The Annals of Statistics
2021, Vol. 49, No. 2, 944–967
https://doi.org/10.1214/20-AOS1986

SUBSPACE ESTIMATION FROM UNBALANCED AND INCOMPLETE DATA MATRICES: $\ell_{2,\infty}$ STATISTICAL GUARANTEES

BY CHANGXIAO CAI[1,*], GEN LI[2], YUEJIE CHI[3], H. VINCENT POOR[1,†] AND YUXIN CHEN[1,‡]

[1] *Department of Electrical Engineering, Princeton University,* [*] *ccai@princeton.edu;* [†] *poor@princeton.edu;* [‡] *yuxin.chen@princeton.edu*

[2] *Department of Electronic Engineering, Tsinghua University, g-li16@mails.tsinghua.edu.cn*

[3] *Department of Electrical and Computer Engineering, Carnegie Mellon University, yuejiechi@cmu.edu*

This paper is concerned with estimating the column space of an unknown low-rank matrix $\boldsymbol{A}^\star \in \mathbb{R}^{d_1 \times d_2}$, given noisy and partial observations of its entries. There is no shortage of scenarios where the observations—while being too noisy to support faithful recovery of the entire matrix—still convey sufficient information to enable reliable estimation of the column space of interest. This is particularly evident and crucial for the highly unbalanced case where the column dimension d_2 far exceeds the row dimension d_1, which is the focal point of the current paper.

We investigate an efficient spectral method, which operates upon the sample Gram matrix with diagonal deletion. While this algorithmic idea has been studied before, we establish new statistical guarantees for this method in terms of both ℓ_2 and $\ell_{2,\infty}$ estimation accuracy, which improve upon prior results if d_2 is substantially larger than d_1. To illustrate the effectiveness of our findings, we derive matching minimax lower bounds with respect to the noise levels, and develop consequences of our general theory for three applications of practical importance: (1) tensor completion from noisy data, (2) covariance estimation/principal component analysis with missing data and (3) community recovery in bipartite graphs. Our theory leads to improved performance guarantees for all three cases.

1. Introduction.

Consider the problem of estimating the *column space* of a low-rank matrix $\boldsymbol{A}^\star = [A^\star_{i,j}]_{1 \le i \le d_1, 1 \le j \le d_2}$, based on noisy and highly incomplete observations of its entries. To set the stage, suppose we observe

$$A_{i,j} = A^\star_{i,j} + N_{i,j} \quad \forall (i,j) \in \Omega, \tag{1.1}$$

where $\Omega \subseteq \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ is the sampling set, and $A_{i,j}$ denotes the observed entry at location (i, j), which is corrupted by noise $N_{i,j}$. In contrast to the classical matrix completion problem that aims to fill in all missing entries [21, 28, 36, 65], the current paper focuses solely on estimating the column space of $\boldsymbol{A}^\star$, which is oftentimes a less stringent requirement.

Motivating applications. A problem of this kind arises in numerous applications. We immediately point out several representative examples as follows, with precise descriptions postponed to Section 4:

- *Tensor completion.* Imagine we seek to estimate a low-rank symmetric tensor from partial observations of its entries [11, 84, 107], a task that spans various applications like visual data inpainting [76] and medical imaging [99]. Consider, for example, an order-3 tensor

Received October 2019; revised March 2020.
MSC2020 subject classifications. Primary 62F10; secondary 62H25.
Key words and phrases. Spectral method, principal component analysis, tensor completion, covariance estimation, missing data, spectral clustering, leave-one-out analysis.

$\boldsymbol{T}^{\star} = \sum_{s=1}^{r} \boldsymbol{w}_s^{\star} \otimes \boldsymbol{w}_s^{\star} \otimes \boldsymbol{w}_s^{\star} \in \mathbb{R}^{d\times d\times d}$, where $\{\boldsymbol{w}_s^{\star}\}$ represents a collection of tensor factors.[1] An alternative representation of $\boldsymbol{T}^{\star}$ can be obtained by unfolding the tensor of interest into a low-rank matrix $\boldsymbol{A}^{\star} \in \mathbb{R}^{d\times d^2}$. Consequently, estimation of the subspace spanned by $\{\boldsymbol{w}_s^{\star}\}$ from partial noisy entries of $\boldsymbol{T}^{\star}$—which serves as a common and crucial step for tensor completion [14, 84]—is equivalent to estimating the column space of $\boldsymbol{A}^{\star}$ from incomplete data; see Section 4.1. Notably, the unfolded matrix becomes extremely fat as the dimension d grows.

- *PCA with missing data.* Suppose we have available a sequence of n independent sample vectors $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, whose covariance matrix exhibits certain low-dimensional structure. Several statistical models fall in the same vein of this model, for example, the generalized spiked model [9] and the factor model [44]. An important task amounts to estimating the principal subspace of the covariance matrix of interest, possibly in the presence of missing data (where we only get to see highly incomplete entries of $\{\boldsymbol{x}_i\}$). If a substantial amount of data are missing, then individual sample vectors cannot possibly be recovered, thus enabling privacy protection for individual data. Fortunately, one might still hope to estimate the principal subspace, provided that a large number of sample vectors are queried (which might yield a fat data matrix $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$); see Section 4.2.
- *Community recovery in bipartite graphs.* Community recovery is often concerned with clustering a collection of individuals or nodes into different communities, based on similarities between pairs of nodes. In many complex networks, such pairwise interactions might only occur when the two nodes involved belong to two disjoint groups (denoted by $\mathcal{U}$ and $\mathcal{V}$, resp.). This calls for community recovery in bipartite networks (sometimes referred to as biclustering) [7, 41, 71, 116]. As we shall detail in Section 4.3, the biclustering problem is tightly connected to subspace estimation; for instance, the column subspace of some biadjacency matrix $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{V}|}$ (which is a noisy copy of a low-rank matrix) reveals the community memberships in $\mathcal{U}$. When the size of $\mathcal{V}$ is substantially larger than that of $\mathcal{U}$, one might encounter a situation where only the nodes in $\mathcal{U}$ (rather than $\mathcal{V}$) can be reliably clustered. This calls for development of "one-sided" community recovery algorithms, that is, the type of algorithms that guarantee reliable clustering of $\mathcal{U}$ without worrying about the clustering accuracy in $\mathcal{V}$.

Contributions. Since we concentrate primarily on estimating the column space of $\boldsymbol{A}^{\star}$, it is natural to expect a reduced sample complexity as well as a weaker requirement on the signal-to-noise ratio, in comparison to the conditions required for reliable reconstruction of the whole matrix—particularly for those highly unbalanced problems with drastically different dimensions d_1 and d_2. Focusing on a spectral method applied to the Gram matrix $\boldsymbol{A}\boldsymbol{A}^{\top}$ with diagonal deletion (whose variants have been studied in multiple contexts [43, 47, 77, 79, 84]), we establish new statistical guarantees in terms of the sample complexity and the estimation accuracy, both of which strengthen prior theory. Our results deliver optimal $\ell_{2,\infty}$ estimation risk bounds with respect to the noise level, which are previously unavailable. All of this is accomplished via a powerful leave-one-out/leave-two-out analysis framework. Further, we develop minimax lower bounds under Gaussian noise, revealing that the sample complexity and the signal-to-noise ratio (SNR) required for spectral methods to achieve consistent estimation are both minimax optimal (up to some logarithmic factor). Finally, we develop concrete consequences of our general theory for all three applications mentioned above, leading to improved performance guarantees.

It is worth noting that low-rank subspace estimation from noisy and incomplete data has been extensively studied in a large number of prior work (e.g., [4, 27, 35, 38, 79, 113,

[1] For any vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^d$, we use $\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$ to denote a $d \times d \times d$ array whose (i, j, k)th entry is given by $a_i b_j c_k$.

117]). While many of these prior results allow d_1 and d_2 to differ, they typically fall short of establishing optimal dependency on d_1 and d_2 in the highly unbalanced scenarios. The focal point of this paper is thus to characterize the effect of such unbalancedness (as reflected by the aspect ratio d_2/d_1) upon consistent subspace estimation.

Paper organization. The rest of this paper is organized as follows. Section 2 formulates the problem and introduces basic definitions and notation. In Section 3, we present our theoretical guarantees for a spectral method, as well as minimax lower bounds. Section 4 applies our general theorem to the aforementioned applications, and corroborate our theory by numerical experiments. Section 5 provides an overview of related prior works. The proof of our main theory and auxiliary lemmas are postponed to the Supplementary Material [13]. We conclude the paper with a discussion of future directions in Section 6.

2. Problem formulation.

2.1. *Models.*

Low-rank matrix. Suppose that the unknown matrix $\boldsymbol{A}^\star \in \mathbb{R}^{d_1 \times d_2}$ is rank-r, where the row dimension d_1 and the column dimension d_2 are allowed to be drastically different. Assume that the (compact) singular value decomposition (SVD) of $\boldsymbol{A}^\star$ is given by

$$\boldsymbol{A}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} = \sum_{i=1}^{r} \sigma_i^\star \boldsymbol{u}_i^\star \boldsymbol{v}_i^{\star\top}. \tag{2.1}$$

Here, $\sigma_1^\star \geq \sigma_2^\star \geq \cdots \geq \sigma_r^\star > 0$ represent the r nonzero singular values of $\boldsymbol{A}^\star$, and $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{r \times r}$ is a diagonal matrix whose diagonal entries are given by $\{\sigma_1^\star, \ldots, \sigma_r^\star\}$. The columns of $\boldsymbol{U}^\star = [\boldsymbol{u}_1^\star, \ldots, \boldsymbol{u}_r^\star] \in \mathbb{R}^{d_1 \times r}$ (resp., $\boldsymbol{V}^\star = [\boldsymbol{v}_1^\star, \ldots, \boldsymbol{v}_r^\star] \in \mathbb{R}^{d_2 \times r}$) are orthonormal, which are the top-r left (resp., right) singular vectors of $\boldsymbol{A}^\star$. We define and denote the condition number of $\boldsymbol{A}^\star$ as $\kappa := \sigma_1^\star / \sigma_r^\star$, and take $d := \max\{d_1, d_2\}$.

Incoherence. Further, we impose certain incoherence conditions on the unknown matrix $\boldsymbol{A}^\star$, which are commonly adopted in the matrix completion literature (e.g., [21, 36, 65]).

DEFINITION 2.1 (Incoherence parameters). *Define the incoherence parameters μ_0, μ_1 and μ_2 as follows*:

$$\mu_0 := \frac{d_1 d_2 \max_{i,j} |A_{i,j}^\star|^2}{\|\boldsymbol{A}^\star\|_{\mathrm{F}}^2}, \tag{2.2a}$$

$$\mu_1 := \frac{d_1}{r} \max_i \|\boldsymbol{U}^{\star\top} \boldsymbol{e}_i\|_2^2 \quad \textit{and} \quad \mu_2 := \frac{d_2}{r} \max_i \|\boldsymbol{V}^{\star\top} \boldsymbol{e}_i\|_2^2, \tag{2.2b}$$

where $\boldsymbol{e}_i$ is the ith standard basis vector of compatible dimensionality.

Intuitively, when μ_0, μ_1 and μ_2 are all small, the energies of the matrices $\boldsymbol{A}^\star$, $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ are (nearly) evenly spread out across all entries, rows and columns. For notational simplicity, we shall set

$$\mu := \max\{\mu_0, \mu_1, \mu_2\}. \tag{2.3}$$

Random sampling and random noise. Suppose that we have only collected noisy observations of the entries of $\boldsymbol{A}^\star$ over a sampling set $\Omega \subseteq \{1, \dots, d_1\} \times \{1, \dots, d_2\}$. Specifically, we observe

$$A_{i,j} = \begin{cases} A^\star_{i,j} + N_{i,j} & (i, j) \in \Omega, \\ 0 & \text{else}, \end{cases} \tag{2.4}$$

where $N_{i,j}$ is the noise at location (i, j). For notational simplicity, we write

$$\boldsymbol{A} = \mathcal{P}_\Omega(\boldsymbol{A}) = \mathcal{P}_\Omega(\boldsymbol{A}^\star) + \mathcal{P}_\Omega(\boldsymbol{N}), \tag{2.5}$$

where $\mathcal{P}_\Omega$ represents the Euclidean projection onto the subspace of matrices supported on Ω. In addition, this paper concentrates on random sampling and random noise as follows.

ASSUMPTION 2.2 (Random sampling). *Each (i, j) is included in the sampling set Ω independently with probability p.*

ASSUMPTION 2.3 (Random noise). *The noise components $\{N_{i,j}\}$ are independent and satisfy the following conditions: for each $1 \le i \le d_1, 1 \le j \le d_2$,*

(1) *(Zero mean)* $\mathbb{E}[N_{i,j}] = 0$;
(2) *(Variance)* $\mathsf{Var}(N_{i,j}) \le \sigma^2$;
(3) *(Magnitude) Each $N_{i,j}$ satisfies either of the following condition*:
  (a) $|N_{i,j}| \le R$;
  (b) *$N_{i,j}$ has a symmetric distribution satisfying $\mathbb{P}\{|N_{i,j}| > R\} \le c_{\mathrm{r}} d^{-12}$ for some universal constant $c_{\mathrm{r}} > 0$.*
  Here, R is some quantity obeying

$$\frac{R^2}{\sigma^2} \le C_{\mathrm{r}} \frac{\min\{p\sqrt{d_1 d_2}, p d_2\}}{\log d} \tag{2.6}$$

for some universal constant $C_{\mathrm{r}} > 0$.

As a remark, Assumption 2.3 allows the largest possible size R of each noise component to be substantially larger than its typical size σ. For example, if $p \asymp 1$, then R can be $\min\{(d_1 d_2)^{1/4}, \sqrt{d_2}\}$ times larger than σ (ignoring log factors). In addition, the $N_{i,j}$'s do not necessarily have identical variance; in fact, our formulation allows us to accommodate the heteroscedasticity of noise (i.e., the scenario where the noise has location-varying variance).

Goal. Given incomplete and noisy observations about $\boldsymbol{A}^\star \in \mathbb{R}^{d_1 \times d_2}$ (cf. (2.4)), we seek to estimate $\boldsymbol{U}^\star \in \mathbb{R}^{d_1 \times r}$ modulo some global rotation. We emphasize once again that the aim here is not to estimate the entire matrix. In truth, there are many unbalanced cases with $d_2 \gg d_1$ such that (1) reliable estimation of $\boldsymbol{U}^\star$ is feasible, but (2) faithful estimation of the whole matrix $\boldsymbol{A}^\star$ is information theoretically impossible.

2.2. *Notation.* We denote $[n] := \{1, \dots, n\}$. For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, we use $\sigma_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{A})$ to represent the ith largest singular value and the ith largest eigenvalue of $\boldsymbol{A}$, respectively. Let $\boldsymbol{A}_{i,:}$ and $\boldsymbol{A}_{:,j}$ denote respectively the ith row and the jth column of $\boldsymbol{A}$. Let $\|\boldsymbol{A}\|$ (resp., $\|\boldsymbol{A}\|_{\mathrm{F}}$) represent the spectral norm (resp., the Frobenius norm) of $\boldsymbol{A}$. We also denote by $\|\boldsymbol{A}\|_{2,\infty} := \max_{i \in [d_1]} \|\boldsymbol{A}_{i,:}\|_2$ and $\|\boldsymbol{A}\|_\infty := \max_{i \in [d_1], j \in [d_2]} |A_{i,j}|$ the $\ell_{2,\infty}$ norm and the entrywise ℓ_∞ norm of $\boldsymbol{A}$, respectively. Similarly, for any tensor $\boldsymbol{T}$, we use $\|\boldsymbol{T}\|_\infty$ to represent the largest magnitude of the entries of $\boldsymbol{T}$. Moreover, we denote by $\mathcal{P}_{\mathsf{diag}}$ the projection onto the subspace that vanish outside the diagonal, and define $\mathcal{P}_{\mathsf{off\text{-}diag}}$ such that $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}) := \boldsymbol{A} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{A})$. Let $\mathcal{O}^{r \times r}$ stand for the set of $r \times r$ orthonormal matrices. In addition, we use $\mathsf{diag}(\boldsymbol{a})$ to represent a diagonal matrix whose (i,i)th entry is equal to a_i. Throughout this paper, the notation $C, C_1, \dots, c, c_1, \dots$ denote absolute positive constants whose values may change from line to line.

For any real-valued functions $f(d_1,d_2)$ and $g(d_1,d_2)$, $f(d_1,d_2) \lesssim g(d_1,d_2)$ or $f(d_1,d_2) = O(g(d_1,d_2))$ mean that $|f(d_1,d_2)/g(d_1,d_2)| \le C_1$ for some constant $C_1 > 0$; $f(d_1,d_2) \gtrsim g(d_1,d_2)$ means that $|f(d_1,d_2)/g(d_1,d_2)| \ge C_2$ for some universal constant $C_2 > 0$; $f(d_1,d_2) \asymp g(d_1,d_2)$ means that $C_1 \le |f(d_1,d_2)/g(d_1,d_2)| \le C_2$ for some universal constants $C_1, C_2 > 0$; $f(d_1,d_2) = o(g(d_1,d_2))$ means that $f(d_1,d_2)/g(d_1,d_2) \to 0$ as $\min\{d_1,d_2\} \to \infty$. In addition, $f(d_1,d_2) \ll g(d_1,d_2)$ (resp., $f(d_1,d_2) \gg g(d_1,d_2)$) means that there exists some sufficiently small (resp., large) constant $c_1 > 0$ (resp., $c_2 > 0$) such that $f(d_1,d_2) \le c_1 g(d_1,d_2)$ (resp. $f(d_1,d_2) \ge c_2 g(d_1,d_2)$) holds true for all sufficiently large d_1 and d_2.

3. Main results.

3.1. *Algorithm: A spectral method with diagonal deletion.* Recall that $\boldsymbol{A} = [A_{i,j}]_{1\le i\le d_1, 1\le j\le d_2}$ is the zero-padded data matrix (see (2.4)). It is easily seen that, under our random sampling model (i.e., Assumption 2.2), $p^{-1}\boldsymbol{A}$ serves as an unbiased estimator of $\boldsymbol{A}^\star$. One might thus expect the left singular subspace of $\boldsymbol{A}$ to form a reasonably good estimator of the subspace spanned by $\boldsymbol{U}^\star$. As it turns out, when $\boldsymbol{A}^\star$ is a very fat matrix (namely, $d_2 \gg d_1$), this approach might fail to work when the sample complexity is not sufficiently large or when the noise size is not sufficiently small.

This paper adopts an alternative route by resorting to the sample Gram matrix $\boldsymbol{A}\boldsymbol{A}^\top$ (properly rescaled). Straightforward calculation reveals that

$$\begin{aligned}\frac{1}{p^2}\mathbb{E}[\boldsymbol{A}\boldsymbol{A}^\top] \\ = \boldsymbol{A}^\star\boldsymbol{A}^{\star\top} + \underbrace{\left(\frac{1-p}{p}\right)\mathcal{P}_{\mathsf{diag}}(\boldsymbol{A}^\star\boldsymbol{A}^{\star\top}) + \frac{1}{p}\mathsf{diag}\left(\left[\sum_{j=1}^{d_2}\mathsf{Var}(N_{i,j})\right]_{1\le i\le d_1}\right)}_{\text{a diagonal matrix}},\end{aligned} \tag{3.1}$$

where $\mathsf{diag}(\boldsymbol{a})$ with $\boldsymbol{a} \in \mathbb{R}^{d_1}$ represents a diagonal matrix whose (i,i)th entry equals a_i. The identity (3.2) implies that the diagonal components of $p^{-2}\mathbb{E}[\boldsymbol{A}\boldsymbol{A}^\top]$ are significantly inflated, which call for special care.

In order to remedy the above-mentioned diagonal inflation issue, we adopt a simple strategy that zeros out all diagonal entries; that is, performing the spectral method on the following matrix:

$$\boldsymbol{G} = \frac{1}{p^2}\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top) \tag{3.2}$$

with $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{M}) := \boldsymbol{M} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{M})$ denoting projection onto the set of zero-diagonal matrices. This clearly satisfies

$$\mathbb{E}[\boldsymbol{G}] = \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}^\star\boldsymbol{A}^{\star\top}) = \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{U}^\star\boldsymbol{\Sigma}^{\star 2}\boldsymbol{U}^{\star\top}).$$

If the diagonal entries of $\boldsymbol{A}^\star\boldsymbol{A}^{\star\top}$ are not too large, then one has $\boldsymbol{A}^\star\boldsymbol{A}^{\star\top} \approx \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}^\star\boldsymbol{A}^{\star\top})$ and, as a result, the rank-r eigen-subspace of $\boldsymbol{G}$ might form a reliable estimate of the subspace spanned by $\boldsymbol{U}^\star$. The procedure is summarized in Algorithm 1.

We remark that this is clearly not a new algorithmic idea. In fact, proper handling of the diagonal entries (e.g., diagonal deletion, diagonal reweighting) has already been recommended in several different applications, including bipartite stochastic block models [47], covariance estimation [43, 77–79], tensor completion [84], to name just a few.

Algorithm 1 The spectral method on the diagonal-deleted Gram matrix

```
Input: sampling set $\Omega$, observed entries $\{A_{i,j} \mid (i, j) \in \Omega\}$, sampling rate $p$, rank $r$.
Compute the (truncated) rank-$r$ eigendecomposition $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ of $\boldsymbol{G}$, where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{r \times r}$, and
    (3.3)    $\boldsymbol{G} := \mathcal{P}_{\text{off-diag}}\left(\frac{1}{p^2}\boldsymbol{A}\boldsymbol{A}^\top\right).$
    Here, $\boldsymbol{A}$ is defined in (2.5) and $\mathcal{P}_{\text{off-diag}}(\boldsymbol{M})$ zeros out the diagonal entries of $\boldsymbol{M}$.
Output $\boldsymbol{U}$ as the subspace estimate, and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{1/2}$ as the spectrum estimate.
```

3.2. *Theoretical guarantees.* In general, one can only hope to estimate $\boldsymbol{U}^\star$ up to global rotation. With this in mind, we introduce the following rotation matrix:

$$\boldsymbol{R} := \arg\min_{\boldsymbol{Q} \in \mathcal{O}^{r \times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{\mathrm{F}}. \tag{3.4}$$

In words, $\boldsymbol{R}$ is the global rotation matrix that best aligns $\boldsymbol{U}$ and $\boldsymbol{U}^\star$. Equipped with this notation, the following theorem delivers upper bounds on the difference between the obtained estimate $\boldsymbol{U}$ and the ground truth $\boldsymbol{U}^\star$. The proof is postponed to the Supplementary Material [13].

THEOREM 3.1. *Assume that the following conditions hold*:

$$p \ge c_0 \max\left\{\frac{\mu\kappa^4 r \log^2 d}{\sqrt{d_1 d_2}}, \frac{\mu\kappa^8 r \log^2 d}{d_2}\right\}, \tag{3.5a}$$

$$\frac{\sigma}{\sigma_r^\star} \le c_1 \min\left\{\frac{\sqrt{p}}{\kappa\sqrt[4]{d_1 d_2}\sqrt{\log d}}, \frac{1}{\kappa^3}\sqrt{\frac{p}{d_1 \log d}}\right\}, \tag{3.5b}$$

$$r \le c_2 \frac{d_1}{\mu_1 \kappa^4}, \tag{3.5c}$$

where $c_0 > 0$ is some sufficiently large constant and $c_1, c_2 > 0$ are some sufficiently small constants. Then with probability at least $1 - O(d^{-10})$, the matrices $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ returned by Algorithm 1 satisfy

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\| \lesssim \mathcal{E}_{\text{general}}, \tag{3.6a}$$

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \kappa^2\sqrt{\frac{\mu r}{d_1}} \cdot \mathcal{E}_{\text{general}}, \tag{3.6b}$$

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\| \lesssim \sigma_r^\star \cdot \mathcal{E}_{\text{general}}, \tag{3.6c}$$

where $\boldsymbol{R}$ is defined in (3.4), *and*

$$\begin{aligned}\mathcal{E}_{\text{general}} :=& \underbrace{\frac{\mu\kappa^2 r \log d}{\sqrt{d_1 d_2}\, p} + \sqrt{\frac{\mu\kappa^4 r \log d}{d_2 p}}}_{\text{missing data effect}} + \underbrace{\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1 d_2}\log d}{p} + \frac{\sigma\kappa}{\sigma_r^\star}\sqrt{\frac{d_1 \log d}{p}}}_{\text{noise effect}} \\ &+ \underbrace{\frac{\mu_1 \kappa^2 r}{d_1}}_{\text{diagonal deletion}}.\end{aligned} \tag{3.7}$$

REMARK 3.2. *If there is no missing data* (*i.e.*, $p = 1$), *then Theorem* 3.1 *holds unchanged if the first two terms on the right-hand side of* (3.7) *are removed.*

In a nutshell, Theorem 3.1 asserts that Algorithm 1 produces reliable estimates of the column subspace of $\boldsymbol{A}^\star$—with respect to both the spectral norm and the $\|\cdot\|_{2,\infty}$ norm—under certain conditions imposed on the sample size and the noise size. For instance, consider the settings where $\mu, \kappa \asymp 1$ and $r \ll d_1 \le d_2$. Then as long as the following condition holds:

$$p \gtrsim \frac{r \log^2 d}{\sqrt{d_1 d_2}} \quad \text{and} \quad \frac{\sigma^2}{\sigma_r^{\star 2}} = o\left(\frac{p}{\sqrt{d_1 d_2} \log d}\right), \tag{3.8}$$

the proposed spectral method achieves consistent estimation with high probability, namely,

$$\begin{aligned} &\min_{\boldsymbol{Q} \in \mathcal{O}^{r\times r}} \frac{\|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|}{\|\boldsymbol{U}^\star\|} = o(1), \qquad \min_{\boldsymbol{Q} \in \mathcal{O}^{r\times r}} \frac{\|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{2,\infty}}{\|\boldsymbol{U}^\star\|_{2,\infty}} = o(1), \\ &\frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|}{\|\boldsymbol{\Sigma}^\star\|} = o(1). \end{aligned} \tag{3.9}$$

Our upper bound (3.7) on the spectral norm error contains five terms. The first two terms of (3.7) are incurred by missing data; the third and the fourth terms of (3.7) represent the influence of observation noise; and the last term of (3.7) arises due to the bias caused by diagonal deletion. In particular, the last term is expected to be vanishingly small in the low-rank and incoherent case. Interestingly, both the missing data effect and the noise effect are captured by two different terms, which we shall interpret in what follows. Note that a primary focus of this paper is to demonstrate the feasibility of obtaining a tight control of the $\ell_{2,\infty}$ statistical error. This is particularly evident for the low-rank, incoherent and well-conditioned case with $r, \mu, \kappa = O(1)$, in which our theory (cf. (3.6a) and (3.6b)) reveals that the $\ell_{2,\infty}$ error can be a factor of $\sqrt{1/d_1}$ smaller than the spectral norm error. The discussion below focuses on this case (namely, $r, \mu, \kappa = O(1)$), with all logarithmic factors omitted for simplicity of presentation.

- Let us first examine the influence of observation noise, which reads

$$\frac{\sigma^2}{\sigma_r^{\star 2}} \frac{\sqrt{d_1 d_2}}{p} + \frac{\sigma}{\sigma_r^\star}\sqrt{\frac{d_1}{p}}. \tag{3.10}$$

  This contains a quadratic term as well as a linear term w.r.t. $\sigma/\sigma_r^\star$. To interpret this, consider, for example, the case without missing data (i.e., $p = 1$) and decompose

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\star\boldsymbol{A}^{\star\top} + \underbrace{\boldsymbol{A}^\star\boldsymbol{N}^\top + \boldsymbol{N}\boldsymbol{A}^{\star\top}}_{\text{linear perturbation}} + \underbrace{\boldsymbol{N}\boldsymbol{N}^\top}_{\text{quadratic perturbation}},$$

  which clearly explains why eigenspace perturbation bounds depend both linearly and quadratically on the noise magnitudes. In general, the quadratic term $\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1 d_2}}{p}$ is dominant when the signal-to-noise ratio (SNR) is not large enough; as the noise decreases to a sufficiently low level, the linear term starts to enter the picture. See Table 1 for a more precise summary. As we shall demonstrate momentarily, the terms (3.10) match the minimax limits (up to some logarithmic factor), meaning that it is generally impossible to get rid of either the linear term or the quadratic term.
- Next, we examine the influence of missing data and assume $\sigma = 0$ to simplify the discussion. If we view $\boldsymbol{N}_{\text{missing}} = \frac{1}{p}\boldsymbol{A} - \boldsymbol{A}^\star$ as a zero-mean perturbation matrix, then one can write

$$\frac{1}{p^2}\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\star\boldsymbol{A}^{\star\top} + \underbrace{\boldsymbol{A}^\star\boldsymbol{N}_{\text{missing}}^\top + \boldsymbol{N}_{\text{missing}}\boldsymbol{A}^{\star\top}}_{\text{linear perturbation}} + \underbrace{\boldsymbol{N}_{\text{missing}}\boldsymbol{N}_{\text{missing}}^\top}_{\text{quadratic perturbation}}.$$

TABLE 1

The dominant term of the noise effect in $\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1 d_2}}{p}+\frac{\sigma}{\sigma_r^{\star}}\sqrt{\frac{d_1}{p}}$ *if* $d_2 \geq d_1$ (*omitting logarithmic factors and assuming* $r, \kappa, \mu \asymp 1$)

	Large-noise (i.e., $\sigma/\sigma_r^\star \gtrsim \sqrt{p/d_2}$)	Small-noise (i.e., $\sigma/\sigma_r^\star \lesssim \sqrt{p/d_2}$)
Dominant term	$\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1 d_2}}{p}$	$\frac{\sigma}{\sigma_r^{\star}}\sqrt{\frac{d_1}{p}}$

Similar to the above noisy case with $p = 1$, this decomposition explains why the influence of missing data also contains two terms (see Table 2)

$$\underbrace{\frac{1}{\sqrt{d_1 d_2 p}}}_{\text{quadratic term in } 1/\sqrt{p}} + \underbrace{\frac{1}{\sqrt{d_2 p}}}_{\text{linear term in } 1/\sqrt{p}}.$$

Comparison with prior results. To demonstrate the effectiveness of our theory, we take a moment to compare them with several prior results. Once again, the discussion below focuses on the case with $\max\{\mu, \kappa, r\} \asymp 1$. To be fair, it is worth noting that most papers discussed below either have different objectives (e.g., aiming at matrix estimation rather than subspace estimation [22, 29, 65]), or work with different (and possibly more general) model assumptions (e.g., square matrices [4] or heteroskedastic noise [113]). Our purpose here is not to argue that our results are always stronger than the previous ones, but rather to point out the insufficiency of prior theory when directly applied to some basic settings.

- To begin with, we compare our spectral norm bound with that required for matrix completion [4, 22, 29, 36, 65] in the noise-free case (i.e., $\sigma = 0$), in order to show how much saving can be harvested when we move from matrix estimation to subspace estimation. Suppose that $d_2 \geq d_1$. As is well known, for both spectral and optimization-based methods, the sample complexities required for faithful matrix completion need to satisfy $pd_1 d_2 \gtrsim d_2 \text{poly} \log d$. In comparison, faithful estimation of the column subspace becomes feasible under the sample size $pd_1 d_2 \gtrsim \sqrt{d_1 d_2} \text{poly} \log d$, which can be much lower than that required for matrix completion (i.e., by a factor of $\sqrt{d_2/d_1}$). Further, we compare our $\|\cdot\|_{2,\infty}$ bound with the theory derived in [4] when $d_2 \gtrsim d_1 \log^2 d$. The theory in [4], Theorem 3.4, requires the sample size and the noise level to satisfy $p \gtrsim d_1^{-1} \log d$ and $\sigma/\sigma_r^\star \lesssim \sqrt{\frac{p}{d_2 \log d}}$, both of which are more stringent requirements than ours (namely, $p \gtrsim \frac{\log^2 d}{\sqrt{d_1 d_2}}$ and $\sigma/\sigma_r^\star \lesssim \frac{\sqrt{p}}{\sqrt[4]{d_1 d_2}\sqrt{\log d}}$). Again, this arises primarily because [4] seeks to estimate the whole matrix as opposed to its column subspace.
- We then compare our results with [84], which studies a diagonal-rescaling algorithm for the noise-free case (i.e., $\sigma = 0$). Combining [84], Theorem 6.2, with the standard Davis–Kahan matrix perturbation theory, we can easily see that their spectral norm bound for

TABLE 2

The dominant term of the missing data effect in $\frac{1}{\sqrt{d_1 d_2 p}}+\frac{1}{\sqrt{d_2 p}}$ *if* $d_2 \geq d_1$ (*omitting logarithmic factors and assuming* $r, \kappa, \mu \asymp 1$)

	High-missingness (i.e., $p \lesssim 1/d_1$)	Low-missingness (i.e., $p \gtrsim 1/d_1$)
Dominant term	$\frac{1}{\sqrt{d_1 d_2 p}}$	$\frac{1}{\sqrt{d_2 p}}$

subspace estimation reads $\frac{\text{poly}\log d}{\sqrt{d_1 d_2 p}} + \frac{\text{poly}\log d}{\sqrt{d_2 p}}$. This coincides with our bound except for the last term of (3.7) (due to the bias incurred by diagonal deletion). In comparison, our theory offers additional $\ell_{2,\infty}$ statistical guarantees and covers the noisy case, thus strengthening the theory presented in [84].

- Additionally, we compare our spectral norm bound with the results derived in [113]. Consider the noiseless case where $\sigma = 0$. It is proven in [113], Theorem 6, (see also the remark that follows) that: if the sample size satisfies $pd_1d_2 \gtrsim \max\{d_1^{1/3}d_2^{2/3}, d_1\}\text{poly}\log d$, then the HeteroPCA estimator is consistent in estimating the column subspace (namely, achieving a relative ℓ_2 estimation error not exceeding $o(1)$). In comparison, our theory claims that Algorithm 1 is guaranteed to yield consistent column subspace estimation as long as the sample size obeys $pd_1d_2 \gtrsim \sqrt{d_1d_2}\text{poly}\log d$. Consequently, if we omit logarithmic terms, then our sample complexity improves upon the theoretical support of HeteroPCA by a factor of $(d_2/d_1)^{1/6}$ if $d_2 \geq d_1$. Once again, the comparison here focuses on the effect of the aspect ratio d_2/d_1, without accounting for the influence of other parameters like μ, κ, r.

SVD applied directly to $\boldsymbol{A}$*?* Finally, another natural spectral method that comes immediately into mind is to compute the rank-r SVD of $\boldsymbol{A}$, and return the matrix containing the r left singular vectors as the column subspace estimate. The ℓ_2 risk analysis of this approach is typically based on classical matrix perturbation theory like Wedin's theorem [105]. We caution, however, that this approach becomes highly suboptimal when the aspect ratio d_2/d_1 grows. Take the case with Gaussian noise and no missing data (i.e., $p = 1$) for example: in order for Wedin's theorem to be applicable, a basic requirement is $\|\boldsymbol{N}\| < \sigma_r^\star$, which translates to the condition $\frac{\sigma}{\sigma_r^\star} \lesssim \frac{1}{\sqrt{d_2}}$ since $\|\boldsymbol{N}\| \lesssim \sigma\sqrt{d_2}$. In comparison, our theory covers the range $\frac{\sigma}{\sigma_r^\star} \lesssim \frac{1}{(d_1d_2)^{1/4}}$ (modulo some log factor), which allows the noise level to be $(d_2/d_1)^{1/4}$ times larger than the upper bound derived for the above SVD approach. The suboptimality of this approach can also be easily seen from numerical experiments as well; see Section 4 for details.

3.3. *Minimax lower bounds.*

It is natural to wonder whether our theoretical guarantees are tight, and whether there are other estimators that can potentially improve the performance of Algorithm 1. To answer these questions, we develop the following minimax lower bounds under Gaussian noise; the proof is deferred to Appendix 12.1.

THEOREM 3.3. *Suppose* $1 \leq r \leq d_1/2$, *and* $N_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. *Define*

$$\mathcal{M}^\star := \{\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2} \mid \mathsf{rank}(\boldsymbol{B}) = r,\ \sigma_r(\boldsymbol{B}) \in [0.9\sigma_r^\star, 1.1\sigma_r^\star]\}.$$

Denote by $\boldsymbol{U}(\boldsymbol{B}) \in \mathbb{R}^{d_1 \times r}$ *the matrix containing the* r *left singular vectors of* $\boldsymbol{B}$. *Then there exists some universal constant* $c_{\mathsf{lb}} > 0$ *such that*

$$\begin{aligned} &\inf_{\widehat{\boldsymbol{U}}} \sup_{\boldsymbol{A}^\star \in \mathcal{M}^\star} \mathbb{E}\Big[\min_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}(\boldsymbol{A}^\star)\|\Big] \\ &\qquad \geq c_{\mathsf{lb}} \min\left\{\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1d_2}}{p} + \frac{\sigma}{\sigma_r^\star}\sqrt{\frac{d_1}{p}}, 1\right\}, \end{aligned} \tag{3.11a}$$

$$\begin{aligned} &\inf_{\widehat{\boldsymbol{U}}} \sup_{\boldsymbol{A}^\star \in \mathcal{M}^\star} \mathbb{E}\Big[\min_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}(\boldsymbol{A}^\star)\|_{2,\infty}\Big] \\ &\qquad \geq c_{\mathsf{lb}} \min\left\{\frac{\sigma^2}{\sigma_r^{\star 2}}\frac{\sqrt{d_1d_2}}{p} + \frac{\sigma}{\sigma_r^\star}\sqrt{\frac{d_1}{p}}, 1\right\}\frac{1}{\sqrt{d_1}}, \end{aligned} \tag{3.11b}$$

where the infimum is taken over all estimators for $\boldsymbol{U}(\boldsymbol{A}^\star)$ *based on the observation* $\mathcal{P}_\Omega(\boldsymbol{A}^\star + \boldsymbol{N})$.

If we again consider the case where $r, \kappa, \mu \asymp 1$, then the above lower bounds (3.11) match the noise effect terms in Theorem 3.1 (or equivalently, (3.10)) up to logarithmic factors. This unveils a fundamental reason why the linear and the quadratic terms in (3.10) are both essential in determining the estimation risk.

Another information-theoretic limit that concerns only the influence of subsampling is supplied as follows; the proof is postponed to Appendix 12.2.

THEOREM 3.4. *Suppose $d_1 \le d_2$ and $p < \frac{1-\epsilon}{\sqrt{d_1 d_2}}$ for any small constant $0 < \epsilon < 1$. With probability approaching one, there exist unit vectors $\boldsymbol{u}^\star, \widetilde{\boldsymbol{u}}^\star \in \mathbb{R}^{d_1}$ and $\boldsymbol{v}^\star, \widetilde{\boldsymbol{v}}^\star \in \mathbb{R}^{d_2}$ such that:*

- $\min \|\boldsymbol{u}^\star \pm \widetilde{\boldsymbol{u}}^\star\|_2 \asymp 1$ *and* $\|\boldsymbol{u}^\star \boldsymbol{v}^{\star\top} - \widetilde{\boldsymbol{u}}^\star \widetilde{\boldsymbol{v}}^{\star\top}\|_{\mathrm{F}} \asymp 1$;
- *one cannot distinguish $\boldsymbol{u}^\star \boldsymbol{v}^{\star\top}$ and $\widetilde{\boldsymbol{u}}^\star \widetilde{\boldsymbol{v}}^{\star\top}$ from the entries in Ω, that is, $\mathcal{P}_\Omega(\boldsymbol{u}^\star \boldsymbol{v}^{\star\top}) = \mathcal{P}_\Omega(\widetilde{\boldsymbol{u}}^\star \widetilde{\boldsymbol{v}}^{\star\top})$.*

In words, Theorem 3.4 asserts that one cannot hope to achieve consistent subspace estimation (in the sense of (3.9)) at all, as soon as the sampling rate p falls below the threshold $1/\sqrt{d_1 d_2}$. Putting Theorems 3.3–3.4 together reveals that: consistent estimation can by no means be guaranteed unless

$$p \gtrsim \frac{1}{\sqrt{d_1 d_2}} \quad \text{and} \quad \frac{\sigma^2}{\sigma_r^{\star 2}} \lesssim \frac{p}{\sqrt{d_1 d_2}}, \tag{3.12}$$

which agrees with our theoretical guarantees (3.8) (up to some logarithmic term). As a result, our minimax lower bounds confirm the near optimality of Algorithm 1 in enabling consistent estimation.

On the other hand, it is widely recognized that spectral methods are typically unable to achieve exact recovery or optimal estimation accuracy in the presence of missing data, even in the balanced case with $d_1 = d_2$. For instance, if there is no noise, namely $\sigma = 0$, the spectral methods fail to achieve perfect recovery as long as $p < 1$ (basically the first two terms of (3.7) do not vanish) [66], whereas exact recovery might sometimes be feasible with the aid of optimization-based approaches [21]. More often than not, spectral methods are employed to produce a rough initial estimate that outperforms the random guess, which can then be refined via other algorithms (e.g., nonconvex optimization algorithms like gradient descent and alternating minimization [14, 66, 80, 102]).

4. Consequences for concrete applications.

We showcase the consequence of Theorem 3.1 in three concrete applications previously introduced in Section 1 in relatively simple settings. Rather than striving for full generality, our purpose is to highlight the broad applicability of our main results.

4.1. *Noisy tensor completion.*

Problem settings. We begin by considering the problem of symmetric tensor completion. Consider an unknown order-3 tensor

$$\boldsymbol{T}^\star = \sum_{s=1}^{r} \boldsymbol{w}_s^\star \otimes \boldsymbol{w}_s^\star \otimes \boldsymbol{w}_s^\star := \sum_{s=1}^{r} (\boldsymbol{w}_s^\star)^{\otimes 3} \in \mathbb{R}^{d \times d \times d},$$

with canonical polyadic (CP) rank r. The goal is to estimate the subspace spanned by $\{\boldsymbol{w}_s^\star\}_{s=1}^r$, based on the noisy tensor $\boldsymbol{T} = [T_{i,j,k}]_{1 \le i,j,k \le d}$ obeying

$$T_{i,j,k} = \begin{cases} T_{i,j,k}^\star + N_{i,j,k}, & (i,j,k) \in \Omega, \\ 0, & (i,j,k) \notin \Omega. \end{cases} \tag{4.1}$$

Algorithm 2 The spectral method for tensor completion

```
Input: sampling set $\Omega$, observed entries $\{T_{i,j,k} \mid (i,j,k) \in \Omega\}$, sampling rate $p$, CP-rank $r$.
Let $\boldsymbol{A} \in \mathbb{R}^{d \times d^2}$ be the mode-1 matricization of the observed tensor $\boldsymbol{T}$ (see (4.1)), namely, set $A_{i,(j-1)d+k} = T_{i,j,k}$ for each $(i,j,k) \in [d]^3$, and employ $\boldsymbol{A}$ as the input of Algorithm 1.
Output $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ returned by Algorithm 1 as the subspace estimate.
```

Here, $T_{i,j,k}$ is the observed entry in location (i, j, k), $N_{i,j,k}$ is the associated independent random noise satisfying Assumption 2.3, and $\Omega \subseteq [d]^3$ stands for a sampling set obtained via uniform random sampling with sampling rate p (namely, each entry is observed independently with probability p).

Algorithm. Observe that the mode-1 matricization of $\boldsymbol{T}^\star$ is given by[2]

$$\boldsymbol{A}^\star = \sum_{s=1}^{r} \boldsymbol{w}_s^\star (\boldsymbol{w}_s^\star \otimes \boldsymbol{w}_s^\star)^\top \in \mathbb{R}^{d \times d^2}, \tag{4.2}$$

indicating that the column subspace of $\boldsymbol{A}^\star$ is essentially the subspace spanned by the tensor factors $\{\boldsymbol{w}_s^\star\}_{s=1}^r$. Therefore, if we denote by $\boldsymbol{A} \in \mathbb{R}^{d \times d^2}$ the mode-1 matricization of $\boldsymbol{T}$, then we can invoke our general spectral method to estimate the column subspace of $\boldsymbol{A}^\star$ given $\boldsymbol{A}$. This procedure is summarized in Algorithm 2.

Theoretical guarantees. In order to provide theoretical support for Algorithm 2, we introduce a few more notation. First, we introduce

$$\kappa_{\mathsf{tc}} := \lambda_{\max}^\star / \lambda_{\min}^\star, \qquad \lambda_{\min}^\star := \min_{1 \le i \le r} \|\boldsymbol{w}_i^\star\|_2^3, \qquad \lambda_{\max}^\star := \max_{1 \le i \le r} \|\boldsymbol{w}_i^\star\|_2^3. \tag{4.3}$$

Note that $\|\boldsymbol{w}_i^\star\|_2^3$ is precisely the Frobenius norm of the rank-1 tensor $\boldsymbol{w}_i^{\star \otimes 3}$—the ith tensor component. Informally, κ_{tc} captures the condition number of the unknown tensor. Additionally, similar to matrix completion, we introduce the following incoherence definitions that enable efficient tensor completion:

DEFINITION 4.1 (Incoherence). *Define the incoherence parameters μ_3, μ_4, μ_5 for the tensor $\boldsymbol{T}^\star$ and its tensor factors $\{\boldsymbol{w}_s^\star\}_{s=1}^r$ as follows*:

$$\mu_3 := \frac{d^3 \|\boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2}, \qquad \mu_4 := \max_{1 \le i \le r} \frac{d \|\boldsymbol{w}_i^\star\|_\infty^2}{\|\boldsymbol{w}_i^\star\|_2^2}, \qquad \mu_5 := \max_{1 \le i \ne j \le r} \frac{d \langle \boldsymbol{w}_i^\star, \boldsymbol{w}_j^\star \rangle^2}{\|\boldsymbol{w}_i^\star\|_2^2 \|\boldsymbol{w}_j^\star\|_2^2}. \tag{4.4}$$

For notational convenience, we also set

$$\mu_{\mathsf{tc}} := \max\{\mu_{3,} \mu_4^2\}. \tag{4.5}$$

Given that the tensor factors $\{\boldsymbol{w}_s^\star\}_{1 \le s \le r}$ are in general not orthogonal to each other, we introduce the following orthonormal matrix $\boldsymbol{U}^\star \in \mathbb{R}^{d \times r}$ to represent the subspace spanned by $\{\boldsymbol{w}_s^\star\}_{1 \le s \le r}$:

$$\boldsymbol{U}^\star := \boldsymbol{W}^\star (\boldsymbol{W}^{\star\top} \boldsymbol{W}^\star)^{-1/2}, \qquad \boldsymbol{W}^\star := [\boldsymbol{w}_1^\star, \ldots, \boldsymbol{w}_r^\star] \in \mathbb{R}^{d \times r}. \tag{4.6}$$

[2] We let $\boldsymbol{a} \otimes \boldsymbol{b} := \begin{bmatrix} a_1 \boldsymbol{b} \\ \vdots \\ a_d \boldsymbol{b} \end{bmatrix}$ represent a d^2-dimensional vector.

Note that the particular choice of $\boldsymbol{U}^\star$ in (4.6) is not pivotal, and can be replaced by any $d_1 \times r$ orthonormal matrix that spans the same column space as $\boldsymbol{W}^\star$. With these in place, we are now ready to quantify the estimation error of this spectral algorithm. The proof is deferred to Appendix 9.1.

COROLLARY 4.2 (Symmetric tensor completion). *Consider the above tensor completion model. There exist some universal constants* $c_0, c_1, c_2 > 0$ *such that if*

$$p \geq c_0 \max\left\{\frac{\mu_{\mathsf{tc}}\kappa_{\mathsf{tc}}^4 r \log^2 d}{d^{3/2}}, \frac{\mu_{\mathsf{tc}}\kappa_{\mathsf{tc}}^8 r \log^2 d}{d^2}\right\}, \tag{4.7a}$$

$$\frac{\sigma}{\lambda_{\min}^\star} \leq c_1 \min\left\{\frac{\sqrt{p}}{\kappa_{\mathsf{tc}} d^{3/4}\sqrt{\log d}}, \frac{1}{\kappa_{\mathsf{tc}}^3}\sqrt{\frac{p}{d\log d}}\right\}, \tag{4.7b}$$

$$r \leq c_2 \min\left\{\frac{d}{\kappa_{\mathsf{tc}}^4 \mu_4}, \frac{1}{\kappa_{\mathsf{tc}}^2}\sqrt{\frac{d}{\mu_5}}\right\}, \tag{4.7c}$$

then with probability exceeding $1 - O(d^{-10})$, *Algorithm* 2 *yields*

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\| \lesssim \mathcal{E}_{\mathsf{tc}}, \tag{4.8a}$$

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \kappa_{\mathsf{tc}}^2\sqrt{\frac{\mu_{\mathsf{tc}} r}{d}} \cdot \mathcal{E}_{\mathsf{tc}}, \tag{4.8b}$$

where $\boldsymbol{R} := \arg\min_{\boldsymbol{Q}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{\mathrm{F}}$ *and*

$$\mathcal{E}_{\mathsf{tc}} := \frac{\mu_{\mathsf{tc}}\kappa_{\mathsf{tc}}^2 r \log d}{d^{3/2} p} + \sqrt{\frac{\mu_{\mathsf{tc}}\kappa_{\mathsf{tc}}^4 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda_{\min}^{\star 2}}\frac{d^{3/2}\log d}{p} + \frac{\sigma\kappa_{\mathsf{tc}}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} + \frac{\mu_4\kappa_{\mathsf{tc}}^2 r}{d}. \tag{4.9}$$

As discussed in several related work (e.g., [14, 59, 84, 106, 107]), once we obtain reliable estimates of the subspace spanned by the tensor factors, we can further exploit the tensor structure to estimate the unknown tensor. Indeed, in many tensor completion algorithms, subspace estimation serves as a crucial initial step for tensor completion. Moreover, while prior works only provide ℓ_2 estimation error bounds, Corollary 4.2 further delivers $\ell_{2,\infty}$ statistical guarantees, which reflect a stronger sense of statistical accuracy. We note that [108], Theorem 4, derived an appealing $\ell_{2,\infty}$ statistical error bound for an algorithm called HOSVD, under the tensor denoising setting. In comparison to the Gaussian noise considered therein, our results accommodate the case with missing data and possibly spiky noise.

Implications. In what follows, we discuss the sample size and the signal-to-noise (SNR) required for achieving consistent tensor estimation (namely, obtaining an $o(1)$ relative estimation error). For convenience of presentation, we again focus on the low-rank, incoherent, and well-conditioned case with $r, \mu, \kappa_{\mathsf{tc}} \asymp 1$. In this case, our results in Corollary 4.2 indicate that

$$\min_{\boldsymbol{Q}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\| = o(1), \qquad \min_{\boldsymbol{Q}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{2,\infty} = o(1/\sqrt{d}) \tag{4.10}$$

with high probability, provided that the sample size and the noise satisfy

$$p \gtrsim \frac{\log^2 d}{d^{3/2}} \quad \text{and} \quad \frac{\sigma}{\lambda_{\min}^\star} = o\left(\sqrt{\frac{p}{d^{3/2}\log d}}\right). \tag{4.11}$$

Several remarks are in order.

- *Sample complexity.* It is widely conjectured that the sample complexity pd^3 required to reconstruct a order-3 tensor in polynomial time—even in the noiseless case—is at least $d^{3/2}$ (or equivalently, $p \gtrsim 1/d^{3/2}$) [11, 84, 107]. Therefore, our theory reveals that spectral methods achieve consistent estimation (w.r.t. both $\|\cdot\|$ and $\|\cdot\|_{2,\infty}$), as long as the sample size is slightly above the (conjectured) computational limit. Moreover, it is easily seen that the bias incurred by deleting the diagonal is much smaller than the error due to missing data, which justifies the rationale that diagonal deletion does not harm the performance by much.
- *Noise level.* It is easily seen that the maximum magnitude of the entries of $\boldsymbol{T}^\star$ in this case is $\|\boldsymbol{T}^\star\|_\infty \asymp \lambda^\star_{\max}/d^{3/2}$. Thus, the noise condition in (4.11) is equivalent to

  $$\frac{\sigma}{\|\boldsymbol{T}^\star\|_\infty} \lesssim \sqrt{pd^{3/2}}.$$

  Taken together with our sample size requirement $p \gtrsim \frac{\log^2 d}{d^{3/2}}$, this condition allows the noise magnitude in each observed entry to significantly exceed the size of the corresponding entry, which covers a broad range of scenarios of practical interest. In addition, in the fully-observed case (i.e., $p = 1$) with i.i.d. Gaussian noise, the authors in [114] showed that the noise size condition (4.11)—up to some log factor—is necessary for any polynomial-time algorithm to achieve consistent estimation, provided that a certain hypergraphic planted clique conjecture holds.

Finally, we remark that in the fully-observed case (i.e., $p = 1$) with i.i.d. Gaussian noise, it can be seen from [114], Theorem 1, that (4.8a) is suboptimal; in fact, the minimax risk consists only of the linear term in σ (namely, $\frac{\sigma}{\lambda^\star_{\min}}\sqrt{d}$, if we omit log factors and assume $r, \mu_{\mathsf{tc}}, \mu_5, \kappa \asymp 1$). This is a typical drawback of the spectral method for tensor estimation, since it falls short of exploiting the low-complexity structure in the row subspace. However, the spectral estimate offers a reasonably good initial estimate for this problem, and one can often employ optimization-based iterative refinement paradigms (like gradient descent [14]) to obtain minimax optimal estimates.

4.2. *PCA with missing data.*

Model and algorithm. Next, we study covariance estimation with missing data, as previously introduced in Section 1. For concreteness, imagine a set of independent sample vectors obeying

$$\boldsymbol{x}_i = \boldsymbol{B}^\star \boldsymbol{f}_i^\star + \boldsymbol{\eta}_i \in \mathbb{R}^d, \qquad \boldsymbol{f}_i^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r), \quad 1 \le i \le n.$$

Here, $\boldsymbol{B}^\star \in \mathbb{R}^{d\times r}$ encodes the r-dimensional principal subspace underlying the data (sometimes referred to as the factor loading matrix in factor models [44, 72]), $\boldsymbol{f}_i^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r)$ represents some random coefficients, and the noise vector $\boldsymbol{\eta}_i = [\eta_{i,j}]_{1\le j\le d}$ consists of independent Gaussian components[3] obeying

$$\mathbb{E}[\eta_{i,j}] = 0 \quad \text{and} \quad \mathsf{Var}[\eta_{i,j}] \le \sigma^2.$$

What we observe is a partial set of entries of $\boldsymbol{x}_i = [x_{i,j}]_{1\le j\le d}$, namely, we only observe $x_{i,j}$ for any $(i, j) \in \Omega$, where Ω is obtained by random sampling with rate p. The goal is to estimate the subspace spanned by $\boldsymbol{B}^\star$, or even $\boldsymbol{B}^\star\boldsymbol{B}^{\star\top}$.

If we write $\boldsymbol{F}^\star = [\boldsymbol{f}_1^\star, \dots, \boldsymbol{f}_n^\star] \in \mathbb{R}^{r\times n}$ and $\boldsymbol{N} = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n] \in \mathbb{R}^{d\times n}$, then it boils down to estimating the column space of $\boldsymbol{A}^\star := \boldsymbol{B}^\star\boldsymbol{F}^\star$ from the data $\mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{B}^\star\boldsymbol{F}^\star + \boldsymbol{N})$. Our spectral method for covariance estimation is summarized in Algorithm 3.

[3] Here, we assume $\boldsymbol{f}_i^\star$ and $\boldsymbol{\eta}_i$ to be Gaussian for simplicity of presentation. The results in this subsection continue to hold if they are sub-Gaussian random vectors.

Algorithm 3 The spectral method for covariance estimation

```
Input: sampling set $\Omega$, observed entries $\{X_{i,j} \mid (i,j) \in \Omega\}$, sampling rate $p$, rank $r$.
Let $\boldsymbol{A} = \mathcal{P}_{\Omega}(\boldsymbol{X}) \in \mathbb{R}^{d\times n}$ with $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$, and use $\boldsymbol{A}$ as the input of Algorithm 1. Let $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ be the estimates returned by Algorithm 1, and set $\boldsymbol{B} := \frac{1}{\sqrt{n}}\boldsymbol{U}\boldsymbol{\Sigma}$.
Output $\boldsymbol{U}$ as the subspace estimate and $\boldsymbol{S} := \boldsymbol{B}\boldsymbol{B}^{\top}$ as the covariance estimate.
```

Theoretical guarantees. In order to present our theory, we make a few more definitions. Without loss of generality, we shall define

$$\boldsymbol{S}^{\star} := \boldsymbol{B}^{\star}\boldsymbol{B}^{\star\top} = \boldsymbol{U}^{\star}\boldsymbol{\Lambda}^{\star}\boldsymbol{U}^{\star\top} \quad \text{and} \quad \boldsymbol{B}^{\star} = \boldsymbol{U}^{\star}\boldsymbol{\Lambda}^{\star 1/2}, \tag{4.12}$$

where $\boldsymbol{U}^{\star} \in \mathbb{R}^{d\times r}$ consists of orthonormal columns and $\boldsymbol{\Lambda}^{\star} = \mathsf{diag}(\lambda_1^{\star}, \dots, \lambda_r^{\star}) \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\lambda_1^{\star} \ge \cdots \ge \lambda_r^{\star} \ge 0$. We also define the condition number and the incoherence parameter as

$$\kappa_{\mathsf{ce}} := \lambda_1^{\star}/\lambda_r^{\star} \quad \text{and} \quad \mu_{\mathsf{ce}} := \frac{d}{r}\|\boldsymbol{U}^{\star}\|_{2,\infty}^2. \tag{4.13}$$

We are now positioned to derive statistical estimation guarantees using our general theorem. The following result is a consequence of Theorem 3.1; the proof is postponed to Appendix 9.2.

COROLLARY 4.3 (Covariance estimation). *Consider the above covariance estimation model with missing data. There exist universal constants $c_0, c_1 > 0$ such that if $r \le c_1 \frac{d}{\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}^2}$ and*

$$\begin{aligned} n \ge c_0 \max\Bigg\{ & \frac{\mu_{\mathsf{ce}}^2\kappa_{\mathsf{ce}}^6 r^2 \log^6(n+d)}{dp^2}, \frac{\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}^5 r \log^3(n+d)}{p}, \\ & \frac{\sigma^4}{\lambda_r^{\star 2}}\frac{\kappa_{\mathsf{ce}}^2 d\log^2(n+d)}{p^2}, \frac{\sigma^2}{\lambda_r^{\star}}\frac{\kappa_{\mathsf{ce}}^3 d\log(n+d)}{p}\Bigg\}, \end{aligned} \tag{4.14}$$

then with probability exceeding $1 - O((n+d)^{-10})$, Algorithm 3 yields

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^{\star}\| \lesssim \mathcal{E}_{\mathsf{ce}}, \tag{4.15a}$$

$$\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^{\star}\|_{2,\infty} \lesssim \kappa_{\mathsf{ce}}^{3/2}\sqrt{\frac{\mu_{\mathsf{ce}} r\log(n+d)}{d}} \cdot \mathcal{E}_{\mathsf{ce}}, \tag{4.15b}$$

$$\|\boldsymbol{S} - \boldsymbol{S}^{\star}\| \lesssim \kappa_{\mathsf{ce}}\lambda_1^{\star} \cdot \mathcal{E}_{\mathsf{ce}}, \tag{4.15c}$$

$$\|\boldsymbol{S} - \boldsymbol{S}^{\star}\|_{\infty} \lesssim \frac{\kappa_{\mathsf{ce}}\mu_{\mathsf{ce}} r\log(n+d)}{d}\lambda_1^{\star} \cdot \mathcal{E}_{\mathsf{ce}}. \tag{4.15d}$$

Here, $\boldsymbol{R} := \arg\min_{\boldsymbol{Q}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^{\star}\|_{\mathrm{F}}$ *and*

$$\begin{aligned} \mathcal{E}_{\mathsf{ce}} := & \frac{\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}^2 r\log^2(n+d)}{\sqrt{dn}p} + \sqrt{\frac{\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}^3 r\log^2(n+d)}{np}} \\ & + \frac{\sigma^2}{\lambda_r^{\star}}\sqrt{\frac{d}{n}}\frac{\log(n+d)}{p} + \frac{\sigma}{\sqrt{\lambda_r^{\star}}}\sqrt{\frac{d}{n}}\sqrt{\frac{\kappa_{\mathsf{ce}}\log(n+d)}{p}} + \frac{\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}} r}{d}. \end{aligned} \tag{4.16}$$

REMARK 4.4. *We make note of a scaling issue that one shall bear in mind when comparing this result with our main theorem. In the settings of Theorem 3.1, the singular values $\{\sigma_i^\star\}_{i=1}^r$ of the truth $\boldsymbol{A}^\star$ do not change as the column dimension d_2 grows. In contrast, in the settings of Corollary 4.3, the singular values of the sample covariance matrix keep growing as we collect more sample vectors, which is equivalent to saying that these singular values scale with the column dimension.*

Discussion. To facilitate interpretation, let us again focus on the case where $\mu_{\mathsf{ce}}, \kappa_{\mathsf{ce}} \asymp 1$. Corollary 4.3 demonstrates that for any given sampling rate p, we can achieve consistent estimation[4] as long as the number n of samples satisfies

$$n \gtrsim \max\left\{\frac{r^2}{dp^2}, \frac{r}{p}, \frac{\sigma^4 d}{\lambda_r^{\star 2} p^2}, \frac{\sigma^2 d}{\lambda_r^\star p}\right\} \text{poly} \log d. \tag{4.17}$$

Throughout this subsection, the sample size refers to n—the number of sample vectors $\{\boldsymbol{x}_i\}_{1 \le i \le n}$ available.

Next, we compare our $\ell_{2,\infty}$ bounds with several prior work for the case with $r \asymp 1$. We emphasize again that the foci and model assumptions of these prior papers might be quite different from ours (e.g., [117] is able to accommodate inhomogeneous sampling patterns), and the advantages of our results discussed below are restricted to the settings considered in this paper. For simplicity, we ignore all log factors.

- Suppose that $\sigma = 1$. In this setting, [117], Theorem 4, demonstrates that if

$$n \gtrsim \max\left\{\frac{1}{p^2}, \frac{d^2}{\lambda_r^{\star 2} p^2}, \frac{d}{\lambda_r^\star p^2}\right\} \text{poly} \log d,$$

  then with high probability one has

$$\min_{\boldsymbol{Q} \in \mathcal{O}^{r \times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{p\sqrt{n}}\left(\frac{1}{\sqrt{\lambda_r^\star}} + \frac{1}{\lambda_r^\star}\right)\left(1 + \frac{\sqrt{d}}{\lambda_r^\star}\right) \text{poly} \log d$$

  In comparison, our sample size requirement for consistent estimation improves upon [117], Theorem 4, by a factor of $\min\{d, p^{-1}\}$. Moreover, our estimation error bound improves upon [117], Theorem 4, by a factor of $\min\{\sqrt{\lambda_r^\star}, \frac{1}{\sqrt{p}}\}$ if $\sqrt{d} \ll \lambda_r^\star \lesssim d$, by a factor of $\frac{\sqrt{d}}{\lambda_r^\star}$ when $\lambda_r^\star \lesssim 1$, and by a factor of $\min\{\frac{\sqrt{d}}{\lambda_r^\star \sqrt{p}}, \sqrt{\frac{d}{\lambda_r^\star}}\}$ if $1 \ll \lambda_r^\star \lesssim \sqrt{d}$.
- In the absence of missing data, the $\ell_{2,\infty}$ error bound presented in [23], Theorem 1.1, reads (ignoring logarithmic terms)

$$\min_{\boldsymbol{Q} \in \mathcal{O}^{r \times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \begin{cases} \sqrt{\dfrac{1}{nd}} & \text{for } \dfrac{\sigma}{\sqrt{\lambda_r^\star}} \lesssim \dfrac{1}{\sqrt{d}}, \\ \dfrac{\sigma^2}{\lambda_r^\star}\sqrt{\dfrac{d}{n}} & \text{for } \dfrac{1}{\sqrt{d}} \ll \dfrac{\sigma}{\sqrt{\lambda_r^\star}} \lesssim 1. \end{cases}$$

  Consequently, our result improves upon the above error bound by a factor of $\frac{\sigma\sqrt{d}}{\sqrt{\lambda_r^\star}}$ if $\frac{1}{\sqrt{d}} \ll \frac{\sigma}{\sqrt{\lambda_r^\star}} \lesssim 1$, while being able to handle the case with larger noise (namely, $\frac{\sigma}{\sqrt{\lambda_r^\star}} \gg 1$).

[4] Here, consistent estimation is declared if $\min_{\boldsymbol{Q} \in \mathcal{O}^{r \times r}} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\| = o(1)$ and $\|\boldsymbol{S} - \boldsymbol{S}^\star\| = o(\lambda_r^\star)$.

4.3. *Community recovery in bipartite stochastic block models.* As it turns out, if we denote by $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{V}|}$ the biadjacency matrix of the observed random bipartite graph or its centered version, then $\boldsymbol{A}^\star := \mathbb{E}[\boldsymbol{A}]$ exhibits a low-rank structure (as we shall elaborate momentarily). Perhaps more importantly, the column subspace of $\boldsymbol{A}^\star$ reveals the community memberships of all nodes in $\mathcal{U}$. As a result, this biclustering problem is tightly connected to subspace estimation given noisy observations of a low-rank matrix. In particular, when the size of $\mathcal{V}$ is substantially larger than that of $\mathcal{U}$, one might encounter a situation where only the nodes in $\mathcal{U}$ (rather than those in $\mathcal{V}$) can be reliably clustered. This calls for development of "one-sided" community recovery algorithms, that is, the type of algorithms that guarantee reliable clustering of $\mathcal{U}$ without worrying about the clustering accuracy in $\mathcal{V}$.

Model. This subsection investigates the problem of biclustering, by considering a bipartite stochastic block model (BSBM) with two disjoint groups of nodes $\mathcal{U}$ and $\mathcal{V}$. Suppose that the nodes in $\mathcal{U}$ (resp., $\mathcal{V}$) form two clusters. For each pair of nodes $(i, j) \in (\mathcal{U}, \mathcal{V})$, there is an edge connecting them with probability depending only on the community memberships of i and j. To be more specific:

- *Biclustering structure.* Consider two disjoint collections of nodes $\mathcal{U}$ and $\mathcal{V}$, which are of size n_u and n_v, respectively. Suppose that each collection of nodes can be clustered into two communities. To be more precise, let $\mathcal{I}_1 \subseteq \mathcal{U}$ and $\mathcal{I}_2 = \mathcal{U}\backslash\mathcal{I}_1$ (resp., $\mathcal{J}_1 \subseteq \mathcal{V}$ and $\mathcal{J}_2 = \mathcal{V}\backslash\mathcal{J}_1$) be two nonoverlapping communities in $\mathcal{U}$ (resp. $\mathcal{V}$) that contain $n_u/2$ (resp. $n_v/2$) nodes each. Without loss of generality, we assume that $\mathcal{I}_1$ contains the first $n_u/2$ nodes of $\mathcal{U}$, and $\mathcal{J}_1$ contains the first $n_v/2$ nodes of $\mathcal{V}$; these are of course *a priori* unknown.
- *Measurement model.* What we observe is a random bipartite graph generated based on the community memberships of the nodes. In the simplest version of BSBMs, a pair of nodes $(i, j) \in (\mathcal{U}, \mathcal{V})$ is connected by an edge independently with probability q_{in} if either $(i, j) \in (\mathcal{I}_1, \mathcal{J}_1)$ or $(i, j) \in (\mathcal{I}_2, \mathcal{J}_2)$ holds, and with probability q_{out} otherwise. Here, $0 \le q_{\mathsf{out}} \le q_{\mathsf{in}} \le 1$ represent the edge densities. If we denote by $\boldsymbol{C} \in \{0, 1\}^{n_u\times n_v}$ the biadjacency matrix of this random bipartite graph, then one has
$$\mathbb{P}\{C_{i,j} = 1\} \overset{\text{ind.}}{=} \begin{cases} q_{\mathsf{in}} & \text{if } (i, j) \in (\mathcal{I}_1, \mathcal{J}_1) \text{ or } (i, j) \in (\mathcal{I}_2, \mathcal{J}_2), \\ q_{\mathsf{out}} & \text{otherwise.} \end{cases}$$

Our goal is to recover the community memberships of the nodes in $\mathcal{U}$, based on the above random bipartite graph. In what follows, we define

$$n := n_u + n_v, \tag{4.18}$$

and declare exact community recovery of $\mathcal{U}$ if the partition of the nodes returned by our algorithm coincides precisely with the true partition $(\mathcal{I}_1, \mathcal{I}_2)$.

While our theory covers a broad range of n_u and n_v, we emphasize the case where $n_v \gg n_u$ (namely, $\mathcal{V}$ contains far more nodes than $\mathcal{U}$). In such a case, it is not uncommon to encounter a situation where one can only hope to recover the community memberships of the nodes in $\mathcal{U}$ but not those in $\mathcal{V}$.

Algorithm. To attempt community recovery, we look at a centered version of the biadjacency matrix[5]

$$\boldsymbol{A} := \boldsymbol{C} - \frac{q_{\mathsf{in}} + q_{\mathsf{out}}}{2}\mathbf{1}_{n_u}\mathbf{1}_{n_v}^\top. \tag{4.19}$$

[5] Here, we assume prior knowledge about q_{in} and q_{out}. Otherwise, the quantity $\frac{q_{\mathsf{in}}+q_{\mathsf{out}}}{2}$ can also be easily estimated.

Algorithm 4 The spectral method for BSBM

```
Input: observed biadjacency matrix $\boldsymbol{C}$, edge probabilities $q_{\mathsf{in}}, q_{\mathsf{out}}$.
Employ $\boldsymbol{A}$ (cf. (4.19)) as the input of Algorithm 1, and let $\boldsymbol{u} = [u_i] \in \mathbb{R}^{n_u}$ be the output returned by Algorithm 1 (which serves as the estimate of the leading left singular subspace of $\boldsymbol{A}^\star$.
Output: for any $i \in \mathcal{U}$, we claim that $i$ belongs to the first community if $u_i > 0$, and the second community otherwise.
```

Recognizing that

$$\begin{aligned} \boldsymbol{A}^\star := \mathbb{E}[\boldsymbol{A}] &= \frac{q_{\mathsf{in}} - q_{\mathsf{out}}}{2} \begin{bmatrix} \mathbf{1}_{n_u/2}\mathbf{1}_{n_v/2}^\top, & -\mathbf{1}_{n_u/2}\mathbf{1}_{n_v/2}^\top \\ -\mathbf{1}_{n_u/2}\mathbf{1}_{n_v/2}^\top, & \mathbf{1}_{n_u/2}\mathbf{1}_{n_v/2}^\top \end{bmatrix} \\ &= \frac{q_{\mathsf{in}} - q_{\mathsf{out}}}{2} \begin{bmatrix} \mathbf{1}_{n_u/2} \\ -\mathbf{1}_{n_u/2} \end{bmatrix} [\mathbf{1}_{n_v/2}^\top, -\mathbf{1}_{n_v/2}^\top], \end{aligned} \tag{4.20}$$

we see that the leading singular vectors of $\boldsymbol{A}^\star$ reveals the community memberships of all nodes. Motivated by this observation, our algorithm for recovering the community memberships in $\mathcal{U}$ proceeds as follows:

Theoretical guarantees and implications. We are now ready to invoke our general theory to demonstrate the effectiveness of the above algorithm, as asserted by the following result.

COROLLARY 4.5 (Bipartite stochastic block model). *Consider the above bipartite stochastic block model. There exists some universal constant* $c_0 > 0$ *such that if*

$$\frac{(q_{\mathsf{in}} - q_{\mathsf{out}})^2}{q_{\mathsf{in}}} \geq c_0 \max\left\{\frac{\log n}{\sqrt{n_u n_v}}, \frac{\log n}{n_v}\right\}, \tag{4.21}$$

then Algorithm 4 achieves exact community recovery of $\mathcal{U}$ *with probability exceeding* $1 - O(n^{-10})$.

We then take a moment to discuss the implications of Corollary 4.5. For simplicity of presentation, we shall focus on the scenario with $q_{\mathsf{in}} \asymp q_{\mathsf{out}} = o(1)$ and $n_u \leq n_v$.

- *Exact recovery via the spectral method alone.* Consider the following sparse regime, where

  $$q_{\mathsf{in}} = \frac{a \log n}{\sqrt{n_u n_v}} \quad \text{and} \quad q_{\mathsf{out}} = \frac{b \log n}{\sqrt{n_u n_v}}$$

  for some absolute positive constants $a \geq b$. Corollary 4.5 demonstrates that we can achieve exact recovery when $\frac{(a-b)^2}{a} \gtrsim 1$. This improves upon prior results presented in [47]. More specifically, the results in [47] only guaranteed *almost* exact recovery of community memberships (namely, obtaining correct community memberships for a fraction $1 - o(1)$ of the nodes). In comparison, our results assert that the spectral estimates alone are sufficient to reveal exact community memberships for all nodes in $\mathcal{U}$; there is no need to invoke further refinement procedures to clean up the remaining errors.
- *Near optimality.* In the balanced case where $n_u \asymp n_v$, the condition $\frac{(a-b)^2}{a} \gtrsim 1$ above is known to be information-theoretically optimal up to a constant factor. In the unbalanced case with $n_v \geq n_u$, prior work has identified a sharp threshold for *detection*—the problem of recovering a fraction $1/2 + \epsilon$ of the community memberships for an arbitrarily small fixed constant $\epsilon > 0$. Specifically, such results reveal a fundamental lower limit that requires $\frac{(q_{\mathsf{in}} - q_{\mathsf{out}})^2}{q_{\mathsf{in}}} \gtrsim \frac{1}{\sqrt{n_u n_v}}$ [46, 47], thus implying the information-theoretic optimality of the spectral method (up to a logarithmic factor).

5. Further related work. A natural class of spectral algorithms to estimate the leading singular subspace of a matrix—when given a noisy and sub-sampled copy of the true matrix—is to compute the leading left singular subspace of the observed data matrix. Despite the simplicity of this idea, this type of spectral methods provably achieves appealing performances for multiple statistical problems when the true matrix is (nearly) square. A partial list of examples includes low-rank matrix estimation and completion [25, 34, 58, 65, 80], community detection [4, 74, 95, 111] and synchronization and alignment [4, 26, 100, 101]. However, the above-mentioned approach might lead to suboptimal performance when the row dimension and the column dimension of the matrix differ dramatically. This issue has already been recognized in multiple contexts, including but not limited to unfolding-based spectral methods for tensor estimation [57, 84, 106, 108, 114] and spectral methods for biclustering [47]. Motivated by this suboptimality issue, an alternative is to look at the "sample Gram matrix" which, as one expects, shares the same leading left singular space as the original observed data matrix. However, in the highly noisy or highly subsampled regime, the diagonal entries of the sample Gram matrix are highly biased, thus requiring special care. Several different treatments of diagonal components have been adopted for different contexts, including proper rescaling [52, 79, 84], deletion [47] and iterative updates [113]. The deletion strategy is perhaps the simplest of this kind, as it does not require estimation of noise parameters. We note, however, that performing more careful iterative updates might be beneficial for certain heteroskedastic noise scenarios; see [113] for details.

An important application of our work is the problem of tensor completion and estimation [48, 51, 56, 68, 76, 87, 94, 96, 98, 109, 110]. Despite its similarity to matrix completion, tensor completion is considerably more challenging; for concreteness and simplicity, we shall only discuss order-3 symmetric rank-r tensors in $\mathbb{R}^{d^3}$. Motivated by the success of matrix completion, a simple strategy is to unfold the observed tensor into a $d \times d^2$ matrix and to apply standard matrix completion methods for completion. However, existing statistical guarantees derived in the matrix completion literature [21, 53, 65] do not lead to useful bounds unless the sample size exceeds the order of rd^2, which far exceeds the requirement for other methods such as the sum-of-squares (SOS) hierarchy [11, 93]. The work by [84] demonstrates that spectral algorithms can also lead to useful estimates under minimal sample size, as long as we look at the "Gram matrix" instead. In addition, such spectral algorithms also play an important role in initializing other nonconvex optimization methods [14, 15, 106, 107].

In addition, there is an enormous literature on covariance estimation and PCA [12, 16, 18, 19, 62, 63, 81, 88, 91]. More recently, a computationally efficient algorithm called *HeteroPCA* has been proposed by [113] to achieve rate-optimal statistical guarantees for PCA in the presence of heteroskedastic noise. When it comes to incomplete data, a variety of methods have been introduced [43, 64, 67]. For instance, Lounici considered estimating the top eigenvector in the setting of sparse PCA in [78], and further proposed an estimator for the covariance matrix in [79]. In [20], bandable and sparse covariance matrices are considered. In addition, most of the prior work considered uniform random subsampling, and the recent work [92, 117] began to account for heterogeneous missingness patterns.

Turning to the problem of community recovery, we note that extensive research has been carried out on stochastic or censored block models, which can be viewed as special cases of unipartite networks [2, 17, 31–33, 37, 50, 54, 55, 60, 61, 83, 85, 86]. The algorithms that enable exact community recovery in these models include two-stage approaches [2, 85] and semidefinite programming [6, 8, 10, 54, 55]. In addition, spectral clustering algorithms have been extensively studied as well [4, 39, 40, 50, 73, 75, 89, 95, 103, 104, 111, 112]. While this class of algorithms was originally developed to yield almost exact recovery (e.g., [2]), the recent work by [4, 74] uncovered that spectral methods alone are sufficient to achieve optimal exact community recovery (a.k.a. achieving strong consistency) for stochastic block models. The interested reader is referred to [1] for an in-depth overview. Our work contributes to this growing literature by justifying the optimality of spectral methods in bipartite stochastic block models [46, 47, 49].

Further, entrywise statistical analysis has recently received significant attention for various statistical problems [4, 5, 23, 24, 27, 30, 42, 45, 74, 82, 90, 97, 108, 115]. For instance, entrywise guarantees for spectral methods are obtained in [27, 42] based on an algebraic Neumann trick, while the results in [4, 30, 115] were established based on a leave-one-out analysis. The work by [27, 69, 70] went one step further by controlling an arbitrary linear form of the eigenvectors or singular vectors of interest. These results, however, typically lead to suboptimal performance guarantees when the row dimension and the column dimension of the matrix are substantially different.

Finally, we recently became aware of [3], which also considers statistical guarantees of PCA beyond the usual ℓ_2 analysis; in particular, they develop an analysis framework that delivers tight ℓ_p perturbation bounds. Note, however, that their results are very different from the ones presented here.

6. Discussion. In this paper, we have investigated the effects of unbalancedness (as reflected by a large aspect ratio d_2/d_1) upon column subspace estimation, and developed tight $\ell_{2,\infty}$ statistical guarantees. Moving forward, there are many directions that are worth pursuing. For example, our current theory is likely suboptimal with respect to the dependence on the rank r and the condition number κ. For instance, the conditions (3.5) and the risk bound (3.7) involve high-order polynomials of κ in multiple places, and the rank r in our current theory cannot exceed the order of d_1/κ^4; all of these might be improvable via more refined analysis. In addition, it is natural to wonder whether we can extend our algorithm and theory to accommodate more general sampling patterns. Going beyond estimation, an important direction lies in statistical inference and uncertainty quantification for subspace estimation, namely, how to construct valid and hopefully optimal confidence regions that are likely to contain the unknown column subspace? It would also be interesting to investigate how to incorporate other structural prior (e.g., sparsity) to further reduce the sample complexity and/or improve the estimation accuracy. Finally, another interesting avenue for future exploration is the extension to distributed or decentralized settings.

Acknowledgments. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661 and IIS-1900140, and the Princeton SEAS Innovation Award. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303 and NSF CCF-1806154 and ECCS-1818571. H. V. Poor is supported in part by the grants NSF DMS-1736417, NSF CCF-1908308 and a Princeton Schmidt Data-X Research Award. C. Cai is supported in part by Gordon Y. S. Wu Fellowships in Engineering.

SUPPLEMENTARY MATERIAL

Numerical experiments and proofs (DOI: 10.1214/20-AOS1986SUPP; .pdf). Numerical experiments and proofs of the results in the paper can be found in the Supplementary Material.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** 6446–6531. MR3827065

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

[3] Abbe, E., Fan, J. and Wang, K. (2020). An ℓ_p theory of PCA and spectral clustering. Preprint. Available at arXiv:2006.14062.

[4] Abbe, E., Fan, J., Wang, K. and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. Preprint. Available at arXiv:1709.09565.

[5] Agarwal, A., Shah, D., Shen, D. and Song, D. (2019). On robustness of principal component regression. In *Advances in Neural Information Processing Systems* 9893–9903.

[6] Agarwal, N., Bandeira, A. S., Koiliaris, K. and Kolla, A. (2017). Multisection in the stochastic block model using semidefinite programming. In *Compressed Sensing and Its Applications*. *Appl*. *Numer*. *Harmon*. *Anal*. 125–162. Birkhäuser/Springer, Cham. MR3751736

[7] Alzahrani, T. and Horadam, K. J. (2016). Community detection in bipartite networks: Algorithms and case studies. In *Complex Systems and Networks*. *Underst*. *Complex Syst*. 25–50. Springer, Heidelberg. MR3586347

[8] Amini, A. A. and Levina, E. (2018). On semidefinite relaxations for the block model. *Ann*. *Statist*. **46** 149–179. MR3766949 https://doi.org/10.1214/17-AOS1545

[9] Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *J*. *Multivariate Anal*. **106** 167–177. MR2887686 https://doi.org/10.1016/j.jmva.2011.10.009

[10] Bandeira, A. S. (2018). Random Laplacian matrices and convex relaxations. *Found*. *Comput*. *Math*. **18** 345–379. MR3777782 https://doi.org/10.1007/s10208-016-9341-9

[11] Barak, B. and Moitra, A. (2016). Noisy tensor completion via the sum-of-squares hierarchy. In *Proceedings of the Conference on Learning Theory* 417–445.

[12] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann*. *Statist*. **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

[13] Cai, C., Li, G., Chi, Y., Poor, H. V. and Chen, Y. (2021). Supplement to "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees." https://doi.org/10.1214/20-AOS1986SUPP

[14] Cai, C., Li, G., Poor, H. V. and Chen, Y. (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems* 1861–1872.

[15] Cai, C., Poor, H. V. and Chen, Y. (2020). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *Proceedings of the International Conference on Machine Learning*. To appear. Available at arXiv:2006.08580.

[16] Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab*. *Theory Related Fields* **161** 781–815. MR3334281 https://doi.org/10.1007/s00440-014-0562-z

[17] Cai, T. T. and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann*. *Statist*. **43** 1027–1059. MR3346696 https://doi.org/10.1214/14-AOS1290

[18] Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann*. *Statist*. **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

[19] Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann*. *Statist*. **40** 2014–2042. MR3059075 https://doi.org/10.1214/12-AOS999

[20] Cai, T. T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *J*. *Multivariate Anal*. **150** 55–74. MR3534902 https://doi.org/10.1016/j.jmva.2016.05.002

[21] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found*. *Comput*. *Math*. **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

[22] Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans*. *Inf*. *Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

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

[24] Cape, J., Tang, M. and Priebe, C. E. (2019). Signal-plus-noise matrix models: Eigenvector deviations and fluctuations. *Biometrika* **106** 243–250. MR3912394 https://doi.org/10.1093/biomet/asy070

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

[26] Chen, Y. and Candès, E. J. (2018). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm*. *Pure Appl*. *Math*. **71** 1648–1714. MR3847751 https://doi.org/10.1002/cpa.21760

[27] Chen, Y., Cheng, C. and Fan, J. (2020). Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *Ann. Statist.* To appear. Available at arXiv:1811.12804.

[28] CHEN, Y. and CHI, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Process. Mag.* **35** 14–31.

[29] CHEN, Y., CHI, Y., FAN, J., MA, C. and YAN, Y. (2019). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. Preprint. Available at arXiv:1902.07698.

[30] CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* **47** 2204–2235. MR3953449 https://doi.org/10.1214/18-AOS1745

[31] CHEN, Y., KAMATH, G., SUH, C. and TSE, D. (2016). Community recovery in graphs with locality. In *Proceedings of the International Conference on Machine Learning* 689–698.

[32] CHEN, Y., LI, X. and XU, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.* **46** 1573–1602. MR3819110 https://doi.org/10.1214/17-AOS1595

[33] CHEN, Y., SUH, C. and GOLDSMITH, A. J. (2016). Information recovery from pairwise measurements. *IEEE Trans. Inf. Theory* **62** 5881–5905. MR3552429 https://doi.org/10.1109/TIT.2016.2600566

[34] CHEN, Y. and WAINWRIGHT, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Available at arXiv:1509.03025.

[35] CHENG, C., WEI, Y. and CHEN, Y. (2020). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. Preprint. Available at arXiv:2001.04620.

[36] CHI, Y., LU, Y. M. and CHEN, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.* **67** 5239–5269. MR4016283 https://doi.org/10.1109/TSP.2019.2937282

[37] CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of the Conference on Learning Theory* 391–423.

[38] CHO, J., KIM, D. and ROHE, K. (2017). Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statist. Sinica* **27** 1921–1948. MR3726772

[39] COJA-OGHLAN, A. (2006). A spectral heuristic for bisecting random graphs. *Random Structures Algorithms* **29** 351–398. MR2254496 https://doi.org/10.1002/rsa.20116

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

[41] DHILLON, I. S. (2001). Co-clustering documents and using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 269–274. ACM, New York.

[42] ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: Spectral analysis beyond Davis–Kahan. In *Proceedigs of the* 29*th Conference on Algorithmic Learning Theory* 321–358.

[43] ELSENER, A. and VAN DE GEER, S. (2019). Sparse spectral estimation with missing and corrupted measurements. *Stat* **8** e229, 11. MR3978409 https://doi.org/10.1002/sta4.229

[44] FAN, J., WANG, K., ZHONG, Y. and ZHU, Z. (2018). Robust high dimensional factor models with applications to statistical machine learning. Preprint. Available at arXiv:1808.03889.

[45] FAN, J., WANG, W. and ZHONG, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** 1-42. MR3827095

[46] FELDMAN, V., PERKINS, W. and VEMPALA, S. (2015). Subsampled power iteration: A unified algorithm for block models and planted csp's. In *Advances in Neural Information Processing Systems* 2836–2844.

[47] FLORESCU, L. and PERKINS, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Proceedings of the Conference on Learning Theory* 943–959.

[48] GANDY, S., RECHT, B. and YAMADA, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Probl.* **27** 025010, 19. MR2765628 https://doi.org/10.1088/0266-5611/27/2/025010

[49] GAO, C., LU, Y., MA, Z. and ZHOU, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* **17** 5602–5630. MR3569248

[50] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** 1980–2024. MR3687603

[51] GHASSEMI, M., SHAKERI, Z., SARWATE, A. D. and BAJWA, W. U. (2017). STARK: Structured dictionary learning through rank-one tensor recovery. In *Proceedings of the* 2017 *IEEE* 7*th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* 1–5. IEEE, New York.

[52] GONEN, A., ROSENBAUM, D., ELDAR, Y. and SHALEV-SHWARTZ, S. (2016). Subspace learning with partial information. *J. Mach. Learn. Res.* **17** 1821–1841. MR3504612

[53] GROSS, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory* **57** 1548–1566. MR2815834 https://doi.org/10.1109/TIT.2011.2104999

[54] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025 https://doi.org/10.1007/s00440-015-0659-z

[55] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inf. Theory* **62** 5918–5937. MR3552431 https://doi.org/10.1109/TIT.2016.2594812

[56] HAO, B., ZHANG, A. and CHENG, G. (2018). Sparse and low-rank tensor estimation via cubic sketchings. Preprint. Available at arXiv:1801.09326.

[57] HOPKINS, S. B., SHI, J., SCHRAMM, T. and STEURER, D. (2016). Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors. In *Proceedings of the* 48*th ACM Symposium on Theory of Computing* 178–191. ACM, New York. MR3536564 https://doi.org/10.1145/2897518.2897529

[58] JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *Proceedings of the* 45*th ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828 https://doi.org/10.1145/2488608.2488693

[59] JAIN, P. and OH, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems* 1431–1439.

[60] JALALI, A., CHEN, Y., SANGHAVI, S. and XU, H. (2011). Clustering partially observed graphs via convex optimization. In *Proceedings of the International Conference on Machine Learning* **11** 1001–1008.

[61] JAVANMARD, A., MONTANARI, A. and RICCI-TERSENGHI, F. (2016). Phase transitions in semidefinite relaxations. *Proc. Natl. Acad. Sci. USA* **113** E2218–E2223. MR3494080 https://doi.org/10.1073/pnas.1523097113

[62] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[63] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[64] JOSSE, J. and HUSSON, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *J. SFdS* **153** 79–99. MR3008600

[65] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inf. Theory* **56** 2980–2998. MR2683452 https://doi.org/10.1109/TIT.2010.2046205

[66] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

[67] KIERS, H. A. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika* **62** 251–266.

[68] KIM, H.-J., OLLILA, E., KOIVUNEN, V. and CROUX, C. (2013). Robust and sparse estimation of tensor decompositions. In *Proceedings of the* 2013 *IEEE Global Conference on Signal and Information Processing* 965–968. IEEE.

[69] KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** 1976–2013. MR3573302 https://doi.org/10.1214/15-AIHP705

[70] KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII. Progress in Probability* **71** 397–423. Springer, Cham. MR3565274 https://doi.org/10.1007/978-3-319-40519-3_18

[71] LARREMORE, D. B., CLAUSET, A. and JACOBS, A. Z. (2014). Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90** 012805.

[72] LAWLEY, D. N. and MAXWELL, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D* (*The Statistician*) **12** 209–229.

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

[74] LEI, L. (2019). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. Preprint. Available at arXiv:1909.04798.

[75] LELARGE, M., MASSOULIÉ, L. and XU, J. (2015). Reconstruction in the labelled stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **2** 152–163. MR3453283 https://doi.org/10.1109/TNSE.2015.2490580

[76] LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 208–220.

[77] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038 https://doi.org/10.1214/12-AOS1018

[78] LOUNICI, K. (2013). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI*. *Progress in Probability* **66** 327–356. Birkhäuser/Springer, Basel. MR3443508

[79] LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437 https://doi.org/10.3150/12-BEJ487

[80] MA, C., WANG, K., CHI, Y. and CHEN, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.* **20** 451–632. MR4099988 https://doi.org/10.1007/s10208-019-09429-9

[81] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121 https://doi.org/10.1214/13-AOS1097

[82] MAO, X., SARKAR, P. and CHAKRABARTI, D. (2017). Estimating mixed memberships with sharp eigenvector deviations. Preprint. Available at arXiv:1709.00407.

[83] MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *Proceedings of the* 46*th ACM Symposium on Theory of Computing* 694–703. ACM, New York. MR3238997

[84] MONTANARI, A. and SUN, N. (2018). Spectral algorithms for tensor completion. *Comm. Pure Appl. Math.* **71** 2381–2425. MR3862094 https://doi.org/10.1002/cpa.21748

[85] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. Preprint. Available at arXiv:1407.1591.

[86] MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461. MR3383334 https://doi.org/10.1007/s00440-014-0576-6

[87] MU, C., HUANG, B., WRIGHT, J. and GOLDFARB, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the International Conference on Machine Learning* 73–81.

[88] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618

[89] O'ROURKE, S., VU, V. and WANG, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra Appl.* **540** 26–59. MR3739989 https://doi.org/10.1016/j.laa.2017.11.014

[90] PANANJADY, A. and WAINWRIGHT, M. J. (2019). Value function estimation in Markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. Preprint. Available at arXiv:1909.08749.

[91] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[92] PAVEZ, E. and ORTEGA, A. (2019). Covariance matrix estimation with non uniform and data dependent missing observations. Preprint. Available at arXiv:1910.00667.

[93] POTECHIN, A. and STEURER, D. (2017). Exact tensor completion with sum-of-squares. In *Proceedings of the Conference on Learning Theory* 1619–1673.

[94] RICHARD, E. and MONTANARI, A. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.

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

[96] ROMERA-PAREDES, B. and PONTIL, M. (2013). A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems* 2967–2975.

[97] RUDELSON, M. and VERSHYNIN, R. (2015). Delocalization of eigenvectors of random matrices with independent entries. *Duke Math. J.* **164** 2507–2538. MR3405592 https://doi.org/10.1215/00127094-3129809

[98] SALMI, J., RICHTER, A. and KOIVUNEN, V. (2009). Sequential unfolding SVD for tensors with applications in array signal processing. *IEEE Trans. Signal Process.* **57** 4719–4733. MR2751759 https://doi.org/10.1109/TSP.2009.2027740

[99] SEMERCI, O., HAO, N., KILMER, M. E. and MILLER, E. L. (2014). Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Process.* **23** 1678–1693. MR3191324 https://doi.org/10.1109/TIP.2014.2305840

[100] SHEN, Y., HUANG, Q., SREBRO, N. and SANGHAVI, S. (2016). Normalized spectral map synchronization. In *Advances in Neural Information Processing Systems* 4925–4933.

[101] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

[102] SUN, R. and LUO, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory* **62** 6535–6579. MR3565131 https://doi.org/10.1109/TIT.2016.2598574

[103] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899 https://doi.org/10.1080/01621459.2012.699795

[104] VU, V. (2018). A simple SVD algorithm for finding hidden partitions. *Combin. Probab. Comput.* **27** 124–140. MR3734334 https://doi.org/10.1017/S0963548317000463

[105] WEDIN, P. (1973). Perturbation theory for pseudo-inverses. *Nordisk Tidskr. Informationsbehandling* (*BIT*) **13** 217–232. MR0336982 https://doi.org/10.1007/bf01933494

[106] XIA, D. and YUAN, M. (2019). On polynomial time methods for exact low-rank tensor completion. *Found. Comput. Math.* **19** 1265–1313. MR4029842 https://doi.org/10.1007/s10208-018-09408-6

[107] XIA, D., YUAN, M. and ZHANG, C.-H. (2017). Statistically optimal and computationally efficient low rank tensor completion from noisy entries. Preprint. Available at arXiv:1711.04934.

[108] XIA, D. and ZHOU, F. (2019). The sup-norm perturbation of HOSVD and low rank tensor denoising. *J. Mach. Learn. Res.* **20** 61–1. MR3960915

[109] YUAN, M. and ZHANG, C.-H. (2016). On tensor completion via nuclear norm minimization. *Found. Comput. Math.* **16** 1031–1068. MR3529132 https://doi.org/10.1007/s10208-015-9269-5

[110] YUAN, M. and ZHANG, C.-H. (2017). Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Trans. Inf. Theory* **63** 6753–6766. MR3707566 https://doi.org/10.1109/TIT.2017.2724549

[111] YUN, S.-Y. and PROUTIERE, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. Preprint. Available at arXiv:1412.7335.

[112] YUN, S.-Y. and PROUTIERE, A. (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems* 965–973.

[113] ZHANG, A., CAI, T. T. and WU, Y. (2018). Heteroskedastic PCA: Algorithm, optimality, and applications. Preprint. Available at arXiv:1810.08316v2.

[114] ZHANG, A. and XIA, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Trans. Inf. Theory* **64** 7311–7338. MR3876445 https://doi.org/10.1109/TIT.2018.2841377

[115] ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28** 989–1016. MR3782406 https://doi.org/10.1137/17M1122025

[116] ZHOU, Z. and AMINI, A. A. (2018). Optimal bipartite network clustering. Preprint. Available at arXiv:1803.06031.

[117] ZHU, Z., WANG, T. and SAMWORTH, R. J. (2019). High-dimensional principal component analysis with heterogeneous missingness. Preprint. Available at arXiv:1906.12125.