The Annals of Statistics
2025, Vol. 53, No. 1, 91–116
https://doi.org/10.1214/24-AOS2456

DEFLATED HETEROPCA: OVERCOMING THE CURSE OF ILL-CONDITIONING IN HETEROSKEDASTIC PCA

BY YUCHEN ZHOU[1,a] AND YUXIN CHEN[2,b]

[1]*Department of Statistics, University of Illinois Urbana-Champaign,* [a]*yuchenz@illinois.edu*
[2]*Department of Statistics and Data Science, The Wharton School, University of Pennsylvania,* [b]*yuxinc@wharton.upenn.edu*

This paper is concerned with estimating the column subspace of a low-rank matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n_1 \times n_2}$ from contaminated data. How to obtain optimal statistical accuracy while accommodating the widest range of signal-to-noise ratios (SNRs) becomes particularly challenging in the presence of heteroskedastic noise and unbalanced dimensionality (i.e., $n_2 \gg n_1$). While the state-of-the-art algorithm HeteroPCA emerges as a powerful solution for solving this problem, it suffers from "the curse of ill-conditioning," namely, its performance degrades as the condition number of $\boldsymbol{X}^\star$ grows. In order to overcome this critical issue without compromising the range of allowable SNRs, we propose a novel algorithm, called Deflated − HeteroPCA, that achieves near-optimal and condition-number-free theoretical guarantees in terms of both ℓ_2 and $\ell_{2,\infty}$ statistical accuracy. The proposed algorithm divides the spectrum of $\boldsymbol{X}^\star$ into well-conditioned and mutually well-separated subblocks, and applies HeteroPCA to conquer each subblock successively. Further, an application of our algorithm and theory to two canonical examples—the factor model and tensor PCA—leads to remarkable improvement for each application.

1. Introduction. In a diverse array of science and engineering applications, we are asked to identify a low-dimensional subspace that best captures the information underlying a large collection of high-dimensional data points, a classical problem that goes by the names of principal component analysis (PCA), subspace estimation, subspace tracking, among others (Johnstone and Paul (2018), Balzano, Chi and Lu (2018), Chen et al. (2021)). A simple yet useful mathematical model is of the following form: imagine we have an unknown large-dimensional matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n_1 \times n_2}$ whose columns are high-dimensional vectors embedded in a r-dimensional subspace (so that $\boldsymbol{X}^\star$ has rank $r \ll \min\{n_1, n_2\}$), and we seek to estimate the *column space* of $\boldsymbol{X}^\star$ from noisy observations:

$$\boldsymbol{Y} = \boldsymbol{X}^\star + \boldsymbol{E} \in \mathbb{R}^{n_1 \times n_2}, \tag{1}$$

where $\boldsymbol{E}$ stands for the noise matrix that contaminates the data. Despite decades-long research, there remain substantial challenges to handle heteroskedastic noise in high dimension, as we shall elaborate on below.

1.1. *Challenges*: *Unbalanced dimensionality and heteroskedasticity.* How to achieve statistically efficient PCA in high dimension is an active research topic that has received much recent interest (Lounici (2014), Johnstone and Paul (2018), Cai et al. (2021), Zhu, Wang and Samworth (2022), Zhang, Cai and Wu (2022), Agterberg, Lubberts and Priebe (2022)). In this paper, we pay particular attention to the case where n_1 and n_2 are both enormous but

Received March 2023; revised April 2024.
MSC2020 subject classifications. Primary 62F10; secondary 62H25.
Key words and phrases. Principal component analysis (PCA), heteroskedastic noise, the curse of ill-conditioning, factor models, tensor PCA.

highly *unbalanced* in the sense that $n_1 \ll n_2$, a scenario that arises frequently in, say, covariance estimation (when there are many noisy samples available) and tensor estimation (when one has to matrice the tensor before estimation). Such unbalanced dimensionality gives rise to unique challenges not present in the complement case: as the signal-to-noise ratio (SNR) keeps decreasing, one might soon enter a regime where consistent estimation of $\boldsymbol{X}^\star$ is no longer infeasible but its column subspace—which is much smaller dimensional than the full matrix—remains estimatable. This regime is often considerably more challenging than the case with $n_2 = O(n_1)$, given that the majority of low-rank matrix estimation algorithms that directly attempt to estimate $\boldsymbol{X}^\star$ become completely off.

One natural strategy that comes into mind is thus to estimate the column subspace of $\boldsymbol{X}^\star$ by calculating the left singular subspace of the observed matrix $\boldsymbol{Y}$ (Cai and Zhang (2018), Abbe et al. (2020), Chen et al. (2021)), which we shall refer to as the *vanilla SVD-based approach* throughout. In the case with $n_1 \ll n_2$, this simple scheme has only been shown to achieve the desired statistical performance when the noise matrix $\boldsymbol{E}$ is composed of i.i.d. entries, but falls short of effectiveness when handling *heteroskedastic* noise (i.e., the scenario where the variances of the entries of $\boldsymbol{E}$ are location-varying) (Zhang, Cai and Wu (2022), Cai et al. (2021)). This issue presents a hurdle to transferring this scheme from theory to practice, due to the ubiquity of heteroskedastic data in applications like social networks, recommendation systems, medical imaging, etc.

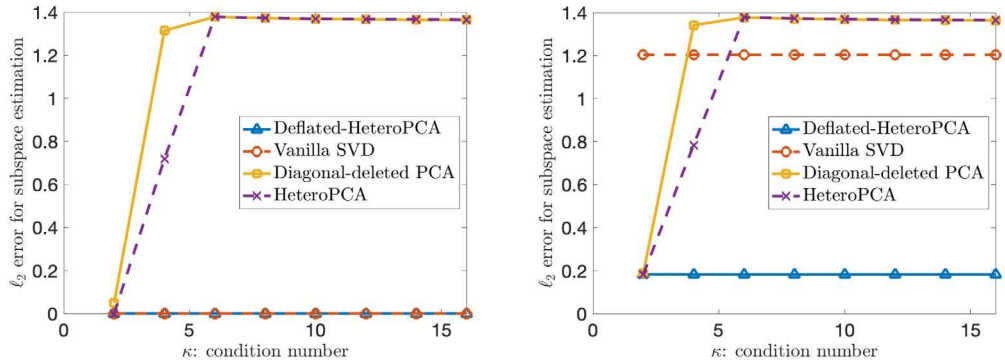
To mitigate this issue, at least two strategies have been proposed that attempt estimation by looking at the empirical covariance matrix (or gram matrix) $\boldsymbol{Y}\boldsymbol{Y}^\top$. Recognizing that large heteroskedastic noise might lead to significant bias in the diagonal of $\boldsymbol{Y}\boldsymbol{Y}^\top$ that distorts estimation, one natural remedy is to zero out (or sometimes rescale) the diagonal entries of $\boldsymbol{Y}\boldsymbol{Y}^\top$ before computing its eigendecomposition (Koltchinskii and Giné (2000), Lounici (2014), Florescu and Perkins (2016), Loh and Wainwright (2012), Montanari and Sun (2018), Elsener and van de Geer (2019), Cai et al. (2021), Ndaoud, Sigalla and Tsybakov (2022)). A more refined iterative procedure called HeteroPCA was subsequently proposed by Zhang, Cai and Wu (2022), which starts with the solution of diagonal-deleted PCA (cf. (10)) and alternates between:

- imputing the diagonal entries of $\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$;
- computing the rank-r eigenspace of $\boldsymbol{Y}\boldsymbol{Y}^\top$ with its diagonal replaced by the imputed values.

See Section 3 for precise descriptions. In both theory and numerical experiments, this iterative paradigm yields enhanced performance compared to diagonal-deleted PCA (Zhang, Cai and Wu (2022), Yan, Chen and Fan (2024)).

1.2. *The curse of ill-conditioning.* Nevertheless, one drawback stands out when running either diagonal-deleted PCA or HeteroPCA in practice; that is, both algorithms become ineffective as the condition number of $\boldsymbol{X}^\star$ (when restricted to its nonzero singular values) grows. Let us illustrate this point more clearly via numerical experiments.

- (*Numerical example*) Consider the case where the unknown signal $\boldsymbol{X}^\star$ has rank $r = 2$ and obeys $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$, where the columns of $\boldsymbol{U}^\star \in \mathbb{R}^{n_1\times 2}$ (resp., $\boldsymbol{V}^\star \in \mathbb{R}^{n_2\times 2}$) are the two left (resp., right) singular vectors of $\boldsymbol{X}^\star$, and $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{2\times 2}$ is a diagonal matrix composed of the two singular values $\sigma_1^\star \geq \sigma_2^\star > 0$ of $\boldsymbol{X}^\star$. Denote by $\kappa = \sigma_1^\star/\sigma_2^\star$ the condition number of $\boldsymbol{\Sigma}^\star$. We conduct a series of experiments based on randomly generated $\boldsymbol{X}^\star$ with $n_2 \gg n_1$, as detailed in the caption of Figure 1. As illustrated in Figure 1, when κ is not too large, both diagonal-deleted PCA and HeteroPCA fail to return reliable estimates of the subspace $\boldsymbol{U}^\star$, even in the noiseless case (i.e., $\boldsymbol{E} = 0$).



(a) Noiseless case (b) Noisy case

FIG. 1. *Subspace estimation error versus condition number κ of $\boldsymbol{\Sigma}^\star$. Here, we set $r = 2$, $n_1 = 200$ and $n_2 = 40{,}000$. The truth $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ has rank 2 with $\boldsymbol{U}^\star \in \mathcal{R}^{n_1\times 2}$ and $\boldsymbol{V}^\star \in \mathcal{R}^{n_2\times 2}$ generated randomly. Plot (a) represents the noiseless case ($\boldsymbol{E} = \boldsymbol{0}$). In plot (b), we choose the two singular values of $\boldsymbol{X}^\star$ as $\sigma_1^\star = \kappa\sigma_2^\star$ and $\sigma_2^\star = 200$, generate $\{\omega_i\}_{1\le i\le n_1}$ independently from* Unif$([0,2])$ *and draw the entries of $\boldsymbol{E} = [E_{i,j}]_{1\le i\le n_1, 1\le j\le n_2}$ independently such that $E_{i,j} \sim \mathcal{N}(0,\omega_i^2)$. We compare multiple subspace estimators here, where* HeteroPCA *is run with* 100 *iterations. For each estimator $\widehat{\boldsymbol{U}}$, we compute the spectral-norm-based error $\|\widehat{\boldsymbol{U}}\boldsymbol{R}_{\widehat{\boldsymbol{U}}} - \boldsymbol{U}^\star\|$ as κ varies, where $\boldsymbol{R}_{\widehat{\boldsymbol{U}}} = \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r,r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}}$; the results are averaged over* 50 *independent runs.*

In summary, both diagonal-deleted PCA and HeteroPCA suffer from the "curse of ill-conditioning," namely they might lead to grossly incorrect subspace estimates as the largest signal component strengthens with all other signal components unchanged. This observation is somewhat counterintuitive; after all, altering the signal this way only serves to increase the SNR, and hence, simplify the task from the information-theoretic perspective. In this sense, the aforementioned curse of ill-conditioning seems to be algorithm specific, although the two algorithms it concerns happen to be the state-of-the-art methods. All this naturally leads to the following question:

Can we overcome the above curse of ill-conditioning without compromising the advantages of both diagonal-deleted PCA and HeteroPCA?

1.3. *This paper.* As it turns out, we can answer the above question in the affirmative, which forms the main contribution of this paper. Our main findings are summarized as follows:

- *Algorithm design.* In an attempt to address the above question, we propose a new algorithm—dubbed as Deflated-HeteroPCA—on the basis of HeteroPCA. In a nutshell, the proposed algorithm divides the spectrum of $\boldsymbol{X}^\star$ into well-conditioned yet mutually well-separated subblocks, and successively applies HeteroPCA to conquer each subblock. This approach counters the adverse influence of ill-conditioning via successive "deflation" (a term borrowed from Dobriban and Owen (2019)), which gradually "deflates" the undesirable bias effect resulting from the diagonal deletion operation.
- *Statistical guarantees.* We develop sharp theoretical guarantees, in terms of both ℓ_2 (spectral-norm-based) and $\ell_{2,\infty}$ estimation errors, for the proposed algorithm. Encouragingly, all of these statistical guarantees are condition-number-free, and match the minimax lower bounds established in Zhang, Cai and Wu (2022) and Cai et al. (2021) (up to some logarithmic factors). To the best of our knowledge, these provide the first near-optimal results in the heteroskedastic PCA setting herein that (i) do not degrade as the condition number of the truth increases, and (ii) accommodate the widest range of SNRs.

- *Consequences in two canonical examples.* To illustrate the utility of our algorithm and theory, we develop concrete consequences of our results for two canonical examples: (a) the factor model, and (b) tensor PCA. We demonstrate that (i) Deflated-HeteroPCA achieves rate-optimal and condition-number-free estimation under the factor model, and (ii) Deflated-HeteroPCA followed by the HOOI algorithm improves upon the state-of-the-art performance guarantees for tensor PCA. Numerical experiments are carried out to corroborate the effectiveness of the proposed algorithm.

Paper organization. The rest of the paper is organized as follows. We formulate the problem precisely in Section 2, and present the proposed algorithm in Section 3. The theoretical guarantees of our algorithm, along with their implications, are presented in Section 4. We develop concrete consequences of our results in two applications in Section 5. Additional numerical experiments are reported in Section 6, and a discussion of further related works is provided in Section 7. The technical proofs are collected in the Supplementary Material (Zhou and Chen (2025)).

1.4. *Notation.* Throughout this paper, we denote $[n] := \{1, \dots, n\}$ for any positive integer n. We let bold capital letters (e.g., $\boldsymbol{X}$) and bold lowercase letters (e.g., $\boldsymbol{x}$) denote matrices and vectors, respectively. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$, $\lambda_i(\boldsymbol{A})$ and $\sigma_i(\boldsymbol{A})$ are used to represent the ith largest eigenvalue (in magnitude) and the ith largest singular value of $\boldsymbol{A}$, respectively. Let $\|\cdot\|_{\mathrm{F}}$ indicate the Frobenius norm and $\|\cdot\|$ the spectral norm. We denote by $\boldsymbol{A}_{i,:}$ and $\boldsymbol{A}_{:,j}$ the ith column and the jth row of $\boldsymbol{A}$, respectively. We also let $\boldsymbol{A}_{:,i:j}$ denote the submatrix of $\boldsymbol{A}$ containing those columns with indices falling in $[i, j]$. Let $\|\boldsymbol{A}\|_{2,\infty} := \max_i \|\boldsymbol{A}_{i,:}\|_2$ denote the $\ell_{2,\infty}$ norm of $\boldsymbol{A}$. We use $\mathcal{O}^{n,r} := \{\boldsymbol{U} \in \mathbb{R}^{n \times r} : \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r\}$ to represent the set containing all $n \times r$ matrices with orthonormal columns. For any $\boldsymbol{U} \in \mathcal{O}^{n,r}$, we define the projection matrix $\mathcal{P}_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$. Let $\boldsymbol{U}_\perp \in \mathcal{O}^{n,n-r}$ denote the orthogonal complement of $\boldsymbol{U}$. We use $\mathcal{P}_{\mathsf{diag}}(\cdot)$ to represent the projection operator that keeps all diagonal entries and sets to zero all nondiagonal entries; meanwhile, we define $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{M}) := \boldsymbol{M} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{M})$ for any $\boldsymbol{M} \in \mathbb{R}^{n \times n}$. For any vector $\boldsymbol{a} = (a_1, \dots, a_n)$, we denote by $\mathsf{diag}(\boldsymbol{a}) \in \mathbb{R}^{n \times n}$ the diagonal matrix whose (i, i)th entry is a_i. For any full-rank matrix $\boldsymbol{H} \in \mathbb{R}^{r \times r}$ with singular value decomposition (SVD) $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we define the sign matrix

$$\mathsf{sgn}(\boldsymbol{H}) := \boldsymbol{U}\boldsymbol{V}^\top. \tag{2}$$

We let $C, c, C_0, c_0, \dots$ denote numerical constants whose values may change from line to line. The boldface calligraphic letters (e.g., $\boldsymbol{\mathcal{X}}$) are used to represent tensors. For any tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and any matrix $\boldsymbol{V}_1 \in \mathbb{R}^{n_1 \times r_1}$, we define the multilinear product $\times_1$ as follows:

$$\boldsymbol{\mathcal{G}} \times_1 \boldsymbol{V}_1 = \left(\sum_{j=1}^{r_1} G_{j,i_2,i_3} V_{i_1,j} \right)_{i_1 \in [n_1], i_2 \in [r_2], i_3 \in [r_3]}.$$

We can define $\times_2$ and $\times_3$ analogously. For any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, let $\mathcal{M}_j(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{n_j \times (n_1 n_2 n_3 / n_j)}$ denote the jth matricization of $\boldsymbol{\mathcal{X}}$ such that for any $(i_1, i_2, i_3) \in [n_1] \times [n_2] \times [n_3]$,

$$[\mathcal{M}_1(\boldsymbol{\mathcal{X}})]_{i_1, i_2 + n_2(i_3 - 1)} = [\mathcal{M}_2(\boldsymbol{\mathcal{X}})]_{i_2, i_3 + n_3(i_1 - 1)} = [\mathcal{M}_3(\boldsymbol{\mathcal{X}})]_{i_3, i_1 + n_1(i_2 - 1)} = X_{i_1, i_2, i_3}.$$

The Frobenius norm of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is defined as

$$\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} = \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} X_{i,j,k}^2 \right)^{1/2}.$$

The notation $f(n_1, n_2) \lesssim g(n_1, n_2)$ or $f(n_1, n_2) = O(g(n_1, n_2))$ means that $|f(n_1, n_2)| \le C g(n_1, n_2)$ holds for some numerical constant $C > 0$; we let $f(n_1, n_2) \gtrsim g(n_1, n_2)$ indicate

that $f(n_1, n_2) \geq C|g(n_1, n_2)|$ for some numerical constant $C > 0$; $f(n_1, n_2) \asymp g(n_1, n_2)$ means that both $f(n_1, n_2) \lesssim g(n_1, n_2)$ and $f(n_1, n_2) \gtrsim g(n_1, n_2)$ hold; we use the notation $f(n_1, n_2) \ll g(n_1, n_2)$ to represent that $f(n_1, n_2) \leq cg(n_1, n_2)$ holds for some sufficiently small constant $c > 0$, and we say $f(n_1, n_2) \gg g(n_1, n_2)$ if $g(n_1, n_2) \ll f(n_1, n_2)$. In addition, we use $f(n_1, n_2) = o(g(n_1, n_2))$ to indicate that $f(n_1, n_2)/g(n_1, n_2) \to 0$ as $\min\{n_1, n_2\} \to \infty$. For any $a, b \in \mathbb{R}$, we define $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$.

2. Problem formulation.

Models and assumptions. Let us present a more precise description of the problem to be studied here. Imagine that we have access to the following noisy data matrix:

$$\boldsymbol{Y} = \boldsymbol{X}^\star + \boldsymbol{E} \in \mathbb{R}^{n_1 \times n_2}, \tag{3}$$

where $\boldsymbol{E} = [E_{i,j}]_{1 \leq i \leq n_1, 1 \leq j \leq n_2}$ is a zero-mean noise matrix composed of independent entries, and $\boldsymbol{X}^\star = [X^\star_{i,j}]_{1 \leq i \leq n_1, 1 \leq j \leq n_2}$ is a rank-r matrix to be estimated. The SVD of the signal matrix $\boldsymbol{X}^\star$ is given by

$$\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} = \sum_{i=1}^{r} \sigma_i^\star \boldsymbol{u}_i^\star \boldsymbol{v}_i^{\star\top} \in \mathbb{R}^{n_1 \times n_2}. \tag{4}$$

Here, $\sigma_1^\star \geq \cdots \geq \sigma_r^\star > 0$ denote the singular values of $\boldsymbol{X}^\star$, $\boldsymbol{u}_i^\star$ (resp., $\boldsymbol{v}_i^\star$) represents the left (resp., right) singular vector associated with $\sigma_i^\star$, and we introduce the matrices $\boldsymbol{\Sigma}^\star = \mathsf{diag}(\sigma_1^\star, \ldots, \sigma_r^\star)$, $\boldsymbol{U}^\star = [\boldsymbol{u}_1^\star, \ldots, \boldsymbol{u}_r^\star] \in \mathcal{O}^{n_1, r}$ and $\boldsymbol{V}^\star = [\boldsymbol{v}_1^\star, \ldots, \boldsymbol{v}_r^\star] \in \mathcal{O}^{n_2, r}$. Clearly, $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ represent the column and row subspaces of $\boldsymbol{X}^\star$, respectively.

Moreover, we introduce additional definitions and assumptions to be used throughout:

- To begin with, let us introduce the following incoherence condition that appears frequently in the low-rank matrix estimation literature (Candès and Recht (2009), Keshavan, Montanari and Oh (2010), Chen et al. (2021)).

DEFINITION 1 (Incoherence). The incoherence parameters μ_1 and μ_2 of $\boldsymbol{X}^\star$ are defined as

$$\mu_1 := \frac{n_1}{r} \max_{1 \leq i \leq n_1} \left\| \boldsymbol{U}^\star_{i,:} \right\|_2^2 \quad \text{and} \quad \mu_2 := \frac{n_2}{r} \max_{1 \leq j \leq n_2} \left\| \boldsymbol{V}^\star_{j,:} \right\|_2^2. \tag{5}$$

It is self-evident that $1 \leq \mu_1 \leq n_1/r$ and $1 \leq \mu_2 \leq n_2/r$. In words, if the incoherence parameter μ_1 (resp., μ_2) is small, then the energy of of $\boldsymbol{U}^\star$ (resp., $\boldsymbol{V}^\star$) would be more or less dispersed across all rows of $\boldsymbol{U}^\star$ (resp., $\boldsymbol{V}^\star$). Throughout this paper, for simplicity we denote

$$\mu = \max\{\mu_1, \mu_2\} \quad \text{and} \quad n := \max\{n_1, n_2\}. \tag{6}$$

- Turning to the zero-mean noise matrix $\boldsymbol{E}$, we first introduce the following parameters:

$$\begin{aligned} \omega_{i,j}^2 &:= \mathsf{Var}[E_{i,j}], \qquad \omega_{\mathsf{max}}^2 := \max_{i,j} \mathsf{Var}[E_{i,j}], \\ \omega_{\mathsf{row}}^2 &:= \max_i \sum_{j=1}^{n_2} \mathsf{Var}[E_{i,j}], \qquad \omega_{\mathsf{col}}^2 := \max_j \sum_{i=1}^{n_1} \mathsf{Var}[E_{i,j}], \end{aligned} \tag{7}$$

where $\omega_{i,j}, \omega_{\mathsf{max}}, \omega_{\mathsf{row}}, \omega_{\mathsf{col}} \geq 0$. Here, we allow the variances $\{\omega_{i,j}^2\}$ to be location-varying, in order to account for *heteroskedasticity* of noise. Moreover, we impose the following assumptions throughout.

ASSUMPTION 1 (Noise). Suppose the noise components satisfy the following properties:

1. The $E_{i,j}$'s are statistically independent and obey $\mathbb{E}[E_{i,j}] = 0$ for all $(i, j) \in [n_1] \times [n_2]$;
2. $\mathbb{P}(|E_{i,j}| > B) \le n^{-12}$, where the quantity B satisfies

$$B \le C_{\mathsf{b}} \frac{\min\{(\omega_{\mathsf{row}}\omega_{\mathsf{col}})^{1/2}, \omega_{\mathsf{row}}\}}{\sqrt{\log n}}$$

for some numerical constant $C_{\mathsf{b}} > 0$.

REMARK 1. Assumption 1 imposes a mild condition on the tails of noise. For instance, if $\omega_{i,j} \asymp \omega_{\mathsf{max}}$ for all i, j, then B is allowed to be as large as $\min\{(n_1 n_2)^{1/4}, \sqrt{n_2}\}\omega_{\mathsf{max}}$ (up to some logarithmic factor), which can be substantially larger than the typical noise level ω_{max}. In comparisons to prior works, (i) this assumption is similar to—in fact slightly weaker than—Cai et al. (2021), Assumption 2 (in that the assumption therein requires noise distributions to be symmetric); (ii) given that Assumption 1 is satisfied if $\{E_{i,j}\}$ are $C\omega_{\mathsf{max}}$-sub-Gaussian and $\omega_{\mathsf{max}} \lesssim \min\{(\omega_{\mathsf{row}}\omega_{\mathsf{col}})^{1/2}, \omega_{\mathsf{row}}\}/\log n$, it is less stringent than the one assumed in Zhang, Cai and Wu (2022), Theorem 4.

Goal. We seek to estimate the column subspace $\boldsymbol{U}^\star$ (up to global rotation) on the basis of $\boldsymbol{Y}$. Our goal is to design an estimator that satisfies the following two desirable properties simultaneously:

(1) it allows for faithful estimation of the column subspace despite the presence of heteroskedasticity and unbalanced dimensionality; we hope to accomplish this for the widest possible range of SNRs;

(2) it achieves the desirable statistical guarantees that do not degrade when the condition number $\kappa = \sigma_1^\star/\sigma_r^\star$ increases.

3. Algorithms. In this section, we proceed to describe the proposed algorithm in attempt to achieve the goal set forth in Section 2, following a brief overview of previous algorithms.

Review: *SVD*, *diagonal-deleted PCA and* HeteroPCA. Before continuing, we briefly review three popular methods that are commonly studied in the literature.

- *The vanilla SVD-based approach.* This approach computes the leading r singular vectors of $\boldsymbol{Y}$, or equivalently, the top-r eigenspace of the Gram matrix $\boldsymbol{Y}\boldsymbol{Y}^\top$, namely

$$\text{(vanilla SVD)} \quad \widehat{\boldsymbol{U}}_{\mathsf{svd}} \leftarrow \mathsf{eigs}_r(\boldsymbol{Y}\boldsymbol{Y}^\top), \tag{8}$$

where $\mathsf{eigs}_r(\cdot)$ stands for the leading rank-r eigensubspace of a matrix. While this approach works well when $n_2 = O(n_1)$, it suffers from some fundamental limitations in the case with $n_2 \gg n_1$ and heteroskedastic noise. To illustrate this point, direct calculation reveals that

$$\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} + \mathsf{diag}\left(\left[\sum_{j=1}^{n_2} \mathbb{E}[E_{i,j}^2]\right]_{1\le i\le n_1}\right). \tag{9}$$

When $n_2 \gg n_1$ and when the noise components are highly heteroskedastic, the set of diagonal entries $\{\sum_{j=1}^{n_2} \mathbb{E}[E_{i,j}^2]\}_{1\le i\le n_1}$ might vary drastically, thereby resulting in a large deviation between the top-r eigenspace of $\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top]$ and that of $\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ (which is the desirable $\boldsymbol{U}^\star$).

Algorithm 1: HeteroPCA($\boldsymbol{G}_{\mathsf{in}}, r, t_{\mathsf{max}}$) (Zhang, Cai and Wu (2022))

input: symmetric matrix $\boldsymbol{G}_{\mathsf{in}}$, rank r, number of iterations t_{max}.
initialization: $\boldsymbol{G}^0 = \boldsymbol{G}_{\mathsf{in}}$.
for $t = 0, 1, \ldots, t_{\mathsf{max}}$ **do**
  $\boldsymbol{U}^t \boldsymbol{\Lambda}^t \boldsymbol{U}^{t\top} \leftarrow$ rank-r leading eigendecomposition of $\boldsymbol{G}^t$.
  $\boldsymbol{G}^{t+1} = \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{G}^t) + \mathcal{P}_{\mathsf{diag}}(\boldsymbol{U}^t \boldsymbol{\Lambda}^t \boldsymbol{U}^{t\top})$.
output: matrix estimate $\boldsymbol{G} = \boldsymbol{G}^{t_{\mathsf{max}}}$ and subspace estimate $\boldsymbol{U} = \boldsymbol{U}^{t_{\mathsf{max}}}$.

- *Diagonal-deleted PCA*. In an effort to rectify the above limitation of the vanilla SVD-based approach, prior works have put forward a solution called "diagonal-deleted PCA," which suppresses the influence of the diagonal entries of $\boldsymbol{Y}\boldsymbol{Y}^\top$ by suppressing them (Koltchinskii and Giné (2000), Florescu and Perkins (2016), Cai et al. (2021), Ndaoud, Sigalla and Tsybakov (2022), Ndaoud (2022), Abbe, Fan and Wang (2022)); that is, this approach outputs

$$\text{(diagonal-deleted PCA)} \quad \widehat{\boldsymbol{U}}_{\mathsf{del}} \leftarrow \mathsf{eigs}_r(\boldsymbol{Y}\boldsymbol{Y}^\top - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)), \tag{10}$$

  where $\mathcal{P}_{\mathsf{diag}}$ denotes Euclidean projection onto the set of diagonal matrices. When the diagonal entries of $\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ are sufficiently small, we have

$$\mathbb{E}[\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)] = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}) \approx \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} = \boldsymbol{U}^\star\boldsymbol{\Sigma}^{\star 2}\boldsymbol{U}^{\star\top},$$

  which forms the rationale of this approach.
- *The* HeteroPCA *algorithm*. The above diagonal-deleted approach can be further improved. Employing (10) as an initialization, Zhang, Cai and Wu (2022) put forward the HeteroPCA algorithm that combines the spectral method with successively refined diagonal estimates; more precisely, HeteroPCA initializes $\boldsymbol{G}$ as $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$, and alternates between the following two steps until convergence:

$$\begin{aligned}\text{(HeteroPCA)} \quad \text{repeat} \quad &\text{(i) } \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \leftarrow \text{rank-}r \text{ eigendecomposition of } (\boldsymbol{G});\\ &\text{(ii) } \boldsymbol{G} \leftarrow \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top) + \mathcal{P}_{\mathsf{diag}}(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top).\end{aligned}$$

  See Algorithm 1 for a complete description of this procedure, with the input matrix (or initialization) chosen to be $\boldsymbol{G}_{\mathsf{in}} = \boldsymbol{Y}\boldsymbol{Y}^\top - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$. The key lies in employing the improved diagonal estimates to help alleviate the bias induced by diagonal deletion.

When the condition number $\sigma_1^\star/\sigma_r^\star$ is large, however, the magnitude of the diagonal entries of $\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ can be substantially larger than, say, the square of the least singular value of $\boldsymbol{X}^\star$ (i.e., $\sigma_r^{\star 2}$). If this is the case, then diagonal-deleted PCA might erase a significant fraction of the useful signal, resulting in loss of effectiveness. This issue carries over to HeteroPCA, as its initialization—which is based on diagonal-deleted PCA—might already be highly unreliable.

The proposed algorithm: Deflated-HeteroPCA. We now describe how to alleviate the above curse of ill-conditioning. One lesson that we have learned from past HeteroPCA theory (Zhang, Cai and Wu (2022), Yan, Chen and Fan (2024)) is that: this procedure works well if (i) the condition number of the truth is well controlled and (ii) the least singular value is not buried by noise. Motivated by this fact, we propose to divide the set of eigenvalues of interest into "well-conditioned" subblocks that are sufficiently separated from each other, and include more subblocks one by one. More precisely, the main ideas of the proposed algorithm are as follows:

Algorithm 2: Deflated-HeteroPCA

input: data matrix $\boldsymbol{Y}$ (cf. (3)), rank r, maximum number of iterations t_i, $i = 1, 2, \ldots$
initialization: $k = 0$, $r_0 = 0$, $\boldsymbol{G}_0 = \mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$.
while $r_k < r$ **do**
  $k = k + 1$.
  select r_k via Equation (12).
  $(\boldsymbol{G}_k, \boldsymbol{U}_k) =$HeteroPCA$(\boldsymbol{G}_{k-1}, r_k, t_k)$.
output: subspace estimate $\boldsymbol{U} = \boldsymbol{U}_k$.

(1) Sequentially identify a collection of ranks $r_0 = 0 < r_1 < r_2 < \cdots < r_{k_{\max}} = r$, which partitions the set of eigenvalues (or singular values) of interest into disjoint subblocks. These points are chosen to ensure that (i) $\sigma^\star_{r_{k-1}+1}/\sigma^\star_{r_k}$ is sufficiently small for each k, and (ii) there is a sufficient gap between $\sigma^\star_{r_k}$ and $\sigma^\star_{r_k+1}$. Given that we do not know the true singular values *a priori*, we shall make careful use of the singular values of our running estimates instead.

(2) In the kth round, we invoke HeteroPCA with the rank r_k and the initialization $\boldsymbol{G}_{k-1}$ to impute the diagonal entries and obtain an improved estimate $\boldsymbol{G}_k$ of the Gram matrix of interest. Here, the first iteration employs the diagonal-deleted version $\boldsymbol{G}_0 = \mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$.

It then boils down to how to select the aforementioned ranks $\{r_k\}$ in a data-driven manner. Toward this end, we look at the following set of ranks in the kth round:[1]

$$\mathcal{R}_k := \Big\{ r' : r_{k-1} < r' \le r,\ \frac{\sigma_{r_{k-1}+1}(\boldsymbol{G}_{k-1})}{\sigma_{r'}(\boldsymbol{G}_{k-1})} \le 4 \text{ and } \sigma_{r'}(\boldsymbol{G}_{k-1}) - \sigma_{r'+1}(\boldsymbol{G}_{k-1}) \ge \frac{1}{r}\sigma_{r'}(\boldsymbol{G}_{k-1}) \Big\}, \tag{11}$$

and select r_k as follows:

$$r_k = \begin{cases} \max \mathcal{R}_k & \text{if } \mathcal{R}_k \neq \varnothing, \\ r & \text{otherwise.} \end{cases} \tag{12}$$

Here, we remind the readers that $\sigma_i(\boldsymbol{G}_{k-1})$ is the ith singular value of $\boldsymbol{G}_{k-1}$. Evidently, the first condition in (11) is imposed to ensure well conditioning of each subblock, whereas the second condition in (11) aims to guarantee a sufficient spectral separation between adjacent subblocks.

In a nutshell, the proposed algorithm counters the bias effect initially incurred by diagonal deletion via successive "deflation," a term that we borrow from Dobriban and Owen (2019) (although the problem considered therein is drastically different). More concretely, we first estimate the first subblock (which contains the largest eigenvalues of interest) by means of the diagonal deletion idea; once we finish estimating the eigensubspace associated with this subblock, we can readily compensate for the contribution of this subblock in the diagonal of interest. This strategy is then repeated subblock by subblock in order to successively reduce—or "deflate"—the original bias in the diagonal. For this reason, we refer to the proposed algorithm as Deflated-HeteroPCA, whose complete details are summarized in Algorithm 2.

The computation cost of Deflated-HeteroPCA (Algorithm 2) is $\widetilde{O}(n_1^2 n_2 + n_1^2 r \sum_{k=1}^{k_{\max}} t_k)$. Here, $\widetilde{O}(b)$ is equivalent to $O(b)$ except that it hides the logarithmic factors. The computational cost of the initialization step is $O(n_1^2 n_2)$. For other steps, the main computation cost

[1] The threshold 4 in (11) can be replaced with any numerical constant $C_{\text{gap}} \ge 4$.

is attributed to the top-r_k eigendecomposition, which amounts to $\widetilde{O}(n_1^2 r)$. Numerically, by setting all t_k's equal to 10, the algorithm performs well and the computational cost simplifies to $\widetilde{O}(n_1^2 n_2 + n_1^2 r k_{\mathsf{max}}) = \widetilde{O}(n_1^2 n_2 + n_1^2 r^2)$ (recall that the number of blocks k_{max} is at most r). As a comparison, the computation cost of HeteroPCA is $\widetilde{O}(n_1^2 n_2 + n_1^2 r t)$, where t is the number of iterations. As a result, it can be seen that Deflated-HeteroPCA does not incur a higher computational burden than HeteroPCA when $r = O(\sqrt{n_2})$.

4. Main theory. In this section, we demonstrate the desirable statistical performance for the proposed algorithm, which enjoys substantially improved dependency on the condition number. Before continuing, we find it helpful to introduce the following rotation matrix for any $\boldsymbol{U} \in \mathcal{O}^{n_1, r}$:

$$\boldsymbol{R}_{\boldsymbol{U}} = \arg\min_{\boldsymbol{R} \in \mathcal{O}^{r,r}} \|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}}, \tag{13}$$

the one that best aligns $\boldsymbol{U}$ with $\boldsymbol{U}^\star$ in the Euclidean sense; after all, it is in general infeasible to resolve the ambiguity brought by global rotation. As is well known in the literature (e.g., Ma et al. (2020), Section D.2.1),

$$\boldsymbol{R}_{\boldsymbol{U}} = \mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star), \tag{14}$$

where $\mathsf{sgn}(\cdot)$ is defined in (2).

4.1. *Spectral-norm-based statistical guarantees.* Let us begin with statistical guarantees based on the spectral norm accuracy. The following theorem asserts that the proposed Deflated-HeteroPCA algorithm enjoys appealing theoretical guarantees in terms of the spectral norm error $\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|$, no matter how large the condition number of $\boldsymbol{\Sigma}^\star$ is. The proof of this theorem is deferred to Section A in the Supplementary Material (Zhou and Chen (2025)).

THEOREM 1. *Suppose that Assumption 1 holds. Assume that*

$$\sigma_r^\star \geq C_0 r(\omega_{\mathsf{col}} + \sqrt{\omega_{\mathsf{col}}\omega_{\mathsf{row}}})\sqrt{\log n}, \tag{15a}$$

$$\mu \leq c_0 \frac{n_1}{r^3}, \tag{15b}$$

$$0 < \mu r \omega_{\mathsf{max}}^2 \leq \omega_{\mathsf{col}}^2 \tag{15c}$$

for some sufficiently large (resp., small) constant $C_0 > 0$ (resp., $c_0 > 0$). If the numbers of iterations obey

$$t_k > \log\left(C \frac{\sigma_{r_{k-1}+1}^{\star 2}}{\sigma_{r_k+1}^{\star 2}}\right), \quad 1 \leq k < k_{\mathsf{max}}, \tag{16a}$$

$$t_{k_{\mathsf{max}}} > \log\left(C \frac{\sigma_{r_{k_{\mathsf{max}}-1}+1}^{\star 2}}{\omega_{\mathsf{max}}^2}\right) \tag{16b}$$

for some large enough constant $C > 0$, then with probability exceeding $1 - O(n^{-10})$, the output returned by Algorithm 2 satisfies

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\| \lesssim \frac{\omega_{\mathsf{col}}\sqrt{\log n}}{\sigma_r^\star} + \frac{\omega_{\mathsf{col}}\omega_{\mathsf{row}}\log n}{\sigma_r^{\star 2}}. \tag{17}$$

Here, $r_0 = 0, r_1, \ldots, r_{k_{\mathsf{max}}}$ are the ranks selected in Algorithm 2 and k_{max} satisfies $r_{k_{\mathsf{max}}} = r$.

We find it helpful to compare our theoretical guarantees with prior theory for this problem. To begin with, the prior theory Zhang, Cai and Wu (2022) only covers the well-conditioned case; when κ is a bounded constant (as assumed therein), our statistical error bound (17) matches the one in Zhang, Cai and Wu (2022), Theorem 4 (up to some logarithmic factors).[2] In addition, when it comes to the case where $\omega_{i,j} \asymp \omega_{\mathsf{max}}$ for all $(i, j) \in [n_1] \times [n_2]$, our error bound (17) simplifies to

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\| \lesssim \frac{\sqrt{n_1 \log n}\omega_{\mathsf{max}}}{\sigma_r^\star} + \frac{\sqrt{n_1 n_2 \log^2 n}\omega_{\mathsf{max}}^2}{\sigma_r^{\star 2}},$$

which matches the minimax lower bounds (Cai et al. (2021), Theorem 2 and Cai and Zhang (2018), Theorem 4) ignoring logarithmic factors. It is noteworthy that when $\omega_{i,j} \asymp \omega_{\mathsf{max}}$ for all $(i, j) \in [n_1] \times [n_2]$ and $r = O(1)$, the signal-to-noise ratio condition (15a) simplifies to

$$\sigma_r^\star \gtrsim [(n_1 n_2)^{1/4} + n_1^{1/2}]\omega_{\mathsf{max}}\sqrt{\log n} \tag{18}$$

which is necessary to ensure—up to logarithmic factor—the existence of a consistent estimator (which means the existence of an estimator $\widehat{\boldsymbol{U}}$ obeying $\|\widehat{\boldsymbol{U}}\boldsymbol{R}_{\widehat{\boldsymbol{U}}} - \boldsymbol{U}^\star\| = o(1)$) (see Cai et al. (2021), Theorem 2).

4.2. *Fine-grained $\ell_{2,\infty}$-norm-based statistical guarantees.* Moving beyond the spectral norm bounds, we proceed to the fine-grained $\ell_{2,\infty}$-norm-based error bounds for column subspace estimation, which further capture how well the estimation error is spread out across the rows (Ma et al. (2020), Chen et al. (2020, 2019, 2021), Agterberg, Lubberts and Priebe (2022), Zhang and Zhou (2024), Cai et al. (2022)). As has been shown in the literature, such $\ell_{2,\infty}$-based subspace estimation guarantees play a crucial role in deriving performance bounds for the subsequent tasks like entrywise covariance estimation, entrywise tensor estimation, exact recovery in a variety of clustering and mixture models (Cai et al. (2021), Yan, Chen and Fan (2024), Abbe et al. (2020), Cai et al. (2021), Abbe, Fan and Wang (2022)).

Before formally presenting our $\ell_{2,\infty}$-norm-based result, we first introduce the following assumption on the noise matrix $\boldsymbol{E}$.

ASSUMPTION 2. Suppose that the noise components satisfy Condition 1 in Assumption 1. In addition, we assume that

$$\mathbb{P}\big(|E_{i,j}| > B\big) \le n^{-12}, \tag{19}$$

where B satisfies, for some universal constant $C_{\mathsf{b}} > 0$, that

$$B \le C_{\mathsf{b}}\omega_{\mathsf{max}}\frac{\min\{(n_1 n_2)^{1/4}, \sqrt{n_2}\}}{\log n}.$$

REMARK 2. Our assumptions on the noise are very mild and they hold across a diverse array of distributions, including

- uniform distributions;
- $C\omega_{\mathsf{max}}$-sub-Gaussian random variables;
- centered Poisson random variables with parameter $\lambda_{\mathsf{max}} = \omega_{\mathsf{max}}^2 \gtrsim \frac{\log^4 n}{\min\{(n_1 n_2)^{1/2}, n_2\}}$;
- centered Bernoulli random variables with $p_{i,j} \in [\frac{\log^2 n}{C_{\mathsf{b}}^2 \min\{(n_1 n_2)^{1/2}, n_2\}}, 1 - \frac{\log^2 n}{C_{\mathsf{b}}^2 \min\{(n_1 n_2)^{1/2}, n_2\}}]$.

[2] Zhang, Cai and Wu (2022) establish estimation guarantees for the $\sin\Theta$ distance $\|\sin\Theta(\widehat{\boldsymbol{U}}, \boldsymbol{U}^\star)\|$, which is (nearly) equivalent to the metric $\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}^\star\|$ (or more precisely, $\|\sin\Theta(\widehat{\boldsymbol{U}}, \boldsymbol{U}^\star)\| \asymp \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}^\star\|$). See Chen et al. (2021), Lemma 2.6 for details.

In addition, it is worth noting that the constant 12 can be replaced by any other constant $c > 2$ to ensure a high probability result. Here, we choose 12 simply to guarantee that the final estimation error bound holds with probability exceeding $1 - O(n^{-10})$. With the logarithmic factors neglected, the only difference between Assumption 2 and (Cai et al. (2021), Assumption 2) is that no symmetric distribution requirement is needed in Assumption 2.

Built upon Assumption 2, we derive the following $\ell_{2,\infty}$-based theoretical guarantees for Deflated-HeteroPCA, with the proof postponed to Section B in the Supplementary Material (Zhou and Chen (2025)).

THEOREM 2. *Suppose that Assumption 2 holds and the signal-to-noise ratio satisfies*

$$\frac{\sigma_r^\star}{\omega_{\max}} \geq C_0 r\big[(n_1 n_2)^{1/4} + n_1^{1/2}\big]\log n, \tag{20a}$$

$$\mu \leq c_0 \frac{n_1}{r^3} \tag{20b}$$

for some large (resp., small) enough constant $C_0 > 0$ (resp., $c_0 > 0$). If the numbers of iterations satisfy (16), then with probability exceeding $1 - O(n^{-10})$, then the estimate returned by Algorithm 2 satisfies

$$\big\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\big\|_{2,\infty} \lesssim \sqrt{\frac{\mu r}{n_1}}\zeta_{\mathsf{op}}, \tag{21a}$$

$$\big\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\big\| \lesssim \zeta_{\mathsf{op}}, \tag{21b}$$

where

$$\zeta_{\mathsf{op}} = \frac{\sqrt{n_1 n_2}\omega_{\max}^2 \log^2 n}{\sigma_r^{\star 2}} + \frac{\sqrt{n_1}\omega_{\max}\log n}{\sigma_r^\star}. \tag{22}$$

Encouragingly, both the $\ell_{2,\infty}$-based and spectral-norm-based estimation guarantees in (21) match the minimax lower bounds previously established in (Cai et al. (2021), Theorem 2) (up to logarithmic factors), thus confirming the near minimax optimality of our results. It can also been seen from (Cai et al. (2021), Theorem 2) that the signal-to-noise ratio requirement (20a) is, in general, essential (ignoring logarithmic factors) in order to enable the plausibility of consistent estimation.

Comparison with prior results. In order to demonstrate the utility of our algorithm and the accompanying theory, we compare our results with past works in the sequel. To ease presentation, the discussion below focuses attention on the case where $\mu, r = O(1)$.

- *Requirement on the condition number κ.* In order to obtain a consistent estimator,[3] all prior theory for both diagonal-deleted PCA (see Cai et al. ((2021), Theorem 1)) and HeteroPCA (see Zhang, Cai and Wu (2022), Theorem 4, Yan, Chen and Fan (2024), Theorem 5 and Agterberg, Lubberts and Priebe (2022), Assumption 4) assumes the condition number κ to obey

$$\text{(prior requirement on } \kappa) \quad \kappa \lesssim n_1^{1/4}, \tag{23}$$

in order to control the bias incurred during the diagonal deletion step. This, however, falls short of accommodating a wider range of condition numbers. In contrast, our result in Theorem 2 does not impose any assumptions on the condition number.

[3]Here, a column subspace estimator $\widehat{\boldsymbol{U}}$ is said to be consistent if $\min_{\boldsymbol{R}\in\mathcal{O}^{r,r}} \|\widehat{\boldsymbol{U}}\boldsymbol{R} - \boldsymbol{U}^\star\| = o(1)$.

- *Statistical error bounds.* We now compare our statistical error bounds with the ones obtained in Cai et al. (2021), Agterberg, Lubberts and Priebe (2022), Yan, Chen and Fan (2024). For notational convenience, define

$$\mathcal{E}_{\mathsf{noise}} := \frac{\sqrt{n_1 n_2}\omega_{\max}^2 \log n}{\sigma_r^{\star 2}} + \frac{\kappa \omega_{\max}\sqrt{n_1 \log n}}{\sigma_r^{\star}}, \tag{24}$$

  which makes it more convenient for us to describe the previous results.
  - Under the signal-to-noise ratio condition

$$\frac{\sigma_r^{\star}}{\omega_{\max}} \gtrsim \big(\kappa (n_1 n_2)^{1/4} + \kappa^3 n_1^{1/2}\big)\sqrt{\log n}, \tag{25}$$

    (Cai et al. (2021), Theorem 1) asserts that the estimate $\widehat{\boldsymbol{U}}_{\mathsf{del}}$ returned by diagonal-deleted PCA obeys, with high probability,

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r,r}} \big\|\widehat{\boldsymbol{U}}_{\mathsf{del}}\boldsymbol{R} - \boldsymbol{U}^{\star}\big\|_{2,\infty} \lesssim \kappa^2 \sqrt{\frac{1}{n_1}}(\mathcal{E}_{\mathsf{noise}} + \mathcal{E}_{\mathsf{diag\text{-}del}}), \tag{26}$$

    where $\mathcal{E}_{\mathsf{diag\text{-}del}}$ is an additional error term due to the bias resulting from diagonal deletion.
  - Focusing on the case where $n_2 \gtrsim n_1$, (Agterberg, Lubberts and Priebe (2022), Theorem 2) establishes an $\ell_{2,\infty}$ error bound for the HeteroPCA estimate $\widehat{\boldsymbol{U}}_{\mathsf{hpca}}$ as follows:

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r,r}} \big\|\widehat{\boldsymbol{U}}_{\mathsf{hpca}}\boldsymbol{R} - \boldsymbol{U}^{\star}\big\|_{2,\infty} \lesssim \sqrt{\frac{1}{n_1}}\mathcal{E}_{\mathsf{noise}}, \tag{27}$$

    albeit under a much more stringent SNR requirement:

$$\sigma_r^{\star} \gg \kappa \omega_{\max}\sqrt{n_2 \log n}. \tag{28}$$

  - (Yan, Chen and Fan (2024), Theorem 5) further shows that under the same SNR condition (25), HeteroPCA yields an estimator $\widehat{\boldsymbol{U}}_{\mathsf{hpca}}$ with the following high-probability $\ell_{2,\infty}$ error bound:

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r,r}} \big\|\widehat{\boldsymbol{U}}_{\mathsf{hpca}}\boldsymbol{R} - \boldsymbol{U}^{\star}\big\|_{2,\infty} \lesssim \kappa^2 \sqrt{\frac{1}{n_1}}\mathcal{E}_{\mathsf{noise}}. \tag{29}$$

  Let us compare our bounds with the above results. Recognizing that $\mathcal{E}_{\mathsf{noise}}$ is at least as large as ζ_{op} if we ignore logarithmic factors, our $\ell_{2,\infty}$ error bound (21a) improves the theoretical guarantees (26) and (29) by at least a factor of κ^2. Additionally, our bound (21a) outperforms the bound (27) in terms of the dependency on κ (ignoring logarithmic factors).
- *SNR requirement.* Let us also briefly make comparisons regarding the SNR required for consistent estimation. To begin with, we make note that the vanilla SVD-based approach (cf. (8)) requires the SNR to exceed (Cai et al. (2021), Zhang, Cai and Wu (2022))

$$\frac{\sigma_r^{\star}}{\omega_{\max}} \gtrsim \sqrt{n_1} + \sqrt{n_2}, \tag{30}$$

  which can be substantially more stringent than the one required in (20a) if $n_2 \gg n_1$. In addition, compared with the SNR requirement imposed in the existing theory for diagonal-deleted PCA and HeteroPCA, our condition (20a) is weaker than the one used in Cai et al. (2021) and Yan, Chen and Fan (2024) (see (25)) by at least a factor of κ, while at the same time being weaker than the condition (28) assumed in Agterberg, Lubberts and Priebe (2022) by a factor of $\kappa(n_2/n_1)^{1/4}$ when $n_2 \gg n_1$.

High-level proof strategy. While the proofs of our main theorems are deferred to the Supplementary Material, we highlight some novelty and technical challenges in our proof. In an attempt to obtain fine-grained $\ell_{2,\infty}$ control while remaining condition-number-free, we develop a new proof strategy that differs drastically from the state-of-the-art techniques based on leave-one-out decoupling arguments (Yan, Chen and Fan (2024), Cai et al. (2021)). Inspired by a spectral representation lemma derived in the recent work Xia (2021) (see also Lemma 1), we proceed by decomposing the difference between the subspaces into an infinite sum of polynomials of the error matrix. With this decomposition at hand, one major part of our proof hinges upon establishing sharp $\ell_{2,\infty}$ bounds on each of the polynomials of the error matrix. The key challenge for this part lies in how to deal with the complicated and accumulated dependence brought by the power of the error matrix, for which we resort to careful induction analyses. We will then single out several sequences of critical quantities and develop intricate arguments to control these quantities in a recursive and inductive manner.

5. Consequences for specific models. To better illustrate the effectiveness of the proposed algorithm, we develop concrete consequences of our theory in Section 4 for two specific models. In each case, we shall begin by describing the model, followed by concrete algorithms and theory tailored to the specific model.

5.1. *Factor models and spiked covariance models.*

Model. A frequently studied model employed to capture low-dimensional structure in high-dimensional sample data is the factor model, which finds applications numerous contexts including finance and econometrics (Lawley and Maxwell (1962), Fan et al. (2020, 2021)), functional magnetic resonance imaging (Chen et al. (2015)) and signal processing (Zhao, Krishnaiah and Bai (1986), Kritchman and Nadler (2008, 2009)), to name just a few. For concreteness, suppose that we observe a collection of n independent sample vectors in $\mathbb{R}^d$ generated as follows:

$$\boldsymbol{y}_j = \boldsymbol{B}^\star \boldsymbol{f}_j + \boldsymbol{\varepsilon}_j \in \mathbb{R}^d, \tag{31a}$$

where $\boldsymbol{B}^\star \in \mathbb{R}^{d\times r}$ represents the factor loading matrix with $r \ll d$, $\{\boldsymbol{f}_j\}$ stands for the latent factor vectors, and $\{\boldsymbol{\varepsilon}_j\}$ denotes the noise vectors. We assume that

$$\boldsymbol{B}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^{\star 1/2} \in \mathbb{R}^{d\times r} \quad \text{and} \quad \boldsymbol{f}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r), \quad 1 \le j \le n, \tag{31b}$$

with $\boldsymbol{U}^\star \in \mathcal{O}^{d,r}$ and $\boldsymbol{\Lambda}^\star = \mathsf{diag}(\lambda_1^\star, \ldots, \lambda_r^\star)$ being a diagonal matrix containing all eigenvalues of $\boldsymbol{B}^\star \boldsymbol{B}^{\star\top}$. Equivalently, one can express it as the following spiked covariance model:

$$\boldsymbol{y}_j = \boldsymbol{x}_j + \boldsymbol{\varepsilon}_j \quad \text{with } \boldsymbol{x}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top}\big), 1 \le j \le n. \tag{32}$$

The noise vectors are allowed to be heteroskedastic, and it is assumed that

- the $\varepsilon_{i,j}$'s are statistically independent, zero-mean and ω-sub-Gaussian,

where $\omega > 0$ is an upper bound on the sub-Gaussian norm of any noise entry. We also assume that

$$\|\boldsymbol{U}^\star\|_{2,\infty} \le \sqrt{\frac{\mu_{\mathsf{pc}} r}{d}}. \tag{33}$$

Our goal is to estimate the subspace $\boldsymbol{U}^\star$ based on the observed vectors $\{\boldsymbol{y}_i\}_{1\le i\le n}$.

5.1.1. *Algorithm and theoretical guarantees.* Taking the data matrix as $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_n] \in \mathbb{R}^{d\times n}$, we can readily invoke Algorithm 2 to estimate the subspace $\boldsymbol{U}^\star$. The performance guarantees are stated below, whose proof is deferred to Section C.1.

COROLLARY 1. *Consider the factor model in* (31). *Assume that*

$$\frac{\lambda_r^\star}{\omega^2} \geq C_1 r^2 \left[\left(\frac{d}{n}\right)^{1/2} + \frac{d}{n}\right] \log^2(n+d), \tag{34a}$$

$$\mu_{\mathsf{pc}} \vee \log(n+d) \leq c_1 \frac{d}{r^3}, \tag{34b}$$

$$r \vee \log(n+d) \leq c_1 n \tag{34c}$$

for some sufficiently large (*resp.*, *small*) *constant* $C_1 > 0$ (*resp.*, $c_1 > 0$). *Suppose that the numbers of iterations obey, for some large enough constant* $C > 0$,

$$t_k \geq \log_2\left(C \frac{\lambda^\star_{r_{k-1}+1}}{\lambda^\star_{r_k+1}}\right) \quad \forall 1 \leq k \leq k_{\mathsf{max}} - 1, \tag{35a}$$

$$t_{k_{\mathsf{max}}} \geq \log\left(C \frac{n\lambda^\star_{r_{k_{\mathsf{max}}-1}+1}}{\omega^2}\right), \tag{35b}$$

where k_{max} *satisfies* $r_{k_{\mathsf{max}}} = r$. *Then with probability exceeding* $1 - O((n+d)^{-10})$, *the output* $\boldsymbol{U}$ *returned by Algorithm 2 satisfies*

$$\begin{aligned} &\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|_{2,\infty} \\ &\lesssim \sqrt{\frac{(\mu_{\mathsf{pc}} + \log(n+d))r}{d}} \left(\frac{\sqrt{d/n}\omega^2 \log^2(n+d)}{\lambda_r^\star} + \frac{\sqrt{d/n}\omega \log(n+d)}{\sqrt{\lambda_r^\star}}\right), \end{aligned} \tag{36a}$$

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\| \lesssim \frac{\sqrt{d/n}\omega^2 \log^2(n+d)}{\lambda_r^\star} + \frac{\sqrt{d/n}\omega \log(n+d)}{\sqrt{\lambda_r^\star}}. \tag{36b}$$

Let us briefly discuss the implications of our results. Consider, for example, the case where $\mathbb{E}[\varepsilon_{i,j}^2] \asymp \sigma^2$ for all $(i,j) \in [d] \times [n]$. The spectral norm bound (36b) matches the minimax limit (see (Zhang, Cai and Wu (2022), Theorems 1 and 4) modulo some logarithmic factor. In addition, recognizing that

$$d\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|_{2,\infty}^2 \geq \|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|_{\mathrm{F}}^2 \geq \|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|^2,$$

we see that the $\ell_{2,\infty}$ bound (36a) is also near optimal when $\mu_{\mathsf{pc}}, r \asymp 1$. Again, our result does not rely on the condition number $\kappa_{\mathsf{pc}} = \lambda_1^\star/\lambda_r^\star$. Moreover, Zhang, Cai and Wu (2022), Theorem 1, assumes that κ_{pc} is bounded by a numerical constant, while Cai et al. (2021), Corollary 2, requires $\kappa_{\mathsf{pc}} \lesssim \sqrt{\frac{d}{\mu r}}$; these form another aspect in which Corollary 1 improves upon the prior literature.

5.2. *Tensor PCA.*

Model. Another canonical example in which column subspace estimation plays a key role is tensor PCA (or low-rank tensor estimation), a problem that has been studied extensively in recent literature (Richard and Montanari (2014), Zhang and Xia (2018), Cai et al. (2021, 2022), Han, Willett and Zhang (2022), Zhou et al. (2022), Han and Zhang (2023)). To be precise, assume that we observe a noisy tensor as follows:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}^\star + \boldsymbol{\mathcal{E}} \in \mathbb{R}^{n_1\times n_2\times n_3}, \tag{37a}$$

where $\boldsymbol{\mathcal{X}}^\star$ is an unknown low-rank tensor to be estimated, and $\boldsymbol{\mathcal{E}}$ represents the noise tensor. We assume that $\boldsymbol{\mathcal{X}}^\star$ has low-Tucker rank in the sense that (Zhang, Cai and Wu (2022), Han and Zhang (2023), Xia, Zhang and Zhou (2022))

$$\boldsymbol{\mathcal{X}}^\star = \boldsymbol{\mathcal{S}}^\star \times_1 \boldsymbol{U}_1^\star \times_2 \boldsymbol{U}_2^\star \times_3 \boldsymbol{U}_3^\star, \tag{37b}$$

where the core tensor $\boldsymbol{\mathcal{S}}^\star$ lies in $\mathbb{R}^{r_1\times r_2\times r_3}$ (with small r_1, r_2, r_3), and the tensor "principal components" $\boldsymbol{U}_i^\star \in \mathcal{O}^{n_i,r_i}$ $(1 \le i \le 3)$ satisfy the incoherence condition

$$\|\boldsymbol{U}_i^\star\|_{2,\infty} \le \sqrt{\frac{\mu r_i}{n_i}}, \quad 1 \le i \le 3. \tag{38}$$

Moreover, the noise tensor $\boldsymbol{\mathcal{E}} = [E_{i,j,k}]_{(i,j,k)\in[n_1]\times[n_2]\times[n_3]}$ is composed of independent entries such that

- the $E_{i,j,k}$'s are statistically independent, zero-mean and ω-sub-Gaussian,

where $\omega > 0$ is an upper bound on the sub-Gaussian norm of each noise entry. The aim is to compute a faithful estimate of the true tensor $\boldsymbol{\mathcal{X}}^\star$ as well as the principal components $\boldsymbol{U}_1^\star$, $\boldsymbol{U}_2^\star$ and $\boldsymbol{U}_3^\star$.

Additional notation. Before presenting the algorithm and our theoretical results, we introduce several useful notation. For any $1 \le i \le 3$ and $1 \le j \le r_i$, we denote by $\sigma_{i,j}^\star$ the jth largest singular value of the ith matricization of $\mathcal{X}$—denoted by $\mathcal{M}_i(\mathcal{X})$. Define

$$\sigma_{\min}^\star := \min\{\sigma_{1,r_1}^\star, \sigma_{2,r_2}^\star, \sigma_{3,r_3}^\star\},$$

and the condition number of the true tensor is then defined as

$$\kappa := \frac{\max\{\sigma_{1,1}^\star, \sigma_{2,1}^\star, \sigma_{3,1}^\star\}}{\sigma_{\min}^\star}.$$

For any $1 \le i \le 3$, we also let $r_{i,1}, r_{i,2}, \ldots, r_{i,k_{\max}^i}$ denote the ranks selected in Algorithm 2 if we apply this algorithm with the input matrix $\boldsymbol{Y} = \mathcal{M}_i(\boldsymbol{\mathcal{Y}})$, the rank r_i and the numbers of iterations $t_{i,1}, \ldots, t_{i,k_{\max}^i}$. As usual, we choose $k_{\max}^i$ such that $r_{k_{\max}^i} = r_i$. In addition, for notational convenience we let

$$n = \max_{1\le i\le 3} n_i \quad \text{and} \quad r = \max_{1\le i\le 3} r_i,$$

and define

$$\boldsymbol{U}_4^\star = \boldsymbol{U}_1^\star \quad \text{and} \quad \boldsymbol{U}_5^\star = \boldsymbol{U}_2^\star.$$

Algorithm and statistical guarantees. In order to apply Deflated-HeteroPCA, let us look at the matrix $\mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star) \in \mathbb{R}^{n_i\times(n_1n_2n_3)/n_i}$, the ith matricization of $\boldsymbol{\mathcal{X}}^\star$. Recognizing that $\boldsymbol{U}_i^\star$ is also the left singular space of $\mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star)$ since

$$\mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star) = \boldsymbol{U}_i^\star \mathcal{M}_i(\boldsymbol{\mathcal{S}}^\star)(\boldsymbol{U}_{i+2}^\star \otimes \boldsymbol{U}_{i+1}^\star),$$

we propose to apply the Deflated-HeteroPCA algorithm to compute an initial subspace estimate $\widehat{\boldsymbol{U}}_i^0$ for $\boldsymbol{U}_i^\star$. Armed with these initial estimates, we invoke the high-order orthogonal iteration (HOOI) algorithm (De Lathauwer, De Moor and Vandewalle (2000), Zhang and Xia (2018)) to iteratively refine the estimates. More specifically, in the tth iteration, we calculate

$$\widehat{\boldsymbol{U}}_i^t = \text{ the first } r \text{ left singular vectors of } \mathcal{M}_i(\boldsymbol{\mathcal{Y}} \times_{i+1} \widehat{\boldsymbol{U}}_{i+1}^{t-1} \times_{i+2} \widehat{\boldsymbol{U}}_{i+2}^{t-1}), \quad 1 \le i \le 3,$$

where $i+1$ and $i+2$ are calculated modulo 3. Once the above iterative procedure converges, we employ the resulting subspace estimates $\widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3$ to construct the following estimator for the true tensor:

$$\widehat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{Y}} \times_1 \mathcal{P}_{\widehat{\boldsymbol{U}}_1} \times_2 \mathcal{P}_{\widehat{\boldsymbol{U}}_2} \times_3 \mathcal{P}_{\widehat{\boldsymbol{U}}_3},$$

where we recall the notation $\mathcal{P}_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$.

Algorithm 3: High-order orthogonal iteration (HOOI) (De Lathauwer, De Moor and Vandewalle (2000), Zhang and Xia (2018))

input: $\mathcal{Y}$, ranks r_1, r_2, r_3, number of iterations $\{t_{i,j}\}_{1\le i\le 3, 1\le j\le k^i_{\mathsf{max}}}$ and t_{max}.

initialization: call Algorithm 2 to compute

$$\widehat{\boldsymbol{U}}_1^0 = \mathsf{Deflated\text{-}HeteroPCA}\big(\mathcal{M}_1(\boldsymbol{\mathcal{Y}}), r_1, t_{1,1}, t_{1,2}, \ldots, t_{1,k^1_{\mathsf{max}}}\big);$$

$$\widehat{\boldsymbol{U}}_2^0 = \mathsf{Deflated\text{-}HeteroPCA}\big(\mathcal{M}_2(\boldsymbol{\mathcal{Y}}), r_2, t_{2,1}, t_{2,2}, \ldots, t_{2,k^2_{\mathsf{max}}}\big);$$

$$\widehat{\boldsymbol{U}}_3^0 = \mathsf{Deflated\text{-}HeteroPCA}\big(\mathcal{M}_3(\boldsymbol{\mathcal{Y}}), r_3, t_{3,1}, t_{3,2}, \ldots, t_{3,k^3_{\mathsf{max}}}\big).$$

while $t < t_{\mathsf{max}}$ **do**

$\widehat{\boldsymbol{U}}_1^t =$ leading r_1 left singular vectors of $\mathcal{M}_1(\boldsymbol{\mathcal{Y}} \times_2 \widehat{\boldsymbol{U}}_2^{t-1} \times_3 \widehat{\boldsymbol{U}}_3^{t-1})$.

$\widehat{\boldsymbol{U}}_2^t =$ leading r_2 left singular vectors of $\mathcal{M}_1(\boldsymbol{\mathcal{Y}} \times_3 \widehat{\boldsymbol{U}}_3^{t-1} \times_1 \widehat{\boldsymbol{U}}_1^{t-1})$.

$\widehat{\boldsymbol{U}}_3^t =$ leading r_3 left singular vectors of $\mathcal{M}_3(\boldsymbol{\mathcal{Y}} \times_1 \widehat{\boldsymbol{U}}_1^{t-1} \times_2 \widehat{\boldsymbol{U}}_2^{t-1})$.

compute $\widehat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{Y}} \times_1 \widehat{\boldsymbol{U}}_1^{t_{\mathsf{max}}} \widehat{\boldsymbol{U}}_1^{t_{\mathsf{max}}\top} \times_2 \widehat{\boldsymbol{U}}_2^{t_{\mathsf{max}}} \widehat{\boldsymbol{U}}_2^{t_{\mathsf{max}}\top} \times_3 \widehat{\boldsymbol{U}}_3^{t_{\mathsf{max}}} \widehat{\boldsymbol{U}}_3^{t_{\mathsf{max}}\top}$.

output: subspace estimates $\widehat{\boldsymbol{U}}_1 = \widehat{\boldsymbol{U}}_1^{t_{\mathsf{max}}}$, $\widehat{\boldsymbol{U}}_2 = \widehat{\boldsymbol{U}}_2^{t_{\mathsf{max}}}$, $\widehat{\boldsymbol{U}}_3 = \widehat{\boldsymbol{U}}_3^{t_{\mathsf{max}}}$ and tensor estimate $\widehat{\boldsymbol{\mathcal{X}}}$.

The whole procedure is summarized in Algorithm 3, where $\mathsf{Deflated\text{-}HeteroPCA}(\boldsymbol{Y}, r, t_1, \ldots, t_{\mathsf{max}})$ is the output of Algorithm 2 with the input matrix $\boldsymbol{Y}$, the rank r and the numbers of iterations $t_1, \ldots, t_{\mathsf{max}}$. The computational cost for the initialization step (Deflated-HeteroPCA) is $\widetilde{O}(n^4 + n^2 r \sum_{i=1}^3 \sum_{j=1}^{k^i_{\mathsf{max}}} t_{i,j})$. For each orthogonal iteration, the computational cost is $\widetilde{O}(n^3 r^2 + n r^3)$. Therefore, the total computational complexity for Algorithm 3 amounts to $\widetilde{O}(n^4 + n^2 r \sum_{i=1}^3 \sum_{j=1}^{k^i_{\mathsf{max}}} t_{i,j} + (n^3 r^2 + n r^3) t_{\mathsf{max}})$. Numerically, the algorithm achieves great performance with all $t_{i,j}$'s and t_{max} set to 10, in which case the computational cost simplifies to $\widetilde{O}(n^4 + n^3 r^2)$. Our main theory for Deflated-HeteroPCA readily leads to the following statistical guarantees for Algorithm 3.

COROLLARY 2. *Consider the tensor PCA model in* (37). *Suppose that* $n_1 \asymp n_2 \asymp n_3 \asymp n$, *and*

$$\frac{\sigma^\star_{\mathsf{min}}}{\omega} \ge C_2 r n^{3/4} \log n, \tag{39a}$$

$$\mu \le c_2 \sqrt{\frac{n}{r^4}} \tag{39b}$$

for some sufficiently large (*resp.*, *small*) *constant* $C_2 > 0$ (*resp.*, $c_2 > 0$). *For any* $1 \le i \le 3$, *if one chooses*

$$t_{i,1} \ge \log_2\left(C \frac{\sigma^{\star 2}_{i,r_{i,k-1}+1}}{\sigma^{\star 2}_{i,r_{i,k}+1}}\right), \quad 1 \le k \le k^i_{\mathsf{max}} - 1, \tag{40a}$$

$$t_{i,k^i_{\mathsf{max}}} \ge \log\left(C \frac{\sigma^{\star 2}_{r_{i,k^i_{\mathsf{max}}-1}+1}}{\omega^2}\right), \tag{40b}$$

then with probability exceeding $1 - O(n^{-10})$, *the initial estimator* $\widehat{\boldsymbol{U}}_i^0$ *satisfies*

$$\|\widehat{\boldsymbol{U}}_i^0 \boldsymbol{R}_{\widehat{\boldsymbol{U}}_i^0} - \boldsymbol{U}_i^\star\|_{2,\infty} \lesssim \frac{\mu r}{\sqrt{n}}\left(\frac{n^{3/2}\omega^2 \log^2 n}{\sigma_{\min}^{\star 2}} + \frac{\sqrt{n}\omega \log n}{\sigma_{\min}^\star}\right), \tag{41a}$$

$$\|\widehat{\boldsymbol{U}}_i^0 \boldsymbol{R}_{\widehat{\boldsymbol{U}}_i^0} - \boldsymbol{U}^\star\| \lesssim \frac{n^{3/2}\omega^2 \log^2 n}{\sigma_{\min}^{\star 2}} + \frac{\sqrt{n}\omega \log n}{\sigma_{\min}^\star}. \tag{41b}$$

In addition, if the number of iterations in HOOI obeys $t_{\max} \geq C(\log(\frac{n}{\sigma_{\min}}) \vee 1)$ *for some large enough constant* $C > 0$, *then with probability exceeding* $1 - O(n^{-10})$ *one has*

$$\|\widehat{\boldsymbol{U}}_i \boldsymbol{R}_{\widehat{\boldsymbol{U}}_i} - \boldsymbol{U}_i^\star\| \lesssim \frac{\sqrt{n_i}\omega}{\sigma_{\min}^\star}, \quad 1 \leq i \leq 3, \tag{42a}$$

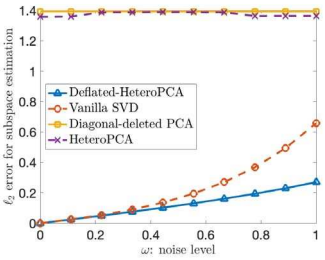
$$\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}^\star\|_{\mathrm{F}}^2 \lesssim (n_1 r_1 + n_2 r_2 + n_3 r_3)\omega^2. \tag{42b}$$

The bounds in (42) are rate optimal, since they match the minimax lower bounds established for the i.i.d. Gaussian noise case in Zhang and Xia (2018), Theorem 3. This confirms that the proposed Deflated-HeteroPCA algorithm serves as an effective paradigm to initialize the HOOI algorithm. It is also noteworthy that when $r = O(1)$, the SNR condition (39) is essential (ignoring logarithmic factor) to ensure that consistent estimation is computable within polynomial time; see Zhang and Xia (2018), Theorem 4.

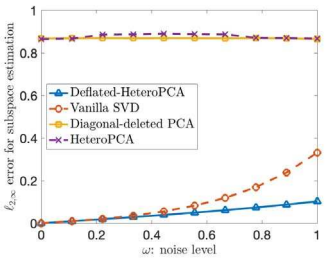
It is then helpful to compare our results with the prior works Zhang and Xia (2018) and Han, Willett and Zhang (2022). First, Zhang and Xia ((2018), Theorem 1) assume that the noise tensor has i.i.d. Gaussian entries, which is clearly much more stringent than our result. Second, while Han, Willett and Zhang ((2022), Theorem 4.1) allow the noise to be heteroskedastic, it requires the condition number of the tensor to be bounded (see the analysis for their main theorems); in comparison, our theory in Corollary 2 suggests that Algorithm 3 succeeds no matter how large the condition number κ is.

6. Numerical experiments. In this section, we conduct additional numerical experiments to verify the practical applicability of our algorithm. All results in this section are averaged over 50 Monte Carlo runs.

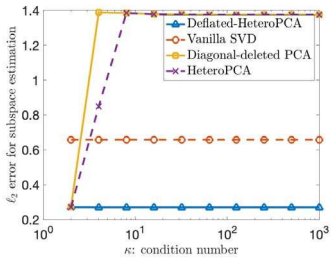
Low-rank subspace estimation from noisy observation. To begin with, we consider the problem of estimating the column subspace of $\boldsymbol{X}^\star$ from the noisy data (3). We randomly generate $\boldsymbol{U}^\star \in \mathcal{O}^{n_1,r}$ and $\boldsymbol{V}^\star \in \mathcal{O}^{n_2,r}$, and $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$, where $\boldsymbol{\Sigma}^\star = \mathsf{diag}(\sigma_1^\star, \ldots, \sigma_r^\star)$. For each $i \in [n_1]$, we independently and uniformly draw $\omega_i \in [0, \omega]$, whereas the $E_{i,j}$'s are independently drawn from $\mathcal{N}(0, \omega_i^2)$. We fix $n_1 = 100$, set $\sigma_r^\star = (n_1 n_2)^{1/4} + n_1^{1/2}$ and consider the following two settings: (i) $r = 3$, $\sigma_1^\star = \kappa\sigma_3^\star$ and $\sigma_2^\star = \sigma_3^\star$; (ii) $r = 5$, $\sigma_1^\star = \kappa\sigma_5^\star$, $\sigma_2^\star = \sigma_3^\star = \kappa^{1/2}\sigma_5^\star$ and $\sigma_4^\star = \sigma_5^\star$. We report the spectral-norm-based error $\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|$ and the $\ell_{2,\infty}$ error $\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|_{2,\infty}$ for each of the following four algorithms: (a) Deflated-HeteroPCA in Algorithm 2, where the numbers of iterations are chosen to be $t_i = 10$; (b) the diagonal-deleted PCA procedure as in (10); (c) HeteroPCA in Algorithm 1, where the number of iterations is taken to be 100; (d) the vanilla SVD-based approach described in (8). The results for $r = 3$ and $r = 5$ are reported in Figures 2 and 3, respectively. As can be seen from the plots, the proposed Deflated-HeteroPCA algorithm significantly outperforms the other three methods, and it is the only algorithm whose performance is unaffected by the condition number κ.



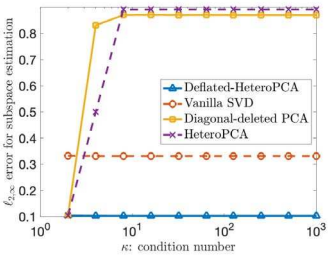
(a) $\kappa = 5, n_2 = 1,000$, ℓ_2 error



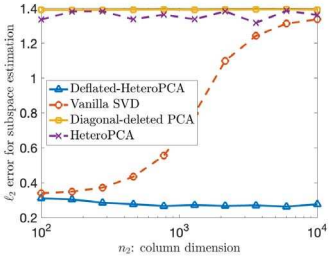
(b) $\kappa = 5, n_2 = 1,000$, $\ell_{2,\infty}$ error



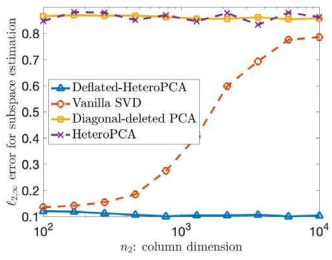
(c) $\omega = 1, n_2 = 1,000$, ℓ_2 error



(d) $\omega = 1, n_2 = 1,000$, $\ell_{2,\infty}$ error

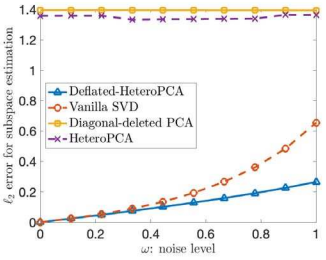


(e) $\omega = 1, \kappa = 5$, ℓ_2 error

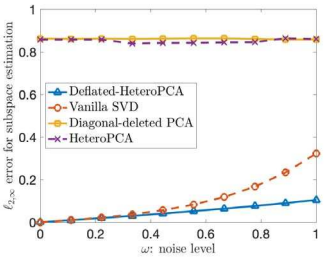


(f) $\omega = 1, \kappa = 5$, $\ell_{2,\infty}$ error

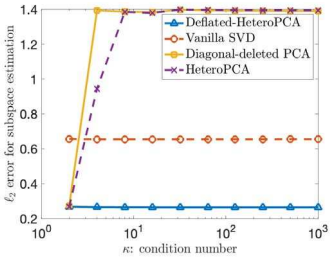
FIG. 2. *Estimation errors of* $\boldsymbol{U}$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *for* $r = 3$. *Plot* (*a*) (*resp.*, (*b*)) *reports the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the noise level* ω (*where* $n_1 = 100$, $n_2 = 1000$, $\kappa = 5$). *Plot* (*c*) (*resp.*, (*d*)) *shows the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the column dimension* κ (*where* $n_1 = 100$, $n_2 = 1000$, $\omega = 1$). *Plot* (*e*) (*resp.*, (*f*)) *displays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the condition number* n_2 (*where* $n_1 = 100$, $\kappa = 5$, $\omega = 1$).



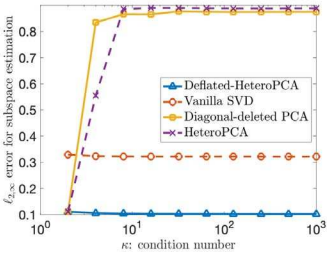
(a) $\kappa = 5, n_2 = 1,000$, ℓ_2 error



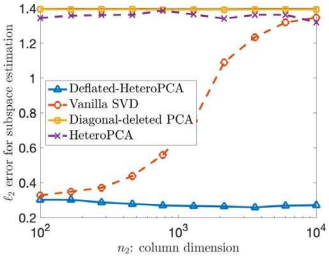
(b) $\kappa = 5, n_2 = 1,000$, $\ell_{2,\infty}$ error



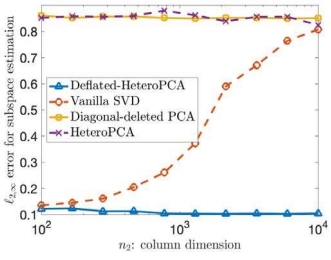
(c) $\omega = 1, n_2 = 1,000$, ℓ_2 error



(d) $\omega = 1, n_2 = 1,000$, $\ell_{2,\infty}$ error

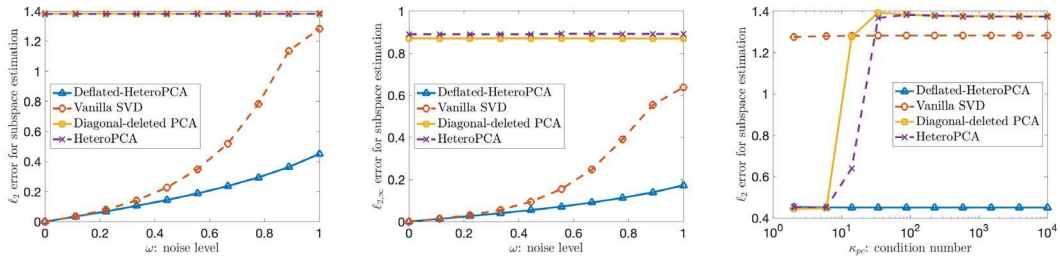


(e) $\omega = 1, \kappa = 5$, ℓ_2 error

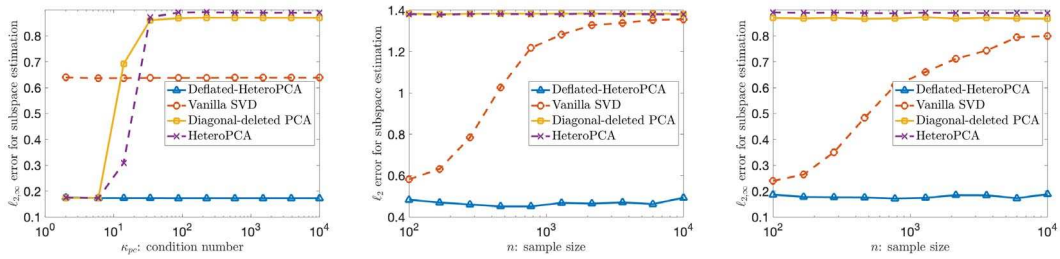


(f) $\omega = 1, \kappa = 5$, $\ell_{2,\infty}$ error

FIG. 3. *Estimation errors of* $\boldsymbol{U}$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *when* $r = 5$. *Plot* (*a*) (*resp.*, (*b*)) *displays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the noise level* ω (*where* $n_1 = 100$, $n_2 = 1000$, $\kappa = 5$). *Plot* (*c*) (*resp.*, (*d*)) *shows the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the condition number* κ (*where* $n_1 = 100$, $n_2 = 1000$, $\omega = 1$). *Plot* (*e*) (*resp.*, (*f*)) *displays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the column dimension* n_2 (*where* $n_1 = 100$, $\kappa = 5$, $\omega = 1$).



(a) $\kappa_{\mathsf{pc}} = 100, n = 1,000$, ℓ_2 error (b) $\kappa_{\mathsf{pc}} = 100, n = 1,000$, $\ell_{2,\infty}$ error (c) $\omega = 1, n = 1,000$, ℓ_2 error



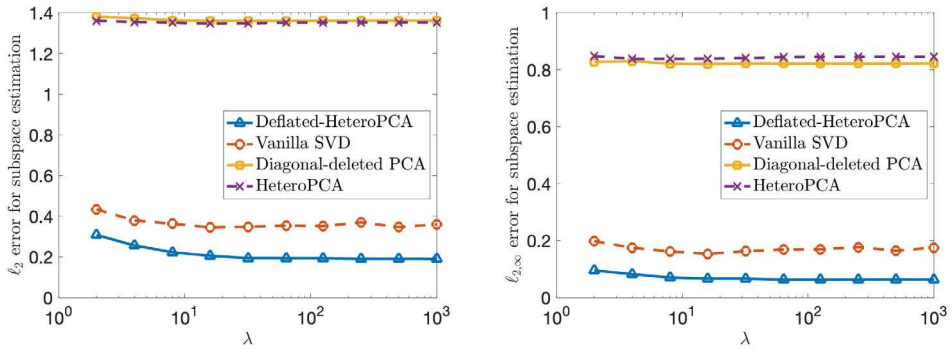
(d) $\omega = 1, n = 1,000$, $\ell_{2,\infty}$ error (e) $\omega = 1, \kappa_{\mathsf{pc}} = 100$, ℓ_2 error (f) $\omega = 1, \kappa_{\mathsf{pc}} = 100$, $\ell_{2,\infty}$ error

FIG. 4. *Estimation errors of* $\boldsymbol{U}$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *under the factor model* (32) *when* $r = 3$. *Plot* (*a*) (*resp.*, (*b*)) *displays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the noise level* ω (*where* $d = 100$, $n = 1000$, $\kappa_{\mathsf{pc}} = 100$). *Plot* (*c*) (*resp.*, (*d*)) *shows the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the condition number* κ_{pc} (*where* $d = 100$, $n = 1000$, $\omega = 1$). *Plot* (*e*) (*resp.*, (*f*)) *displays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the sample size* n (*where* $d = 100$, $\kappa_{\mathsf{pc}} = 100$, $\omega = 1$).

Factor model. We then turn attention to the factor model (32). We consider the case with $d = 100$, $r = 3$ and randomly generate the subspace $\boldsymbol{U}^\star \in \mathcal{O}^{d,3}$ and $\boldsymbol{F} = [\boldsymbol{f}_1 \ \cdots \ \boldsymbol{f}_n] \in \mathbb{R}^{3\times n}$ with i.i.d. standard Gaussian entries. We set the diagonal matrix $\boldsymbol{\Lambda}^\star = \mathsf{diag}(\lambda_1^\star, \lambda_2^\star, \lambda_3^\star)$ with $\lambda_1^\star = \kappa\lambda_3^\star$ and $\lambda_2^\star = \lambda_3^\star = (d/n)^{1/2} + d/n$. The noise matrix is generated in the same way as in the previous setting. We report in Figure 4 the ℓ_2 and $\ell_{2,\infty}$ errors for the principal subspace for the four methods, Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA and Vanilla SVD. The numerical results suggest that the proposed Deflated-HeteroPCA algorithm achieves the best performance among all these methods, which is not affected as κ_{pc} varies.

Poisson PCA. We consider the Poisson PCA problem (Zhang, Cai and Wu (2022), Liu, Dobriban and Singer (2018)): suppose that the truth $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} \in \mathbb{R}^{n_1\times n_2}$ is a rank-r matrix with positive entries. Our goal is to estimate the column subspace $\boldsymbol{U}^\star \in \mathbb{R}^{n_1\times r}$ based on the observations $\boldsymbol{Y} \in \mathbb{R}^{n_1\times n_2}$, where each entry $Y_{i,j}$ of $\boldsymbol{Y}$ is an independent random variable following a Poisson distribution with mean $X_{i,j}^\star$, that is, $Y_{i,j} \sim \mathsf{Poisson}(X_{i,j}^\star)$. More specifically, we fix $n_1 = 100$, $n_2 = 1000$, $r = 3$ and generate random matrices $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{n_1\times 3}$ and $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n_2\times 3}$ with i.i.d. standard Gaussian entries. We let $\overline{\boldsymbol{U}} \in \mathbb{R}^{n_1\times 3}$ (resp., $\overline{\boldsymbol{V}} \in \mathbb{R}^{n_2\times 3}$) denote the matrix with entries $\overline{U}_{i,j} = |\widetilde{U}_{i,j}|$ (resp., $\overline{V}_{i,j} = |\widetilde{V}_{i,j}|$). We define $\overline{\boldsymbol{\Lambda}} = \frac{1}{5}\mathsf{diag}(\lambda^2, \lambda, \lambda)$ and let $\boldsymbol{X}^\star = \overline{\boldsymbol{U}}\overline{\boldsymbol{\Lambda}}\overline{\boldsymbol{V}}^\top$. The empirical ℓ_2 and $\ell_{2,\infty}$ errors for the subspace estimation for the four methods, Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA and Vanilla SVD are illustrated in Figure 5. It is clearly seen that Deflated-HeteroPCA outperforms the other three methods.

Tensor PCA. Finally, we conduct numerical experiments for the tensor PCA model (37). We fix $n = 50$ and $r = 3$, and introduce a quantity $\sigma^\star = n^{3/4}$. The subspaces $\boldsymbol{U}_1^\star \in \mathcal{O}^{100,3}$, $\boldsymbol{U}_2^\star \in \mathcal{O}^{100,3}$ and $\boldsymbol{U}_3^\star \in \mathcal{O}^{100,3}$ are generated randomly, and the core tensor $\mathcal{S}^\star \in \mathbb{R}^{3\times 3}$ is a diagonal

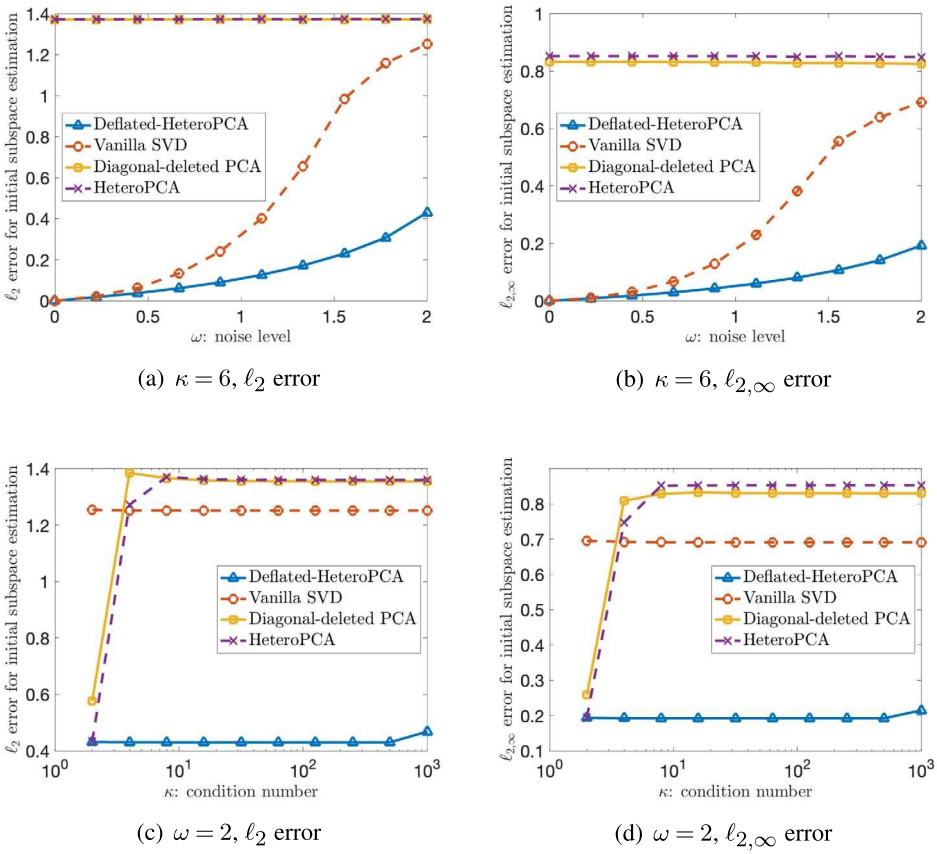


(a) $\kappa = 5, n_2 = 1,000$, ℓ_2 error (b) $\kappa = 5, n_2 = 1,000$, $\ell_{2,\infty}$ error

FIG. 5. *Estimation errors of* $\boldsymbol{U}$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *under the Poisson PCA model. Plot* (*a*) (*resp.*, (*b*)) *reports the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus* λ (*where* $n_1 = 100$, $n_2 = 1000$, $r = 3$).

tensor with entries $S_{1,1,1} = \kappa\sigma^\star$ and $S_{2,2,2} = S_{3,3,3} = \sigma^\star$. The noise tensor is generated in the following way: we first generate three random vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, where $\{\alpha_i\}$, $\{\beta_j\}$, $\{\gamma_k\}$ are independently drawn from $[0, 1]$. We then generate each $E_{i,j,k}$ independently from $\mathcal{N}(0, \omega^2\alpha_i^2\beta_j^2\gamma_k^2)$. The above four subspace estimation methods are applied to obtain initial subspace estimates, followed by 50 iterations of HOOI to refine the subspace estimators and construct the final tensor estimates. Figures 6 and 7 report the initial subspace estimation errors and the final subspace/tensor estimation errors, respectively. We can see from these plots that the Deflated-HeteroPCA algorithm produces faithful initial estimators in terms of both the ℓ_2 and $\ell_{2,\infty}$ errors, outperforming the other three methods. Moreover, compared with the other three methods, the Deflated-HeteroPCA algorithm serves as a more effective initialization scheme that can help one achieve more reliable subspace and tensor estimators.

7. Related works. This paper is closely related to the problem of matrix denoising, which aims to estimate either a low-rank matrix or its column subspace based on noisy observations and spans a diverse array of applications (Chen et al. (2021)). In addition to the examples of factor models and tensor estimation (Cai and Zhang (2018), Cai et al. (2021), Zhu, Wang and Samworth (2022), Richard and Montanari (2014), Zhang and Xia (2018), Cai et al. (2021)), it can also help us understand and solve several clustering problems (Rohe, Chatterjee and Yu (2011), Florescu and Perkins (2016), Cai et al. (2021), Chen, Liu and Ma (2022), Cai and Zhang (2018), Löffler, Zhang and Zhou (2021), Ndaoud (2022), Srivastava, Sarkar and Hanasusanto (2023), Han et al. (2022), Zhang and Zhou (2024)). When it comes to the task of estimating the whole matrix, a number of methods have been put forward and thoroughly studied in the literature, including but not limited to singular value hard thresholding (Gavish and Donoho (2014), Chatterjee (2015)), singular value soft thresholding (Cai, Candès and Shen (2010), Koltchinskii, Lounici and Tsybakov (2011), Donoho and Gavish (2014)) and singular value shrinkage (Nadakuditi (2014), Gavish and Donoho (2017)). Turning to the task of subspace estimation, the vanilla SVD-based approach (see (8)) has been commonly used and widely studied in the literature (Koltchinskii and Xia (2016), Cai and Zhang (2018), Bao, Ding and Wang (2021), Xia (2021), Chen et al. (2021)). How to perform uncertainty quantification for this approach has also been demonstrated in the previous work (see Chen et al. (2021)). In the scenario where the matrix dimensions are extremely unbalanced and the noise is heteroskedastic, however, such estimators can be highly suboptimal for subspace estimation. As already mentioned previously, the diagonal-deleted PCA



(a) $\kappa = 6$, ℓ_2 error

(b) $\kappa = 6$, $\ell_{2,\infty}$ error

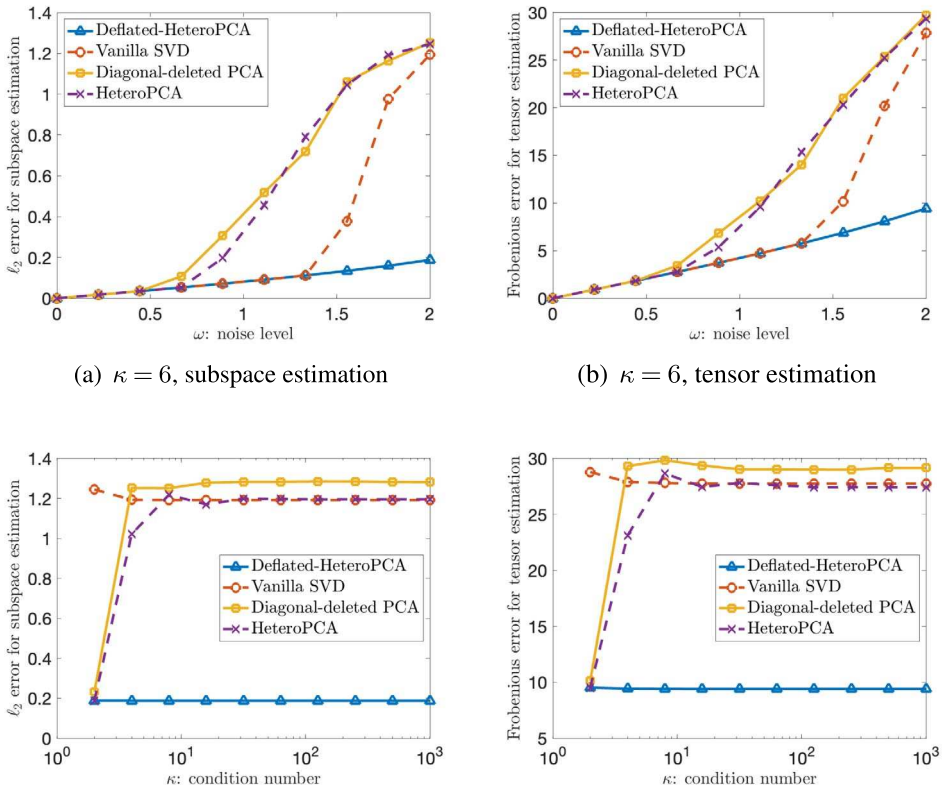
(c) $\omega = 2$, ℓ_2 error

(d) $\omega = 2$, $\ell_{2,\infty}$ error

FIG. 6. *Initial estimation errors of* $\widehat{\boldsymbol{U}}_1^0$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *under the tensor SVD model* (37). *Plot* (*a*) (*resp.*, (*b*)) *diplays the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the noise level* ω (*where* $n_1 = n_2 = n_3 = 50$, $r = 3$, $\kappa = 6$). *Plot* (*c*) (*resp.*, (*d*)) *shows the* ℓ_2 (*resp.*, $\ell_{2,\infty}$) *error versus the condition number* κ (*where* $n_1 = n_2 = n_3 = 50$, $r = 3$, $\omega = 2$).

and HeteroPCA algorithms have been proposed to improve the performance over the vanilla SVD approach (Cai et al. (2021), Zhang, Cai and Wu (2022), Agterberg, Lubberts and Priebe (2022), Yan, Chen and Fan (2024)). In fact, it has also been shown in Yan, Chen and Fan (2024) that the HeteroPCA admits a nonasymptotic distributional theory, which paves the way to construction of fine-grained confidence regions for this problem. Another family of effective algorithms—which can even accommodate the case when there is additional prior structure on the low-rank factors—is approximate message passing (Montanari and Venkataramanan (2021), Deshpande, Abbe and Montanari (2017), Feng et al. (2022), Li, Fan and Wei (2023), Li and Wei (2022), Montanari and Wu (2024)), for which the existing theory often requires more stringent assumptions on the noise components (e.g., i.i.d. Gaussian). It is also worth mentioning that how to accelerate optimization-based low-rank estimation algorithms in spite of ill conditioning has been an active research topic as well, which oftentimes involves proper preconditioning (Tong, Ma and Chi (2021), Xu et al. (2023)); the statistical guarantees therein, however, are still dependent on the condition number.

With regards to the factor model, one can easily find numerous works on this topic. The model (32) has been extensively studied under the names of spiked covariance models (Johnstone (2001), Paul (2007), Bai and Ding (2012), Wang and Fan (2017), Donoho, Gavish and Johnstone (2018), Perry et al. (2018), Bao et al. (2022)) and factor models (Lawley and Maxwell (1962), Bai and Li (2012), Fan, Liao and Wang (2016), Bai and Wang (2016)). Focusing on principal component estimation under heteroskedastic noise, Hong, Balzano and



(a) $\kappa = 6$, subspace estimation (b) $\kappa = 6$, tensor estimation

(c) $\omega = 2$, subspace estimation (d) $\omega = 2$, tensor estimation

FIG. 7. *Final estimation errors of* $\widehat{\boldsymbol{U}}_1$ *and* $\widehat{\boldsymbol{\mathcal{X}}}$ *for* Deflated-HeteroPCA, Diagonal-deleted PCA, HeteroPCA *and* Vanilla SVD *under the tensor SVD model* (37). *We report* (*a*) (*resp.*, (*b*)) ℓ_2 (*resp.*, *Frobenious*) *error of* $\widehat{\boldsymbol{U}}_1$ (*resp.* $\widehat{\boldsymbol{\mathcal{X}}}$) *versus noise level* ω (*where* $n_1 = n_2 = n_3 = 50$, $r = 3$, $\kappa = 6$); (*c*) (*resp.*, (*d*)) ℓ_2 (*resp.*, *Frobenious*) *error of* $\widehat{\boldsymbol{U}}_1$ (*resp.*, $\widehat{\boldsymbol{\mathcal{X}}}$) *versus condition number* κ (*where* $n_1 = n_2 = n_3 = 50$, $r = 3$, $\omega = 2$).

Fessler (2016, 2018), Hong et al. (2023) investigate the case where the noise components within each noise vector $\boldsymbol{\varepsilon}_j$ are i.i.d., and develop asymptotic analysis for PCA and a variant called Weighted PCA. Turning to nonasymptotic analysis, the theoretical performances of diagonal-deleted PCA (Cai et al. (2021)) and HeteroPCA have been investigated in (Cai et al. (2021), Zhang, Cai and Wu (2022), Yan, Chen and Fan (2024)). It is also worth noting that principal component estimation in the presence of missing data encounters additional challenges (Cai et al. (2021), Zhang, Cai and Wu (2022), Zhu, Wang and Samworth (2022), Pavez and Ortega (2021), Yan, Chen and Fan (2024)), which is beyond the scope of this work.

8. Discussion. This paper has studied subspace estimation from noisy low-rank matrices in the presence of unbalanced dimensionality and heteroskedastic noise. Recognizing a curse of ill-conditioning that appears in two cutting-edge algorithms, we have developed a new algorithm called Deflated-HeteroPCA to strengthen the state-of-the-art statistical performance in the face of a large condition number, without compromising the range of SNRs that can be accommodated. We have demonstrated that the proposed estimator enjoys nearly rate-optimal statistical guarantees (in terms of both the spectral-norm error and the more fine-grained $\ell_{2,\infty}$-based error), which are unaffected by the underlying condition number (regardless of how large it is). When applied to two concrete statistical models (i.e., factor models and tensor PCA), our theory has led to remarkable improvement over the prior art (particularly for the ill-conditioned scenarios).

Our work suggests several potential avenues for future investigation. For example, the signal-to-noise ratio conditions (15a) and (20a) in our theory remain suboptimal when it comes to their dependency on the rank r. How to tighten this rank dependency calls for a more refined analysis or a more powerful algorithm. Another direction worthy of future studies is the case with missing data (i.e., suppose we only have access to highly incomplete observations of the entries of the data matrix $\boldsymbol{Y}$ in (1)). It would be of great interest to extend our approach and develop a computationally efficient estimator that enjoys condition-number-free and rate-optimal estimation guarantees in the presence of missing data. Furthermore, note that the independent noise assumption plays an important role on our current theoretical analysis. Having said that, our method has potential to deal with more general correlated noise distributions (e.g., the one arising in network data). Our follow-up work Zhou and Chen (2023) applied a clustering method based on Deflated-HeteroPCA to the flight route network data, which demonstrates superior clustering performance compared to prior algorithms. We leave more extensive theoretical studies for the case with correlated data to future investigation.

Funding. This work is supported in part by the Alfred P. Sloan Research Fellowship and NSF Grants CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773.

SUPPLEMENTARY MATERIAL

Supplement: Proofs (DOI: 10.1214/24-AOS2456SUPPA; .pdf). All technical proofs of the results in this paper can be found in the Supplementary Material (Zhou and Chen (2025)).

Supplement to "Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA" (DOI: 10.1214/24-AOS2456SUPPB; .zip). Supplementary information.

REFERENCES

ABBE, E., FAN, J. and WANG, K. (2022). An ℓ_p theory of PCA and spectral clustering. *Ann. Statist.* **50** 2359–2385. MR4474494 https://doi.org/10.1214/22-aos2196

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

AGTERBERG, J., LUBBERTS, Z. and PRIEBE, C. E. (2022). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Trans. Inf. Theory* **68** 4618–4650. MR4449064 https://doi.org/10.1109/tit.2022.3159085

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966

BAI, J. and WANG, P. (2016). Econometric analysis of large factor models. *Ann. Rev. Econ.* **8** 53–80.

BAI, Z. and DING, X. (2012). Estimation of spiked eigenvalues in spiked models. *Random Matrices Theory Appl.* **1** 1150011, 21. MR2934717 https://doi.org/10.1142/S2010326311500110

BALZANO, L., CHI, Y. and LU, Y. M. (2018). Streaming PCA and subspace tracking: The missing data case. *Proc. IEEE Inst. Electr. Electron. Eng.* **106** 1293–1310. https://doi.org/10.1109/JPROC.2018.2847041

BAO, Z., DING, X., WANG, J. and WANG, K. (2022). Statistical inference for principal components of spiked covariance matrices. *Ann. Statist.* **50** 1144–1169. MR4404931 https://doi.org/10.1214/21-aos2143

BAO, Z., DING, X. and WANG, K. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *Ann. Statist.* **49** 370–392. MR4206682 https://doi.org/10.1214/20-AOS1960

CAI, C., LI, G., CHI, Y., POOR, H. V. and CHEN, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *Ann. Statist.* **49** 944–967. MR4255114 https://doi.org/10.1214/20-aos1986

CAI, C., LI, G., POOR, H. V. and CHEN, Y. (2022). Nonconvex low-rank tensor completion from noisy data. *Oper. Res.* **70** 1219–1237. MR4409613

CAI, J.-F., CANDÈS, E. J. and SHEN, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20** 1956–1982. MR2600248 https://doi.org/10.1137/080738970

CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. MR3766946 https://doi.org/10.1214/17-AOS1541

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

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

CHEN, P. H., CHEN, J., YESHURUN, Y., HASSON, U., HAXBY, J. V. and RAMADGE, P. J. (2015). A reduced-dimension fMRI shared response model. *Adv. Neural Inf. Process. Syst.* **2015** 460–468.

CHEN, S., LIU, S. and MA, Z. (2022). Global and individualized community detection in inhomogeneous multilayer networks. *Ann. Statist.* **50** 2664–2693. MR4500621 https://doi.org/10.1214/22-aos2202

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). *Spectral Methods for Data Science: A Statistical Perspective. Foundations and Trends® in Machine Learning* **14** 566–806.

CHEN, Y., CHI, Y., FAN, J., MA, C. and YAN, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* **30** 3098–3121. MR4167625 https://doi.org/10.1137/19M1290000

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2021). Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data. *Ann. Statist.* **49** 2948–2971. MR4338899 https://doi.org/10.1214/21-aos2066

DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21** 1324–1342. MR1780276 https://doi.org/10.1137/S0895479898346995

DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

DOBRIBAN, E. and OWEN, A. B. (2019). Deterministic parallel analysis: An improved method for selecting factors and principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 163–183. MR3904784

DONOHO, D. and GAVISH, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.* **42** 2413–2440. MR3269984 https://doi.org/10.1214/14-AOS1257

DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

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

FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC Press, Boca Raton.

FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767 https://doi.org/10.1214/15-AOS1364

FAN, J., WANG, K., ZHONG, Y. and ZHU, Z. (2021). Robust high-dimensional factor models with applications to statistical machine learning. *Statist. Sci.* **36** 303–327. MR4255196 https://doi.org/10.1214/20-sts785

FENG, O. Y., VENKATARAMANAN, R., RUSH, C., SAMWORTH, R. J. et al. (2022). *A Unifying Tutorial on Approximate Message Passing. Foundations and Trends® in Machine Learning* **15** 335–536.

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

GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory* **60** 5040–5053. MR3245370 https://doi.org/10.1109/TIT.2014.2323359

GAVISH, M. and DONOHO, D. L. (2017). Optimal shrinkage of singular values. *IEEE Trans. Inf. Theory* **63** 2137–2152. MR3626861 https://doi.org/10.1109/TIT.2017.2653801

HAN, R., LUO, Y., WANG, M. and ZHANG, A. R. (2022). Exact clustering in tensor block model: Statistical optimality and computational limit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1666–1698. MR4515554

HAN, R., WILLETT, R. and ZHANG, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *Ann. Statist.* **50** 1–29. MR4382094 https://doi.org/10.1214/21-AOS2061

HAN, Y. and ZHANG, C.-H. (2023). Tensor principal component analysis in high dimensional CP models. *IEEE Trans. Inf. Theory* **69** 1147–1167. MR4564648 https://doi.org/10.1109/tit.2022.3203972

HONG, D., BALZANO, L. and FESSLER, J. A. (2016). Towards a theoretical analysis of PCA for heteroscedastic data. In 2016 54*th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 496–503. IEEE.

HONG, D., BALZANO, L. and FESSLER, J. A. (2018). Asymptotic performance of PCA for high-dimensional heteroscedastic data. *J. Multivariate Anal.* **167** 435–452. MR3830656 https://doi.org/10.1016/j.jmva.2018.06.002

HONG, D., YANG, F., FESSLER, J. A. and BALZANO, L. (2023). Optimally weighted PCA for high-dimensional heteroscedastic data. *SIAM J. Math. Data Sci.* **5** 222–250. MR4567414 https://doi.org/10.1137/22M1470244

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

JOHNSTONE, I. M. and PAUL, D. (2018). PCA in high dimensions: An orientation. *Proc. IEEE Inst. Electr. Electron. Eng.* **106** 1277–1292. https://doi.org/10.1109/JPROC.2018.2846730

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

KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185 https://doi.org/10.2307/3318636

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

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, Berlin. MR3565274 https://doi.org/10.1007/978-3-319-40519-3_18

KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemom. Intell. Lab. Syst.* **94** 19–32.

KRITCHMAN, S. and NADLER, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.* **57** 3930–3941. MR2683143 https://doi.org/10.1109/TSP.2009.2022897

LAWLEY, D. N. and MAXWELL, A. E. (1962). Factor analysis as a statistical method. *J. R. Stat. Soc., Ser. D, Stat.* **12** 209–229.

LI, G., FAN, W. and WEI, Y. (2023). Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302930120, 7. MR4637851

LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. arXiv preprint. Available at arXiv:2208.03313.

LIU, L. T., DOBRIBAN, E. and SINGER, A. (2018). *e*PCA: High dimensional exponential family PCA. *Ann. Appl. Stat.* **12** 2121–2150. MR3875695 https://doi.org/10.1214/18-AOAS1146

LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.* **49** 2506–2530. MR4338373 https://doi.org/10.1214/20-aos2044

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

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

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

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

MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

MONTANARI, A. and WU, Y. (2024). Fundamental limits of low-rank matrix estimation with diverging aspect ratios. *Ann. Statist.* **52** 1460–1484. MR4804816 https://doi.org/10.1214/24-aos2400

NADAKUDITI, R. R. (2014). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inf. Theory* **60** 3002–3018. MR3200641 https://doi.org/10.1109/TIT.2014.2311661

NDAOUD, M. (2022). Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.* **50** 2096–2126. MR4474484 https://doi.org/10.1214/22-aos2178

NDAOUD, M., SIGALLA, S. and TSYBAKOV, A. B. (2022). Improved clustering algorithms for the bipartite stochastic block model. *IEEE Trans. Inf. Theory* **68** 1960–1975. MR4395508 https://doi.org/10.1109/tit.2021.3130683

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

PAVEZ, E. and ORTEGA, A. (2021). Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Trans. Inf. Theory* **67** 1201–1215. MR4232009 https://doi.org/10.1109/tit.2020.3039118

PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.* **46** 2416–2451. MR3845022 https://doi.org/10.1214/17-AOS1625

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

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

SRIVASTAVA, P. R., SARKAR, P. and HANASUSANTO, G. A. (2023). A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. *Oper. Res.* **71** 224–244. MR4560196 https://doi.org/10.1287/opre.2022.2317

TONG, T., MA, C. and CHI, Y. (2021). Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.* **22** Paper No. 150, 63. MR4318506

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457 https://doi.org/10.1214/16-AOS1487

XIA, D. (2021). Normal approximation and confidence region of singular subspaces. *Electron. J. Stat.* **15** 3798–3851. MR4298986 https://doi.org/10.1214/21-ejs1876

XIA, D., ZHANG, A. R. and ZHOU, Y. (2022). Inference for low-rank tensors—no need to debias. *Ann. Statist.* **50** 1220–1245. MR4404934 https://doi.org/10.1214/21-aos2146

XU, X., SHEN, Y., CHI, Y. and MA, C. (2023). The power of preconditioning in overparameterized low-rank matrix sensing. arXiv preprint. Available at arXiv:2302.01186.

YAN, Y., CHEN, Y. and FAN, J. (2024). Inference for heteroskedastic PCA with missing data. *Ann. Statist.* **52** 729–756. MR4744194 https://doi.org/10.1214/24-aos2366

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

ZHANG, A. R., CAI, T. T. and WU, Y. (2022). Heteroskedastic PCA: Algorithm, optimality, and applications. *Ann. Statist.* **50** 53–80. MR4382008 https://doi.org/10.1214/21-aos2074

ZHANG, A. Y. and ZHOU, H. Y. (2024). Leave-one-out singular subspace perturbation analysis for spectral clustering. *Ann. Statist.* **52** 2004–2033. MR4829478 https://doi.org/10.1214/24-aos2418

ZHAO, L. C., KRISHNAIAH, P. R. and BAI, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20** 1–25. MR0862239 https://doi.org/10.1016/0047-259X(86)90017-5

ZHOU, Y. and CHEN, Y. (2023). Heteroskedastic tensor clustering. arXiv preprint. Available at arXiv:2311.02306.

ZHOU, Y. and CHEN, Y. (2025). Supplement to "Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA." https://doi.org/10.1214/24-AOS2456SUPPA, https://doi.org/10.1214/24-AOS2456SUPPB

ZHOU, Y., ZHANG, A. R., ZHENG, L. and WANG, Y. (2022). Optimal high-order tensor SVD via tensor-train orthogonal iteration. *IEEE Trans. Inf. Theory* **68** 3991–4019. MR4433265 https://doi.org/10.1109/tit.2022.3152733

ZHU, Z., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 2000–2031. MR4515564