Information Diffusion

Fan Zhou, Ping Li, and Zhixin Zhou
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{zhoufan066, pingli98, zhixin0825}@gmail.com

Abstract

Let $\mathbf{A} = \mathbf{L}_0 + \mathbf{S}_0$, where $\mathbf{L}_0 \in \mathbb{R}^{d\times d}$ is low rank and $\mathbf{S}_0$ is a perturbation matrix. We study the principal subspace estimation of $\mathbf{L}_0$ through observations $\mathbf{y}_j = f(\mathbf{A})\mathbf{x}_j$, $j = 1, \ldots, n$, where $f : \mathbb{R} \to \mathbb{R}$ is an unknown polynomial and $\mathbf{x}_j$'s are i.i.d. random input signals. Such models are widely used in graph signal processing to model information diffusion dynamics over networks with applications in network topology inference and data analysis. We develop an estimation procedure based on nuclear norm penalization, and establish upper bounds on the principal subspace estimation error when $\mathbf{A}$ is the adjacency matrix of a random graph generated by $\mathbf{L}_0$. Our theory shows that when the signal strength is strong enough, the exact rank of $\mathbf{L}_0$ can be recovered. By applying our results to blind community detection, we show that consistency of spectral clustering can be achieved for some popular stochastic block models. Together with the experimental results, our theory show that there is a fundamental limit of using the principal components obtained from diffused graph signals which is commonly adapted in current practice. Finally, under some structured perturbation $\mathbf{S}_0$, we build the connection between this model with spiked covariance model and develop a new estimation procedure. We show that such estimators can be optimal under the minimax paradigm.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

We consider a matrix perturbation model where

$$\mathbf{A} = \mathbf{L}_0 + \mathbf{S}_0, \ \mathbf{A} \in \mathbb{R}^{d\times d} \tag{1.1}$$

where $\mathbf{L}_0$ is low rank and positive semi-definite (PSD) and $\mathbf{S}_0$ is a perturbation matrix. The problem of interest in this article is to study the principal subspace estimation of $\mathbf{L}_0$ given observations

$$\mathbf{y}_j = f(\mathbf{A})\mathbf{x}_j, \ j = 1, ..., n \tag{1.2}$$

with $\mathbf{x}_j$'s being i.i.d. copies of a Gaussian random vector $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $f : \mathbb{R} \to \mathbb{R}$ being an unknown polynomial which belongs to certain polynomial class $\mathcal{F}$.

One major motivation behind model (1.2) originates from network data analysis and graph signal processing (Sandryhaila and Moura, 2013; Shuman et al., 2013; Ortega et al., 2018). Model (1.2) is widely used to model information diffusion dynamics over networks. In particular, let one sample observation $\mathbf{y} = (y_1, ..., y_d)^T \in \mathbb{R}^d$ be a zero-mean graph signal in which the ith element y_i denotes the signal value at node i of an *unknown graph* $\mathcal{G}$ with graph-shift operator (GSO). Common choices of GSO can be either the adjacency matrix $\mathbf{A}$ of the graph or the Laplacian matrix $\mathbf{L} := \mathrm{diag}(\mathbf{A}) - \mathbf{A}$. The GSO can be used to define linear graph filters. Typically, these graph filters are linear graph signal operators that can be expressed as matrix polynomials of $\mathbf{A}$: $f(\mathbf{A}) = \sum_{\ell=0}^{T} \beta_\ell \mathbf{A}^\ell$. For a given excitation graph signal $\mathbf{x} \in \mathbb{R}^d$, the output of the filter is $\mathbf{y} = f(\mathbf{A})\mathbf{x}$, which is exactly model (1.2). Graphical data is widely used to capture network information such as the underlying dependency and/or similarity structure between the data points. For example, neural activities at different regions of the brain can be viewed on a graph where the regions are represented by nodes and the edge weights between the nodes encode the functional or structural connectivity levels among the corresponding regions (Honey et al.,

2007). Additionally, when a society is affected by an epidemic, a graph can show the individual interactions and data at each node can measure the infection level of each individual. The corresponding principal components of the network can be crucial in network topology inference (Segarra et al., 2017) or community detection (Wai et al., 2019).

According to Cayley-Hamilton theorem, any matrix polynomial of $\mathbf{A}$ can be represented as

$$f(\mathbf{A}) := \sum_{\ell=0}^{T} \beta_\ell \mathbf{A}^\ell = \mathbf{U}\Big(\sum_{\ell=0}^{T} \beta_\ell \mathbf{\Lambda}^\ell\Big)\mathbf{U}^T \tag{1.3}$$

where $\mathbf{A} := \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the eigen-decomposition of $\mathbf{A}$. It suggests that $\mathbf{U}$ is an invariant parameter of $\mathbf{A}$ under such diffusion process. Moreover, if $\mathbf{S}_0$ is a small perturbation under model (1.1), it's reasonable to use $\mathbf{U}$ to estimate the principal subspace of $\mathbf{L}_0$ which is of our interest. One major difficulty in solving this problem comes from the low rank constraint of $\mathbf{L}_0$. Apparently, both perturbation in (1.1) and diffusion in (1.2) can introduce noise and redundant information. Thus how to accurately locate the underlying low dimensional principal space becomes essential. Our first idea is to use nuclear norm penalization for low rank information retrieval. This classical tool was originally introduced by Candès and Recht (2009) in the classical matrix completion problem and has been very successful to study low rank matrix recovery during the past decade (Recht, 2011; Koltchinskii et al., 2011; Liu and Li, 2014; Chatterjee, 2015; Cai and Li, 2020) and the references therein. Our second idea is to use tools developed in principal component analysis (PCA) in covariance estimation. Indeed, model (1.2) suggests that the principal components of $\mathbf{A}$ is the same as that of the covariance matrix of the Gaussian random vector $\mathbf{y}$, which is relatively well understood.

There are several well studied models closely related to ours. Robust PCA was introduced by Candès et al. (2011) where they considered model (1.1) with $\mathbf{L}_0$ being a low rank component and $\mathbf{S}_0$ being a sparse perturbation. Clearly, the diffusion process in our model (1.2) introduces further correlation between $\mathbf{L}_0$ and $\mathbf{S}_0$ which makes the sparsity pattern of $\mathbf{S}_0$ hard to capture. So we don't assume any sparsity on $\mathbf{S}_0$. Another well studied topic related to ours is spiked covariance model, where $\mathbf{L}_0$ is a PSD low rank component with compact spectrum and $\mathbf{S}_0$ is a multiple of identity, see Paul (2007); Johnstone and Lu (2009). When $\mathbf{A}$ is equipped with this structure, we are able to develop estimation procedure that is minimax optimal under model (1.2). We notice that in a recent related work (Wai et al., 2019), the authors studied a special case of model (1.2) with $\mathbf{A}$ being the Laplacian matrix of the underlying graph focusing on the algorithmic and application aspects of blind community detection, while ours is focused on the theoretical understanding of the statistical estimation procedure. There are also other works which focus on minimax estimation problems that relate nonlinear functions and matrix (Gao et al., 2015; Zhou, 2019). Works that focus on principal subspace estimation under matrix perturbation model can be found in Cai and Zhang (2018); Xia and Zhou (2019) and the references therein.

Our major **contribution** is on the theory front. In Section 3 we introduce our low rank estimation procedure and establish a major result that serve as the cornerstone to derive bounds on the principal subspace estimation error. We show that when the signal strength is strong enough, the exact rank of $\mathbf{L}_0$ can be recovered. In Section 4, we derive bounds on principal subspace estimation under random graph setting and apply it in blind community detection in stochastic block model (SBM) (Holland et al., 1983) which is one of the most important model in network data analysis. We show that for some popular SBM, it implies consistency of spectral clustering. In Section 5, we turn to some structured perturbation and propose a new estimation procedure. We show that our model under this setting is closely related to the spiked covariance model (Johnstone and Lu, 2009) and minimax optimal rates are proved. In section 6, we conduct numerical simulation study to validate our analysis. Together with the upper bounds on principal subspace estimation proved for random graph setting, our experimental results show that there is a fundamental limit of using principal components of $f(\mathbf{A})$ to estimate those of $\mathbf{L}_0$ which is commonly adapted in practice. To our best knowledge, this is the first work that established those theoretical results in graph signal processing.

2 Preliminaries

2.1 Notations

Throughout this paper, we use boldface uppercase letter $\mathbf{X}$ to denote a matrix and boldface lowercase letter $\mathbf{x}$ to denote a vector. Given a matrix $\mathbf{A}$, we always use the form $\mathbf{A} = \sum_{j=1}^{d} \sigma_j(\mathbf{A})\mathbf{u}_j(\mathbf{A}) \otimes \mathbf{v}_j(\mathbf{A})$ to denote its singular value decomposition (SVD); we denote by $\|\mathbf{A}\|_{op} := \sigma_1(\mathbf{A})$ its spectral or operator norm; denote by $\|\mathbf{A}\|_F$ its Frobenius norm; denote by $\|\mathbf{A}\|_*$ its nuclear norm; denote by $\|\mathbf{A}\|_\infty$ its max-norm, i.e. $\|\mathbf{A}\|_\infty := \max_{i,j} \mathbf{A}_{ij}$. Moreover, if $\mathbf{A}$ is PSD, we denote by $\mathbf{r}(\mathbf{A}) := \mathrm{tr}(\mathbf{A})/\|\mathbf{A}\|_{op}$ the effective rank of $\mathbf{A}$. We denote by $[d] := \{1, 2, \ldots, d\}$. Given nonnegative a and b, $a \lesssim b$ or $a = O(b)$ means that $a \le Cb$ with a numerical constant C, and $a \asymp b$ or $a = \Omega(b)$ means that $a \lesssim b$ and $b \lesssim a$. $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2.2 Matrix Perturbation Model

Assumption 1. Assume that $\mathbf{A}$ is symmetric and yields a decomposition $\mathbf{A} = \mathbf{L}_0 + \mathbf{S}_0$ where $\mathbf{L}_0$ is positive semidefinite (PSD) with $\text{rank}(\mathbf{L}_0) = r \ll d$. Further, we assume that $\mathbf{L}_0$ has the following SVD:

$$\mathbf{L}_0 = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^T = \sum_{j=1}^{r} \lambda_j(\mathbf{L}_0)\mathbf{u}_j(\mathbf{L}_0) \otimes \mathbf{u}_j(\mathbf{L}_0), \quad (2.1)$$

with $\mathbf{U}_r \in \mathbb{R}^{d\times r}$, $\mathbf{\Lambda}_r \in \mathbb{R}^{r\times r}$, $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$.

Remark 1. Note that the PSD assumption on $\mathbf{L}_0$ is necessary. One reason is that there exist polynomials f that can cause identifiability issue on estimation of the principal components of $\mathbf{L}_0$ under model (1.2). We will have a detailed discussion on this in Section 2.3.

Assumption 2 (Random graph)**.** Let $\mathbf{A}$ be an adjacency matrix of a random graph of d nodes generated by $\mathbf{L}_0$ such that $\mathbb{E}[\mathbf{A}] = \mathbf{L}_0 - \text{diag}(\mathbf{L}_0)$, and each edge of the random graph occurs independently. Assume that $d \cdot \|\mathbf{L}_0\|_\infty \leq s$ for some $s \geq c_0 \log d$ with some constant $c_0 > 0$.

Example 1. One important example under Assumptions 1 and 2 is a popular stochastic block model (SBM) widely used in community detection. In this case, the adjacency matrix $\mathbf{A}$ is generated from $\mathbf{L}_0$ where $\mathbf{L}_0 = \mathbf{Z}\mathbf{B}\mathbf{Z}^T$. Here $\mathbf{Z} \in \mathbb{M}_{d,r}$ is the membership matrix where each row has one entry equals 1 and the rest are 0's. $\mathbf{B} \in \mathbb{R}^{r\times r}$ is the connectivity matrix. Especially,

$$\mathbf{B} = \alpha_d \mathbf{B}_0; \quad \mathbf{B}_0 = \lambda \mathbf{I}_r + (1-\lambda)\mathbf{1}_r \otimes \mathbf{1}_r, \quad \lambda \in (0,1),$$

where $\mathbf{I}_r \in \mathbb{R}^{r\times r}$ is the identity matrix and $\mathbf{1}_r \in \mathbb{R}^r$ is a vector of 1's.

Assumption 2 is standard in network analysis and community detection. Briefly, Example 1 exemplifies the edge probability within the same community is $\alpha_d\lambda$ and that across different communities is $\alpha_d(1-\lambda)$. The quantity s in Assumption 2 is an upper bound on the expected node degree of the random graph and characterizes the sparsity of the network. Since interesting networks in reality are mostly sparse. The following wonderful result proved by Lei and Rinaldo (2015) provides an upper bound on the size of such perturbation.

Proposition 2.1. (Spectral bound of binary symmetric random matrices) Suppose that Assumption 2 holds under model (1.1). Then for any $t > 0$ there exists a constant $C = C(t, c_0)$ such that with probability at least $1 - d^{-t}$

$$\|\mathbf{S}_0\|_{op} \leq C\sqrt{s}. \quad (2.2)$$

2.3 Polynomial with Homogenous Decaying Coefficients

The polynomial f in model (1.2) we consider belongs to the following polynomial class.

$$\mathcal{F}(\ell, \sigma, \lambda) := \Big\{ f(x) = \sum_{i=0}^{\ell} a_i x^i : |a_0| \leq \sigma;\ a_i = O(|\lambda|^{-(i-1)}),\ a_i \geq 0 \quad \forall i \geq 1 \Big\}.$$

The constant coefficient a_0 serves as a bias intercept. The degree ℓ characterizes the diffusion depth. When $\mathbf{A}$ is an adjacency matrix, higher order term in $f(\mathbf{A})$ models the interaction along longer path in the network. Throughout this paper, we assume that ℓ is a fixed constant. In diffusion dynamics over networks, the decaying coefficients can model the situation that further nodes have less effect on a given one.

Consider

$$\begin{aligned} f(\mathbf{A}) = f(\mathbf{L}_0 + \mathbf{S}_0) &= \sum_{i=0}^{\ell} a_i (\mathbf{L}_0 + \mathbf{S}_0)^i \\ &= f(\mathbf{L}_0) + S_f(\mathbf{L}_0, \mathbf{S}_0) \end{aligned} \quad (2.3)$$

where $S_f(\mathbf{L}_0, \mathbf{S}_0) := f(\mathbf{A}) - f(\mathbf{L}_0)$. One can immediately realize that given the full eigen-decomposition of $\mathbf{L}_0 = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ with $\mathbf{U} \in \mathbb{R}^{d\times d}$ and $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$, then $f(\mathbf{L}_0) = \mathbf{U} f(\mathbf{\Lambda}) \mathbf{U}^T$ which shares the same eigenspace with $\mathbf{L}_0$. Especially, if $f(0) = a_0 \neq 0$, $f(\mathbf{L}_0)$ is not necessarily low rank. However, one can decompose $f(\mathbf{L}_0) := \mathbf{L}_1 + a_0\mathbf{I}_d$ which implies that $f(\mathbf{L}_0)$ has a very simple decomposition form, i.e. a low rank component plus a scalar matrix. Moreover, $\mathbf{L}_1$ has a simple form of SVD: $\mathbf{L}_1 = \mathbf{U}_r(f(\mathbf{\Lambda}_r) - a_0\mathbf{I}_r)\mathbf{U}_r^T$ when $f \in \mathcal{F}(\ell, \sigma, \lambda)$. It essentially means that $\mathbf{L}_1$ perfectly preserves the eigenspace and rank information of $\mathbf{L}_0$. Meanwhile, given that the exact values of a_i's are unknown to us, which is the typical case in real world applications, the eigenvalue information is lost. Therefore, it is only reasonable for one to recover its eigenspace or rank information.

Given $f \in \mathcal{F}(\ell, \sigma, \lambda_1)$, another direct observation is that $\|\mathbf{L}_1\|_{op} = \sigma_1(\mathbf{L}_1) = \sum_{i=1}^{\ell} a_i \lambda_1^i \asymp \lambda_1$. This observation partly motivates our PSD assumption on $\mathbf{L}_0$. If $\mathbf{L}_0$ is not PSD, and the largest eigenvalue $\lambda_1 < 0$. Even with a simple polynomial function $\tilde{f} \in \mathcal{F}(\ell, \sigma, \lambda)$ such as $\tilde{f}(x) := |\lambda_1|^{-1}x^2 + x$, we can have $\tilde{f}(\lambda_1) = 0$, which means $\mathbf{L}_1$ loses the most important principal component information of $\mathbf{L}_0$. This will cause some identifiability issue of the problem thus we don't consider it here.

3 Information Retrieval via Nuclear Norm Penalization

In Section 2.3, we discussed that $f(\mathbf{A})$ can be decomposed into a low rank component plus some perturbation, and the low rank component preserves the principal subspace of $\mathbf{L}_0$. In this section, we construct an estimator using nuclear norm penalization to estimate the low rank component. We prove a major result below which serves as a cornerstone for us to derive bounds on principal subspace estimation.

Recall from Section 2.3, we can rewrite $f(\mathbf{A})$ as $f(\mathbf{A}) = \mathbf{L}_1 + \widetilde{S}_f(\mathbf{L}_0, \mathbf{S}_0)$, where $\widetilde{\mathbf{S}}_f(\mathbf{L}_0, \mathbf{S}_0) := S_f(\mathbf{L}_0, \mathbf{S}_0) + a_0\mathbf{I}_d$. Then the covariance matrix of $\mathbf{y}$ can be represented as

$$f^2(\mathbf{A}) = \mathbf{L}_2 + \bar{S}_f(\mathbf{L}_0, \mathbf{S}_0) \tag{3.1}$$

where we use $\mathbf{L}_2 = \mathbf{L}_1^2$ for the simplicity of representation and denote

$$\bar{S}_f(\mathbf{L}_0, \mathbf{S}_0) := \mathbf{L}_1\widetilde{S}_f(\mathbf{L}_0, \mathbf{S}_0) + \widetilde{S}_f(\mathbf{L}_0, \mathbf{S}_0)\mathbf{L}_1 + \widetilde{S}_f^2(\mathbf{L}_0, \mathbf{S}_0).$$

From (3.1), we can see that the covariance matrix can be decomposed into a low rank part plus some remainder. It is a natural idea to use nuclear norm penalization to extract the low rank component. Dealing with low rank estimation through nuclear norm minimization/penalization has been a standard approach along with the prosperous development of matrix completion and low rank recovery (Recht et al., 2010; Gross, 2011; Candes and Plan, 2010; Koltchinskii et al., 2011; Liu and Li, 2016; Shen and Li, 2016) and the references therein. Since the underlying true covariance matrix of $\mathbf{y}$ is not available, the sample covariance matrix $\widehat{\mathbf{\Sigma}} := n^{-1}\sum_{j=1}^n \mathbf{y}_j \otimes \mathbf{y}_j$ can serve as a good surrogate. Consider the following optimization problem

$$\widehat{\mathbf{L}} := \arg\min_{\mathbf{L}\in\mathbb{D}} \left\|\widehat{\mathbf{\Sigma}} - \mathbf{L}\right\|_F^2 + \varepsilon\|\mathbf{L}\|_*. \tag{3.2}$$

where $\mathbb{D}$ is a closed convex subset of the space of PSD matrices. The following lemma characterizes the performance of estimator (3.2) measured by Frobenius norm $\|\widehat{\mathbf{L}} - \mathbf{L}_2\|_F$.

Lemma 1. Suppose that Assumption 1 holds under model (1.2) with $\|\mathbf{S}_0\|_{op} \le \delta$, and $f \in \mathcal{F}(\ell, \sigma, \lambda_1)$ with $\lambda_1 := \lambda_1(\mathbf{L}_0) \ge C_1(\sigma \vee \delta)$ for some absolute constant $C_1 > 0$. Let $\widehat{\mathbf{L}}$ be the solution to (3.2). For any t_1, $t_2 > 0$, take

$$\varepsilon \ge C(\ell)\max\left\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2) + t_1}{n}}, \lambda_1(\sigma+\delta)\Big(1\vee\sqrt{\frac{d+t_2}{n}}\Big)\right\} \tag{3.3}$$

Then with probability at least $1 - (e^{-t_1} + 5e^{-t_2})$

$$\|\widehat{\mathbf{L}} - \mathbf{L}_2\|_F \le$$

$$C^*(\ell)\sqrt{r}\max\left\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2) + t_1}{n}}, \lambda_1(\sigma+\delta)\Big(1\vee\sqrt{\frac{d+t_2}{n}}\Big)\right\}. \tag{3.4}$$

Especially,

$$\mathbb{E}\|\widehat{\mathbf{L}} - \mathbf{L}_2\|_F \le$$

$$c^*(\ell)\sqrt{r}\max\left\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2))}{n}}, \lambda_1(\sigma+\delta)\Big(1\vee\sqrt{\frac{d}{n}}\Big)\right\}.$$

where $C(\ell)$, $C^*(\ell)$ and $c^*(\ell)$ are some constants depending on ℓ.

The proof of Lemma 1 (and missing proofs for other theorems) can be found in the supplementary material. The key challenge of the proof is to bound $\|\bar{S}_f(\mathbf{L}_0, \mathbf{S}_0)\|_{op}$ whose structure is quite complicated. Lemma 1 indicates that when the sample size n is large enough, $\|\widehat{\mathbf{L}} - \mathbf{L}_2\|_F \lesssim \sqrt{r}\lambda_1(\sigma+\delta)$. It shows that the estimation error of $\widehat{\mathbf{L}}$ is only controlled by the size of the perturbation δ when n is large enough. The intuition behind this is that large sample size n can only contribute to better covariance estimation. While the covariance estimation is accurate enough, it is the matrix perturbation that controls the estimation accuracy. Currently, we don't know whether bound (3.4) is optimal or not since to prove the lower bound, one needs some advanced mathematical tools that are not available as far as we are concerned. However, as we shall see in Section 6, our numerical experiments validate this phenomenon and show bound (3.4) could be very tight. On the other hand, as we shall see in Section 5, it is possible for us to design new estimators to further denoise this perturbation for some structured $\mathbf{S}_0$, and get estimators that can be not only consistent but also minimax optimal. This result is crucial for us to derive bounds on principal subspace estimation under our random graph setting (Assumption 2).

In the following theorem, we show that when the signal strength λ_r, the smallest singular value of $\mathbf{L}_0$ is strong enough, one can recover the exact rank of $\mathbf{L}_0$ with high probability.

Theorem 3.1. *Under the same condition as in Lemma 1, let $\widehat{\mathbf{L}}$ be the solution to (3.2) with ε taken as in Lemma 1. Suppose that for some $0 < \eta < 1$, $\widehat{\mathbf{L}}'$ be the solution to (3.2) with $\varepsilon' = \varepsilon/(1-\eta)$, namely, for some t_1, $t_2 > 0$*

$$\varepsilon' \ge \frac{C(\ell)}{1-\eta}\max\left\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2) + t_1}{n}}, \lambda_1(\sigma+\delta)\Big(1\vee\sqrt{\frac{d+t_2}{n}}\Big)\right\}. \tag{3.5}$$

Set $\widehat{r} := \mathrm{rank}(\widehat{\mathbf{L}}')$. Then with probability at least $1 - (e^{-t_1} + 5e^{-t_2})$

$$\widehat{r} \le r.$$

Moreover, if

$$\min_{j:\sigma_j(\mathbf{L}_2)\neq 0}\sigma_j(\mathbf{L}_2) \ge \frac{C(\ell)}{1-\eta}\max\left\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_0)}{n}}, \lambda_1(\sigma+\delta)\right\}, \tag{3.6}$$

then with the same probability

$$\widehat{r} \ge r.$$

Remark 2. Recall that the smallest singular value of $\mathbf{L}_2$ is $\min_{j:\sigma_j(\mathbf{L}_2)\neq 0}\sigma_j(\mathbf{L}_2) \asymp (\lambda_r^2+\lambda_r\sigma)$. An immediate observation is that when $\lambda_r \asymp \lambda_1$, then (3.6) can easily hold when the sample size n is large enough. Thus the exact rank r can be recovered with high probability.

4 Principal Subspace Estimation of Random Graph

4.1 Bounds on Principal Subspace

As we have discussed in Section 3, $\mathbf{L}_2$ preserves the principal subspace information of $\mathbf{L}_0$. Consider the eigen-decomposition (or equivalently SVD when $\mathbf{L}_2$ is PSD) of $\mathbf{L}_2 = \mathbf{U}_r\mathbf{\Lambda}\mathbf{U}_r^T$. A natural estimator of $\mathbf{U}_r$ is the first r-leading singular vectors $\widehat{\mathbf{U}}_r$ of $\widehat{\mathbf{L}}$. Denote by $\widehat{\mathbf{L}} := \widehat{\mathbf{U}}_r\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}_r^T$ the SVD of $\widehat{\mathbf{L}}$ with $\widehat{\mathbf{U}}_r \in \mathbb{R}^{d\times r}$. In Theorem 4.1, we derive an upper bound on $\|\widehat{\mathbf{U}}_r - \mathbf{U}_r\mathbf{Q}\|_F$ where $\mathbf{Q} \in \mathbb{R}^{r\times r}$ is an orthogonal matrix. It is a common metric used for principal subspace estimation under the perturbation model (1.1) and Assumption 2. The proof (in the supplementary material) follows from Lemma 1 and is an application of Davis-Kahan $\sin\Theta$-Theorem (Davis and Kahan, 1970; Yu et al., 2014).

Theorem 4.1. *Suppose that Assumption 1 and 2 hold under model (1.2), and $f \in \mathcal{F}(\ell,\sigma,\lambda_1)$ with $\lambda_1 \geq C_1(\sigma\vee\delta)$ for some constant $C_1 > 0$. Let $\widehat{\mathbf{L}}$ be the solution to (3.2) and $\widehat{\mathbf{U}}_r$ be the principal subspace with columns being the first r leading eigenvectors. Take*

$$\varepsilon \geq C_1(\ell)\max\Big\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2)}{n}}, \lambda_1(\sigma+\sqrt{s})\Big(1\vee\sqrt{\frac{d}{n}}\Big)\Big\} \tag{4.1}$$

Then for any orthogonal matrix $\mathbf{Q}\in\mathbb{R}^{r\times r}$, with probability at least $1-d^{-1}$

$$\|\widehat{\mathbf{U}}_r - \mathbf{U}_r\mathbf{Q}\|_F \leq$$
$$C_1^*(\ell)\frac{\sqrt{r}}{(\lambda_r^2\vee\lambda_r\sigma)}\max\Big\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2)}{n}}, \lambda_1(\sigma+\sqrt{s})\Big(1\vee\sqrt{\frac{d}{n}}\Big)\Big\}, \tag{4.2}$$

where $C_1^*(\ell)$ and $C_1(\ell)$ are constants depending on ℓ.

Remark 3. One implication of Theorem 4.1 is that when $\lambda_1 \asymp \lambda_r$, $\sigma \leq \sqrt{s}$, and the sample size n is large enough, then the bound in (4.2) implies that the following rate holds with high probability

$$\|\widehat{\mathbf{U}}_r - \mathbf{U}_r\mathbf{Q}\|_F \lesssim_\ell \frac{\sqrt{s}\sqrt{r}}{\lambda_r}.$$

In Section 4.2 we apply this result to blind community detection in SBM and show that it can imply consistency of spectral clustering for some popular SBMs.

Another common metric used in the literature to measure the distance between two subspaces is $\|\widehat{\mathbf{U}}_r\widehat{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\|_F$. Compared with the previous metric, this one is more straight forward to compute in practice. As a consequence, we show an upper bound on $\|\widehat{\mathbf{U}}_r\widehat{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\|_F$ in the sequel theorem which again follows from of Lemma 1.

Theorem 4.2. *Under the same condition of Theorem 4.1, let $\widehat{\mathbf{L}}$ be the solution to (3.2) and $\widehat{\mathbf{U}}_r$ be the principal subspace with columns being the first r leading eigenvectors. Take*

$$\varepsilon \geq C_2(\ell)\max\Big\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2)}{n}}, \lambda_1(\sigma+\sqrt{s})\Big(1\vee\sqrt{\frac{d}{n}}\Big)\Big\} \tag{4.3}$$

Then with probability at least $1-(e^{-t_1}+5e^{-t_2})$

$$\|\widehat{\mathbf{U}}_r\widehat{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\|_F \leq$$
$$\frac{C_2^*(\ell)\sqrt{r}}{\lambda_r^2\vee\lambda_r\sigma}\max\Big\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_2)+t_1}{n}}, \lambda_1(\sigma+\sqrt{s})\Big(1\vee\sqrt{\frac{d+t_2}{n}}\Big)\Big\}, \tag{4.4}$$

where $C_2^*(\ell)$ and $C_2(\ell)$ are constants depending on ℓ.

Remark 4. As one might have noticed, the term $\lambda_1(\sigma+\sqrt{s})(1\vee\sqrt{d/n})$ in bound (4.4) presents a fundamental limit of using $\widehat{\mathbf{U}}_r$ to estimate $\mathbf{U}_r$ under such diffusion models. Despite the fact that this estimator is commonly adopted in the literature, it indicates that even if the sample size n goes to infinity, it will not contribute much to improving the accuracy on the estimation of $\mathbf{U}_r$ once n is bigger than d. As we shall see in Section 6, our numerical simulation results verified this observation. Thus, this result suggests that when one intends to use $\widehat{\mathbf{U}}_r$ under this model, too many samples (when $n \gg d$) can be redundant.

4.2 Application in Blind Community Detection

Community detection has been one of the central topics in network data analysis while most of the works are based on a single observation of the network, see Newman and Girvan (2004); Amini et al. (2013); Bickel and Chen (2009); Sussman et al. (2012); Lei and Rinaldo (2015) and the references therein. Recently, we also see a surge of interests in generalization of such topics to tensor-valued data (Paul and Chen, 2016; Jing et al., 2020). We consider blind community detection in SBM. In SBM, a network with d nodes and r communities is parameterized by two matrices: 1. a membership matrix $\mathbf{Z} \in \mathbb{M}_{d,r}$ and 2. a symmetric connectivity matrix $\mathbf{P} \in [0,1]^{r\times r}$. The nodes are indexed by $[d]$ and the communities are indexed by $[r]$. For the membership matrix, a node $i \in [d]$ belongs to the kth community if and only if $\mathbf{Z}_{ik} = 1$ and $\mathbf{Z}_{i\ell} = 0$ when $\ell \neq k$. For the connectivity matrix, the entry $\mathbf{P}_{k\ell}$ is the edge probability between any nodes in community ℓ and community k. The adjacency matrix $\mathbf{A} \in \{0,1\}^{d\times d}$ that represents the network is generated

by

$$\begin{aligned}&\text{If } \mathbf{Z}_{ik} = \mathbf{Z}_{j\ell} = 1, \text{ then}\\ &\mathbf{A}_{ij} = \mathbf{A}_{ji} \sim \text{Bern}(\mathbf{P}_{k\ell}) \text{ independently } \forall i > j;\\ &\mathbf{A}_{ii} = 0, \quad \forall i \in [n]\end{aligned} \tag{4.5}$$

Especially, we can write

$$\mathbb{E}[\mathbf{A}] = \mathbf{Z}\mathbf{P}\mathbf{Z}^T - \text{diag}(\mathbf{Z}\mathbf{P}\mathbf{Z}^T). \tag{4.6}$$

Note that $\text{rank}(\mathbf{Z}\mathbf{P}\mathbf{Z}^T) \leq r$ and $\text{diag}(\mathbf{Z}\mathbf{P}\mathbf{Z}^T)$ with bounded operator norm has little effects. The goal of community recovery is to recover the membership matrix $\mathbf{Z}$ up to column permutations. A standard way to achieve this goal is to use a simple method called spectral clustering (Von Luxburg, 2007; Rohe et al., 2011; Balakrishnan et al., 2011; Fishkind et al., 2013; Rohe et al., 2011; Zhou and Li, 2020). In the classical setting, adjacency matrix $\mathbf{A}$ is given and so is the number of community r. Under our model, we only observe the graph signals $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n) \in \mathbb{R}^{d\times n}$, so the adjacency matrix $\mathbf{A}$ is latent. As a result, we can not apply spectral clustering algorithm through the adjacency matrix $\mathbf{A}$ directly. However, according to our discussion in Section 4, we can apply spectral clustering algorithm to $\widehat{\mathbf{U}}_r$ instead. Our estimation procedure is summarized in Algorithm 1.

Algorithm 1: Spectral-Clustering with approximate k-means

1: $\widehat{\mathbf{U}}_r \leftarrow$ top-r left singular vectors of $\mathbf{Y}$;
2: $\widehat{\mathbf{Z}} \leftarrow (1+\epsilon)$-approximation solution to k-means with r clusters on rows of $\widehat{\mathbf{U}}_r$;
3: Output $\widehat{\mathbf{Z}}$.

Remark 5. There are two computational concerns in practice: 1. we simply use the top-r left singular vectors of $\mathbf{Y}$ which is exactly the top-r left singular vectors of $\widehat{\mathbf{L}}$. Since it is shown in Koltchinskii et al. (2011) that the solution to (3.2) is just $\widehat{\mathbf{\Sigma}}$ with a soft threshold on its singular values. It means that $\widehat{\mathbf{U}}_r$ is again the top-r singular vectors of $\widehat{\mathbf{\Sigma}}$. Computationally, it is more stable to get $\widehat{\mathbf{U}}_r$ from $\mathbf{Y}$. 2. the original spectral clustering algorithm is applying the k-means clustering algorithm on the rows of $\widehat{\mathbf{U}}_r$. However, it is known that finding a global minimizer of such a procedure is NP-hard (Aloise et al., 2009). Instead, one can solve the $(1+\epsilon)$ approximate k-means that is computationally tractable and whose solution is within a constant fraction of the optimal value (Kumar et al., 2004).

Once we get the estimated membership matrix $\widehat{\mathbf{Z}} \in \mathbb{M}_{d,r}$, we consider a popular measure of estimation error in community detection called overall relative error: $L(\widehat{\mathbf{Z}}, \mathbf{Z}) := d^{-1} \min_{\mathbf{J}\in\mathbb{O}_r} \|\widehat{\mathbf{Z}}\mathbf{J} - \mathbf{Z}\|_0$, where $\mathbb{O}_r$ denotes the set of all $r \times r$ permutation matrices. In short, $L(\widehat{\mathbf{Z}}, \mathbf{Z})$ measures the overall proportion of misclassified nodes. The following theorem characterizes the error bound on $L(\widehat{\mathbf{Z}}, \mathbf{Z})$.

Theorem 4.3. *Suppose that the conditions of Theorem 4.1 hold, and let $\mathbf{A}$ be an adjacency matrix generated from SBM parametrized by $\mathbf{L}_0 = \mathbf{Z}\mathbf{P}\mathbf{Z}^T$. Assume that $\lambda_1 \asymp \lambda_r$ and $\sigma < \sqrt{s}$. Especially, let $\mathbf{P} = \alpha_d \mathbf{P}_0$ with $\|\mathbf{P}_0\|_\infty = 1$ and its smallest singular value $\lambda > 0$ be a constant. Let $\widehat{\mathbf{Z}}$ be the output of Algorithm 1. Then there exists an absolute constant c such that if*

$$\frac{(2+\epsilon)rd}{d_{\min}^2 \lambda^2 \alpha_d} < c \tag{4.7}$$

Then with probability at least $1 - d^{-1}$,

$$L(\widehat{\mathbf{Z}}, \mathbf{Z}) \leq \tilde{c}(\ell, \epsilon) \frac{r d_{\max}}{d_{\min}^2 \lambda^2 \alpha_d}. \tag{4.8}$$

where $\tilde{c}(\ell)$ is a constant depending on ℓ and ϵ. $d_{\max}$ and $d_{\min}$ are the largest and smallest community size respectively.

Remark 6. When $d_{\max} \asymp d_{\min}$ and $d\alpha_d = \Omega(\log d)$, then bound (4.8) with $r = o(\sqrt{\log d})$ implies $L(\widehat{\mathbf{Z}}, \mathbf{Z}) = o_p(1)$, which means the communities can be consistently recovered. This shows that our result (4.2) can be used to achieve consistency of spectral clustering for blind community detection of this important SBM with balanced community sizes. Similar results were proved by Lei and Rinaldo (2015) when $\mathbf{A}$ and r are given.

5 Principal Subspace Estimation under Structured Perturbation

As we have discussed in Section 3, the term $\lambda_1(\sigma + \delta)$ in bounds (3.4) controls the estimation accuracy of the estimator (3.2) as the sample size $n \to \infty$. The term σ is introduced by a_0. The other term $\sqrt{s}$ is caused by perturbation matrix $\mathbf{S}_0$. The major reason that large sample size n could not decrease the error to zero is that the estimator (3.2) fails to capture the complicated structure of $\bar{S}_f(\mathbf{L}_0, \mathbf{S}_0)$ with a random perturbation $\mathbf{S}_0$. In this section, we study some structured perturbation for which we are able to develop new estimators for denoising. Therefore consistency can be achieved. Such a model is closely related to the so called spiked covariance model (SPM) which draws a lot of attention during the past decade. We refer to Johnstone and Paul (2018) as a good survey on this topic. We further show that the estimator we get is minimax optimal when the new model is equivalent to SPM.

We consider the case $\mathbf{S}_0 = \epsilon \mathbf{I}_d$ for some $\epsilon > 0$. In the following proposition we show that under such perturbation $f^2(\mathbf{A})$ can be uniquely decomposed into a low

rank component plus a multiple of identity. What is more important is that the low rank component preserves the exact principal subspace information of $\mathbf{L}_0$.

Proposition 5.1. Suppose that Assumption 1 holds under model (1.2) with $\mathbf{S}_0 = \epsilon\mathbf{I}_d$. Let $r < d/2$ and $\lambda_1 \geq C_1(\sigma \vee \epsilon)$ for some absolute constant $C_1 > 0$.. Then for any $f \in \mathcal{F}(\ell, \sigma, \lambda_1)$, there exists a unique decomposition of $f^2(\mathbf{A})$ such that

$$f^2(\mathbf{A}) = f^2(\mathbf{L}_0 + \epsilon\mathbf{I}_d) := g(\mathbf{L}_0) + f^2(\epsilon) \tag{5.1}$$

where $g(\mathbf{L}_0) := \tilde{g}^2(\mathbf{L}_0) + 2f(\epsilon)\tilde{g}(\mathbf{L}_0)$ with $\tilde{g} \in \mathcal{F}(\ell, 0, \lambda_1)$.

$g(\mathbf{L}_0)$ preserves the rank information and eigenspace of $\mathbf{L}_0$. When $(\sigma + \epsilon) < \lambda_1$ and $\lambda_1 \asymp \lambda_r$, the covariance matrix $\mathbf{\Sigma} = f^2(\mathbf{A}) = g(\mathbf{L}_0) + f^2(\epsilon)$ of $\mathbf{y}$ is exactly a spiked covariance matrix. We propose an improved estimator of the low rank component $g(\mathbf{L}_0)$. Denote the spectrum of the sample covariance $\widehat{\mathbf{\Sigma}}$ by $\widehat{\sigma}(\mathbf{\Sigma}) := \{\sigma_1(\widehat{\mathbf{\Sigma}}), ..., \sigma_d(\widehat{\mathbf{\Sigma}})\}$. Take $\tilde{a}_0^2 := \text{Med}\{\sigma_1(\widehat{\mathbf{\Sigma}}), ..., \sigma_d(\widehat{\mathbf{\Sigma}})\}$ as the median of the singular values of $\widehat{\mathbf{\Sigma}}$. Consider the following optimization problem:

$$\widetilde{\mathbf{L}} = \arg\min_{\mathbf{L} \in \mathbb{D}} \left\| \widehat{\mathbf{\Sigma}} - (\mathbf{L} + \tilde{a}_0^2)\mathbf{I}_d \right\|_F^2 + \varepsilon \|\mathbf{L}\|_*. \tag{5.2}$$

The intuition behind this estimator is that $\tilde{a}_0^2$ serves as a good estimate of $f^2(\epsilon)$ as long as $\widehat{\mathbf{\Sigma}}$ is close to $\mathbf{\Sigma}$ given $\text{rank}(\mathbf{L}_0) < d/2$. To see why, we denote by $\Delta\mathbf{\Sigma} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$, then by the classical Wely's inequality from matrix perturbation theory (Stewart, 1990, page 203)

$$\sup_{j \geq 1} |\sigma_j(\widehat{\mathbf{\Sigma}}) - \sigma_j(\mathbf{\Sigma})| \leq \sup_{j \geq 1} |\sigma_j(\Delta\mathbf{\Sigma})| = \|\Delta\mathbf{\Sigma}\|_{op}. \tag{5.3}$$

Thus as long as $\|\Delta\mathbf{\Sigma}\|_{op}$ is small, $|\tilde{a}_0^2 - f^2(\epsilon)|$ is also small. In the following theorem, we prove an upper bound on estimation of the low rank component.

Theorem 5.2. *Suppose that Assumption 1 holds under model (1.2) with $\mathbf{S}_0 = \epsilon\mathbf{I}_d$ and $r < d/2$, and $f \in \mathcal{F}(\ell, \sigma, \lambda_1)$ with $\lambda_1 \geq C_1(\sigma \vee \epsilon)$ for some absolute constant $C_1 > 0$. Let $\widetilde{\mathbf{L}}$ be the solution to (5.2). For any t_1, $t_2 > 0$ take*

$$\varepsilon \geq C_3(\ell) \max\left\{ \lambda_1^2 \sqrt{\frac{\mathbf{r}(\mathbf{L}_0) + t_1}{n}}, \lambda_1(\sigma + \epsilon)\sqrt{\frac{d + t_2}{n}} \right\} \tag{5.4}$$

Then with probability at least $1 - (e^{-t_1} + 5e^{-t_2})$

$$\|\widetilde{\mathbf{L}} - g(\mathbf{L}_0)\|_{op} \leq C_3^*(\ell) \max\left\{ \lambda_1^2 \sqrt{\frac{\mathbf{r}(\mathbf{L}_0) + t_1}{n}}, \lambda_1(\sigma + \epsilon)\sqrt{\frac{d + t_2}{n}} \right\}. \tag{5.5}$$

Especially,

$$\mathbb{E}\|\widetilde{\mathbf{L}} - g(\mathbf{L}_0)\|_{op} \leq c_3^*(\ell) \max\left\{ \lambda_1^2 \sqrt{\frac{\mathbf{r}(\mathbf{L}_0)}{n}}, \lambda_1(\sigma + \epsilon)\sqrt{\frac{d}{n}} \right\}.$$

where $C_3(\ell)$, $C_3^*(\ell)$ and $c_3^*(\ell)$ are some constants depending on ℓ.

Remark 7. When the sample size $n \to \infty$, $\mathbb{E}\|\widetilde{\mathbf{L}} - g(\mathbf{L}_0)\|_{op} \to 0$ which shows that $\widetilde{\mathbf{L}}$ is a consistent estimator of the low rank component $g(\mathbf{L}_0)$.

In the following, we derive a minimax lower bound when $\lambda_1 \asymp \lambda_r$, which shows the optimality of estimator (5.2). The techniques used to prove the minimax lower bound are based on those used to prove minimax lower bounds for spiked covariance estimation (Vu and Lei, 2012; Birnbaum et al., 2013; Cai et al., 2015).

Firstly, we define the following parameter space which contains $\mathbf{A}$

$$\begin{aligned}\mathbf{\Theta}_0 :=& \{\mathbf{A} \in \mathbb{S}_d : \mathbf{A} = \mathbf{L}_0 + \epsilon\mathbf{I}_d,\ \mathbf{L}_0 \geq 0,\\ & \text{rank}(\mathbf{L}_0) \leq r,\ \|\mathbf{L}_0\|_{op} = \lambda_1 > |\epsilon| > 0\}.\end{aligned}$$

We consider the following parameter space:

$$\begin{aligned}\mathbf{\Theta} := \{g(\mathbf{L}_0) :& f(\mathbf{A}) = g(\mathbf{L}_0) + f^2(\epsilon),\\ & f \in \mathcal{F}(\ell, \sigma, \lambda_1),\ \mathbf{A} \in \mathbf{\Theta}_0\}.\end{aligned}$$

Theorem 5.3. *Under model (1.2), for some constant $\bar{c}(\ell) > 0$, the following minimax lower bound holds*

$$\inf_{\widetilde{\mathbf{L}}} \sup_{g(\mathbf{L}_0) \in \mathbf{\Theta}} \mathbb{E}\left\|\widetilde{\mathbf{L}} - g(\mathbf{L}_0)\right\|_{op} \geq \bar{c}(\ell) \max\left\{ \lambda_1^2 \sqrt{\frac{r}{n}}, \lambda_1(\sigma + \varepsilon)\sqrt{\frac{d}{n}} \right\}.$$

Remark 8. Note that when $\lambda_1 \asymp \lambda_r$, we have $\mathbf{r}(\mathbf{L}_0) \asymp r$. It shows that the estimator defined in (5.2) for the low rank component is actually minimax optimal.

Now we switch to prove the upper bounds on the principal subspace estimation.

Theorem 5.4. *Under the same condition of Theorem 5.2, let $\widetilde{\mathbf{U}}_r$ be the r-leading singular vectors of $\widetilde{\mathbf{L}}$. Then with the same probability*

$$\left\|\widetilde{\mathbf{U}}_r\widetilde{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\right\|_{op} \leq \frac{\tilde{C}(\ell)}{\lambda_r^2 + \lambda_r(\sigma + \epsilon)} \max\left\{ \lambda_1^2 \sqrt{\frac{\mathbf{r}(\mathbf{L}_0)}{n}}, \lambda_1(\sigma + \epsilon)\sqrt{\frac{d}{n}} \right\}, \tag{5.6}$$

where $\tilde{C}(\ell)$ is a constant depending on ℓ.

Remark 9. When $d = o(n)$, $\widetilde{\mathbf{U}}_r$ becomes a consistent estimator of $\mathbf{U}_r$. Especially, when $\lambda_1 \asymp \lambda_r$, bound (5.6) can reproduce the minimax optimal rate of principal subspace estimation for spike covariance model, see Cai et al. (2015).

Similar to Theorem 4.2, we can get an upper bound on $\|\widetilde{\mathbf{U}}_r\widetilde{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\|_F$.

Theorem 5.5. *Under the same condition of Theorem 5.4, let $\widetilde{\mathbf{U}}_r$ be the r-leading singular vectors of $\widetilde{\mathbf{L}}$. Then with the same probability*

$$\big\|\widetilde{\mathbf{U}}_r\widetilde{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\big\|_F \le \frac{\check{C}(\ell)\sqrt{r}}{\lambda_r^2+\lambda_r(\sigma+\epsilon)} \max\Big\{\lambda_1^2\sqrt{\frac{\mathbf{r}(\mathbf{L}_0)}{n}}, \lambda_1(\sigma+\epsilon)\sqrt{\frac{d}{n}}\Big\}. \tag{5.7}$$

where $\check{C}(\ell)$ is a constant depending on ℓ.

Remark 10. Bound (5.7) is very different from (4.4) as the right hand side goes to zero when $n \to \infty$. Thus it makes $\widetilde{\mathbf{U}}_r\widetilde{\mathbf{U}}_r^T$ a consistent estimator of $\mathbf{U}_r\mathbf{U}_r^T$. As we shall see in Section 6, our experimental results show that once n exceeds d, $\big\|\widetilde{\mathbf{U}}_r\widetilde{\mathbf{U}}_r^T - \mathbf{U}_r\mathbf{U}_r^T\big\|_F$ decreases rapidly as we increase n.

6 Numerical Simulation

In this section, we present numerical simulation results to validate our theoretical analysis. We create an underlying low rank parameter matrix $\mathbf{L}_0$ as described in Theorem 4.3:

$$\mathbf{L}_0 = \mathbf{Z}\mathbf{P}\mathbf{Z}^T, \tag{6.1}$$

where $\mathbf{Z}$ is a randomly generated membership matrix and $\mathbf{P}$ is generated as

$$\mathbf{P} = \alpha\big(\lambda\mathbf{I}_r + (1-\lambda)\mathbf{1}_r \otimes \mathbf{1}_r\big), \quad \lambda \in (0,1). \tag{6.2}$$

Clearly, this $\mathbf{L}_0$ satisfies Assumption 1. We choose

$$f(x) = \lambda_1^{-1}x^2 + x + \sigma$$

where λ_1 is the largest singular value of $\mathbf{L}_0$ and $\sigma < \lambda_1$ is a small constant. One can check that $f \in \mathcal{F}(2,\sigma,\lambda_1)$. We consider two types of perturbation model studied in this paper: 1. $\mathbf{A}_1 = \mathbf{L}_0 + \mathbf{S}_1$ is the adjacency matrix of a random graph generated by $\mathbf{L}_0$ as explained in (4.5) in Section 4.2. 2. $\mathbf{A}_2 = \mathbf{L}_0 + \mathbf{S}_2$ is the spiked covariance model, where $\mathbf{S}_2 = \epsilon\mathbf{I}_d$. Here we take $\epsilon = \|\mathbf{S}_1\|_{op}$ so that the noise levels of the two perturbation matrices are the same. By solving (3.2) and (5.2) respectively, we get $\widehat{\mathbf{L}}$ and $\widetilde{\mathbf{L}}$. Then we use their leading singular vectors as estimators of the singular vectors $\mathbf{U}$ of $\mathbf{L}_0$, we denote them by $\widehat{\mathbf{U}}$ and $\widetilde{\mathbf{U}}$ respectively. The metric we use to measure the distance between two principal subspaces are

$$\begin{aligned} err_1 &:= \big\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}\mathbf{U}^T\big\|_F; \\ err_2 &:= \big\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T - \mathbf{U}\mathbf{U}^T\big\|_F. \end{aligned} \tag{6.3}$$

We plot the estimation errors against different sample sizes for both cases with different values of d. The results are presented in the first row of Fig. 1. As we can see, when d is large compared with n, we can not see much difference between the estimation error for both cases. However, when we keep increasing n, we can see that err_2 decreases rapidly while we can hardly see any improvement in error reduction of err_1 for large n. This difference reflects the term $\lambda_1(\sigma+\sqrt{s})\big(1\vee\sqrt{d/n}\big)$ in bound (4.4) which presents a fundamental limit of using $\widehat{\mathbf{U}}_r$ to estimate $\mathbf{U}_r$. To take a closer look at the evolution of err_1 as n increases, we plot it in the second row of Fig. 1. As we can see, for each d, when $d > n$, we can observe relatively rapid decrease in err_1 as n increases. However, once $d < n$ the curve in each figure flattens, which validates bound (4.4). These experimental results shows that bound (5.7) can be quite tight, which essentially puts a question mark on the effectiveness of using $\widehat{\mathbf{U}}_r$ to estimate $\mathbf{U}$ which is commonly adapted in the community.

As we have mentioned, the result we proved in Theorem 1 is quite general. It can be applied to any subgaussian perturbation $\mathbf{S}_0$. In fact, the random graph case in Section 4 we studied is a special case of subgaussian perturbation. In this section, we plot the error against sample size for gaussian perturbations and show that bound (3.4) is still accurate. We set $\mathbf{S}_0$ to be a random matrix with independent zero-mean gaussian entries. In the forth and the fifth row of Fig. 1, we plot how err_1 evolves with different sample size and different noise levels, i.e. $\delta := \|\mathbf{S}_0\|_{op}$. As we can see, err_1 behaves similarly as that of the random graph case: 1. as the sample size n exceeds the dimension d, there is no substantial reduction in err_1; 2. err_1 is determined by the noise level $\|\mathbf{S}_0\|_{op}$ with large n. These two phenomena align well with bound (4.2) and may indicate that the results of Lemma 1 and Theorem 4.1 are tight for general subgaussian perturbations under this model.

As we have learned from the bounds in Theorem 5.4 and Theorem 5.5, the proposed estimator is consistent regardless of the size of perturbation in this case. So we conducted experiments with different values of ϵ. The outcome of err_2 with $\epsilon = 10,\ 1,\ 0.1$ are presented in the third row of Fig. 1. For each case, $\|\mathbf{S}_0\|_{op} = 10,\ 1,\ 0.1$. As we can see, err_2 converges to 0 as long as $n \to \infty$. This validates our results in Theorem 5.4 and Theorem 5.5. It shows that under such perturbation, even if the noise level $\delta := \|\mathbf{S}_0\|_{op}$ is large, as long as we have enough samples, the principal subspace can still be estimated accurately. However, as we have discussed, this is very different from other random perturbations such as the examples shown above.

7 Conclusion

In this article, we studied principal subspace estimation through diffused network data where theoretical results are rarely explored. One major result we get

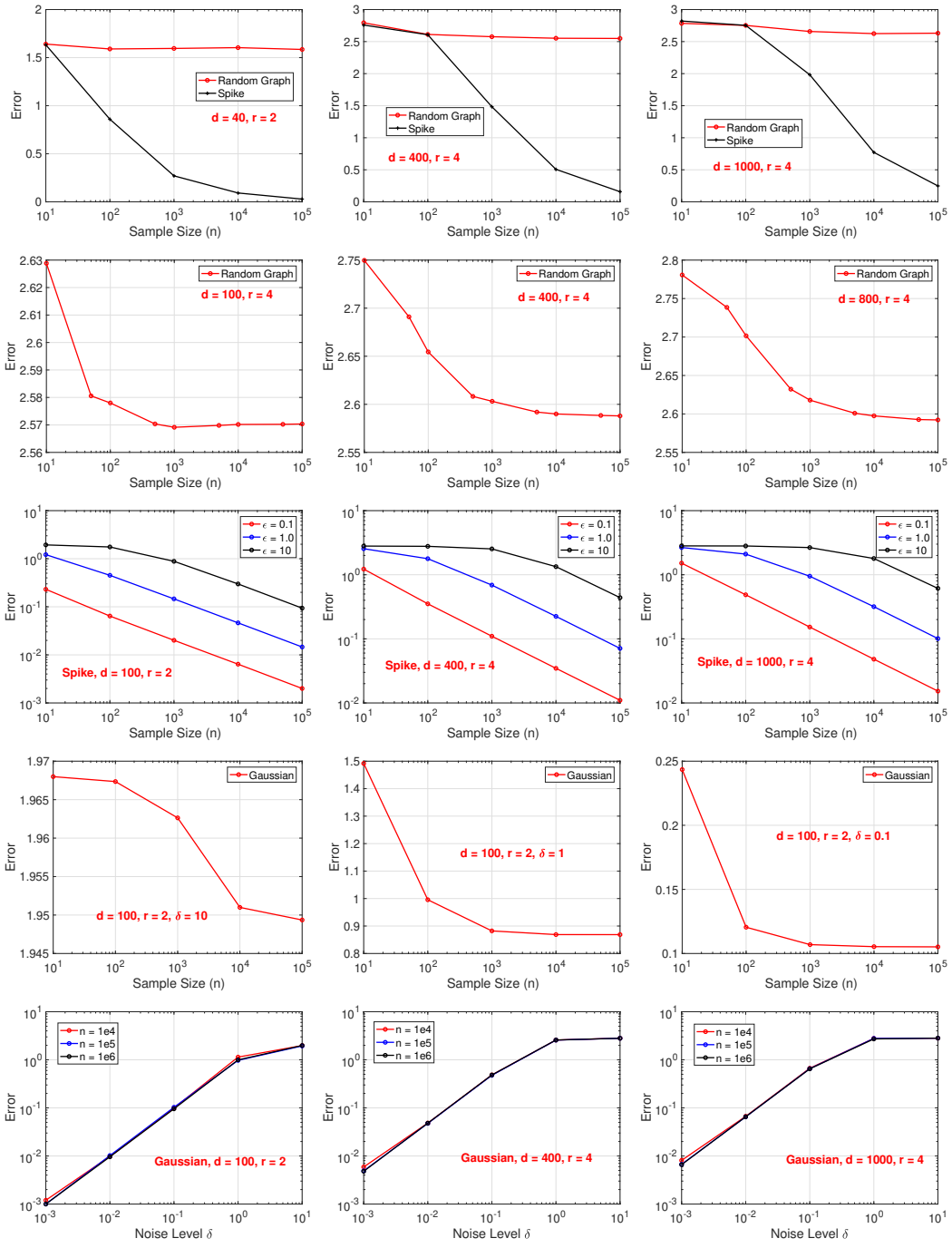

Figure 1: Principal Subspace Estimation Error

is the upper bound on principal subspace estimation for random graphs. We showed that such bounds can serve as the theoretical guarantee of consistency of spectral clustering in blind community detection for some popular SBMs. By combining our analysis with numerical simulation results, we showed that our upper bounds are very tight and it implies a fundamental limit of using principal components obtained from diffused graph signals to estimate the underlying network. We further show that with some structured perturbation this model can be connected to SPM, where minimax optimal estimators can be constructed for principal subspace estimation.

References

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

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

Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

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

Aharon Birnbaum, Iain M Johnstone, Boaz Nadler, and Debashis Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055, 2013.

T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, 2015.

Yunfeng Cai and Ping Li. Solving the robust matrix completion problem via a system of nonlinear equations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4162–4172, Online [Palermo, Sicily, Italy], 2020.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *Annals of Statistics*, 43(6):2624–2652, 2015.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Christopher J. Honey, Rolf Kötter, Michael Breakspear, and Olaf Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, 104(24):10240–10245, 2007.

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multi-layer networks via regularized tensor decomposition. *arXiv preprint arXiv:2002.04457*, 2020.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Iain M Johnstone and Debashis Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

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

Guangcan Liu and Ping Li. Recovery of coherent data via low-rank dictionary pursuit. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1206–1214, Montreal, Canada, 2014.

Guangcan Liu and Ping Li. Low-rank matrix completion in the presence of high coherence. *IEEE Trans. Signal Process.*, 64(21):5623–5633, 2016.

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

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2):3807–3870, 2016.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12 (Dec):3413–3430, 2011.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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

Aliaksei Sandryhaila and José MF Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013.

Santiago Segarra, Antonio G Marques, Gonzalo Mateos, and Alejandro Ribeiro. Network topology inference from spectral templates. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):467–483, 2017.

Jie Shen and Ping Li. Learning structured low-rank representation via matrix factorization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 500–509, Cadiz, Spain, 2016.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Gilbert W Stewart. *Matrix perturbation theory*. Citeseer, 1990.

Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107 (499):1119–1128, 2012.

Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Vincent Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 21–23 Apr 2012.

Hoi-To Wai, Santiago Segarra, Asuman E Ozdaglar, Anna Scaglione, and Ali Jadbabaie. Blind community detection from low-rank excitations of a graph filter. *IEEE Transactions on Signal Processing*, 2019.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Dong Xia and Fan Zhou. The sup-norm perturbation of hosvd and low rank tensor denoising. *Journal of Machine Learning Research*, 20(61):1–42, 2019.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.

Fan Zhou. Nonparametric estimation of low rank matrix valued function. *Electronic Journal of Statistics*, 13(2):3851–3892, 2019.

Zhixin Zhou and Ping Li. Rate optimal Chernoff bound and application to community detection in the stochastic block models. *Electronic Journal of Statistics*, 14(1):1302 – 1347, 2020.