Submitted to the Annals of Statistics

INFERENCE FOR HETEROSKEDASTIC PCA WITH MISSING DATA

BY YULING YAN[1], YUXIN CHEN[2], JIANQING FAN[3],

[1]*Institute for Data, Systems, and Society, Massachusetts Institute of Technology, yulingy@mit.edu*

[2]*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, yuxinc@wharton.upenn.edu*

[3]*Department of Operations Research and Financial Engineering, Princeton University, jqfan@princeton.edu*

This paper studies how to construct confidence regions for principal component analysis (PCA) in high dimension, a problem that has been vastly under-explored. While computing measures of uncertainty for nonlinear/nonconvex estimators is in general difficult in high dimension, the challenge is further compounded by the prevalent presence of missing data and heteroskedastic noise. We propose a novel approach to performing valid inference on the principal subspace, on the basis of an estimator called HeteroPCA (Zhang, Cai and Wu, 2022). We develop non-asymptotic distributional guarantees for HeteroPCA, and demonstrate how these can be invoked to compute both confidence regions for the principal subspace and entrywise confidence intervals for the spiked covariance matrix. Our inference procedures are fully data-driven and adaptive to heteroskedastic random noise, without requiring prior knowledge about the noise levels.

1. Introduction. The applications of modern data science frequently ask for succinct representations of high-dimensional data. At the core of this pursuit lies principal component analysis (PCA), which serves as an effective means of dimension reduction and has been deployed across a broad range of domains (Jolliffe, 1986; Johnstone and Paul, 2018; Vaswani, Chi and Bouwmans, 2018; Fan et al., 2021). In reality, data collection could often be far from ideal — for instance, the acquired data might be subject to random contamination and contain incomplete observations — which inevitably affects the fidelity of PCA and calls for additional care when interpreting the results. To enable informative assessment of the influence of imperfect data acquisition, it would be desirable to accompany the PCA estimators in use with valid measures of uncertainty or "confidence".

1.1. *Problem formulation.* To allow for concrete and precise studies, the present paper concentrates on a tractable model that captures the effects of random heteroskedastic noise and missing data in PCA. In what follows, we start by formulating the problem, in the hope of facilitating more precise discussions.

Model. Imagine we are interested in n independent random vectors $\boldsymbol{x}_j = [x_{1,j}, \cdots, x_{d,j}]^\top \in \mathbb{R}^d$ drawn from the following distribution[1]

$$\boldsymbol{x}_j \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{S}^\star\right), \qquad 1 \le j \le n, \tag{1.1}$$

MSC2020 subject classifications: Primary 62H25; secondary 62E17.

Keywords and phrases: principal component analysis, confidence regions, missing data, uncertainty quantification, heteroskedastic data, subspace estimation.

[1]All results in this paper continue to hold if the sample vectors are generated such that $\boldsymbol{x}_j = \boldsymbol{U}^\star(\boldsymbol{\Lambda}^\star)^{1/2}\boldsymbol{f}_j$, where $\{\boldsymbol{f}_j\}_{j=1}^n$ are independent *sub-Gaussian* random vectors in $\mathbb{R}^r$ satisfying $\boldsymbol{E}[\boldsymbol{f}_j] = \boldsymbol{0}$, $\mathbb{E}[\boldsymbol{f}_j\boldsymbol{f}_j^\top] = \boldsymbol{I}_r$ and $\|\boldsymbol{f}_j\|_{\psi_2} = O(1)$. Here, $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm (Vershynin, 2018).

where the unknown covariance matrix $\boldsymbol{S}^\star \in \mathbb{R}^{d\times d}$ is assumed to be rank-r $(r < n)$ with eigen-decomposition

$$\boldsymbol{S}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top}. \tag{1.2}$$

Here, the orthonormal columns of $\boldsymbol{U}^\star \in \mathbb{R}^{d\times r}$ constitute the r leading eigenvectors of $\boldsymbol{S}^\star$, whereas $\boldsymbol{\Lambda}^\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose diagonal entries are composed of the non-zero eigenvalues of $\boldsymbol{S}^\star$. In other words, these vectors $\{\boldsymbol{x}_j\}_{1\le j\le n}$ are randomly drawn from a low-dimensional subspace when r is small. What we have available are partial and randomly corrupted observations of the entries of the above vectors. Specifically, suppose that we only get to observe

$$y_{l,j} = x_{l,j} + \eta_{l,j} \qquad \text{for all } (l,j) \in \Omega \tag{1.3}$$

over a subsampled index set $\Omega \subseteq [d]\times[n]$ (with $[n] \coloneqq \{1,\cdots,n\}$), where $\eta_{l,j}$ represents the noise that contaminates the observation in this location. Throughout this paper, we focus on the following random sampling and random noise models.

- *Random sampling*: each index (l,j) is contained in Ω independently with probability p;
- *Heteroskedastic random noise with unknown variance*: the noise components $\{\eta_{l,j}\}$ are independently generated sub-Gaussian random variables obeying

$$\mathbb{E}[\eta_{l,j}] = 0, \qquad \mathbb{E}[\eta_{l,j}^2] = \omega_l^{\star 2}, \qquad \text{and} \qquad \|\eta_{l,j}\|_{\psi_2} = O(\omega_l^\star),$$

  where $\{\omega_l^\star\}_{1\le l\le d}$ denote the standard deviations that are *a priori* unknown, and $\|\cdot\|_{\psi_2}$ stands for the sub-Gaussian norm of a random variable (Vershynin, 2018). The noise levels $\{\omega_l^\star\}_{1\le l\le d}$ are allowed to vary across locations, so as to model the so-called *heteroskedasticity* of noise.

This model can be viewed as a generalization of the spiked covariance model (Johnstone, 2001; Baik, Ben Arous and Péché, 2005; Paul, 2007; Donoho, Gavish and Johnstone, 2018; Nadler, 2008; Cai et al., 2021; Bao et al., 2022a) to account for missing data and heteroskedastic noise. With the observed data $\{y_{l,j} \mid (l,j)\in\Omega\}$ in hand, can we perform statistical inference on the orthonormal matrix $\boldsymbol{U}^\star$ — which embodies the ground-truth r-dimensional principal subspace underlying the vectors $\{\boldsymbol{x}_j\}_{1\le j\le n}$ — and make inference on the underlying covariance matrix $\boldsymbol{S}^\star$. Mathematically, the task can often be phrased as constructing valid confidence intervals/regions for both $\boldsymbol{U}^\star$ and $\boldsymbol{S}^\star$ based on the incomplete and corrupted observations $\{y_{l,j} \mid (l,j)\in\Omega\}$. Noteworthily, this model is frequently studied in econometrics and financial modeling under the name of factor models (Fan and Yao, 2017; Fan, Li and Liao, 2021; Fan et al., 2021; Bai and Wang, 2016; Gagliardini, Ossola and Scaillet, 2019), and is closely related to the noisy matrix completion problem where we also quantify uncertainty of missing entries (Candès and Recht, 2009; Candès and Plan, 2010; Keshavan, Montanari and Oh, 2010a; Chi, Lu and Chen, 2019).

Inadequacy of prior works. While methods for estimating principal subspace are certainly not in shortage (e.g., Lounici (2014); Zhang, Cai and Wu (2022); Cai and Zhang (2018); Balzano, Chi and Lu (2018); Zhu, Wang and Samworth (2022); Cai et al. (2021); Li et al. (2021)), methods for constructing confidence regions for principal subspace remain vastly under-explored. The fact that the estimators in use for PCA are typically nonlinear and nonconvex presents a substantial challenge in the development of a distributional theory, let alone uncertainty quantification. As some representative recent attempts, Xia (2021); Bao, Ding and Wang (2021) established normal approximations of the distance between the true subspace and its estimate for the matrix denoising setting, while Koltchinskii, Löffler and Nickl (2020) further established asymptotic normality of some debiased estimator for

linear functions of principal components. These distributional guarantees pave the way for the development of statistical inference procedures for PCA. However, it is noteworthy that these results required the noise components to either be i.i.d. Gaussian or at least exhibit matching moments (up to the 4th order), which fell short of accommodating heteroskedastic noise. The challenge is further compounded when statistical inference needs to be conducted in the face of missing data, a scenario that is beyond the reach of these prior works.

1.2. *Our contributions.* In light of the insufficiency of prior results, this paper takes a step towards developing data-driven inference and uncertainty quantification procedures for PCA, in the hope of accommodating both heteroskedastic noise and missing data. Our inference procedures are built on an estimator called HeteroPCA recently proposed by Zhang, Cai and Wu (2022), which is an iterative algorithm in nature and will be detailed in Section 2. The main contributions of this paper are summarized as follows.

- *Distributional theory for PCA and covariance estimation.* We derive, in a non-asymptotic manner, row-wise distributional characterizations of the principal subspace estimate returned by HeteroPCA (see Theorem 1), as well as entrywise distributional guarantees of the estimate for the covariance matrix estimate of $\{\boldsymbol{x}_l\}_{1\le l\le n}$ (see Theorem 3). These distributional characterizations take the form of tractable Gaussian approximations centered at the ground truth.
- *Fine-grained confidence regions and intervals.* Our distributional theory in turns allows for construction of row-wise confidence region for the subspace $\boldsymbol{U}^\star$ (see Algorithm 3 and Theorem 2) as well as entrywise confidence intervals for the matrix $\boldsymbol{S}^\star$ (see Algorithm 4 and Theorem 4). The proposed inference procedures are fully data-driven and do not require prior knowledge of the noise levels.

Along the way, we have significantly strengthened the estimation guarantees for HeteroPCA in the presence of missing data. It is noteworthy that all of our theory allows the observed data to be highly incomplete and covers heteroskedastic noise, which is previously unavailable.

1.3. *Paper organization.* The remainder of the paper is organized as follows. In Section 2, we introduce the estimation algorithms available in prior literature. Section 3 develops a suite of distributional theory for HeteroPCA and demonstrates how to use it to construct fine-grained confidence regions and confidence intervals for the unknowns; the detailed proofs of our theorems are deferred to the appendices. In Section 4, we carry out a series of numerical experiments to confirm the validity and applicability of our theoretical findings. Section 5 gives an overview of several related works. Section 6 takes a detour to analyze two intimately related problems, which will then be utilized to establish our main results. We conclude the paper with a discussion of future directions in Section 7. Most of the proof details are deferred to the appendices.

1.4. *Notation.* Before proceeding, we introduce several notation that will be useful throughout. We let $f(n)\lesssim g(n)$ or $f(n)=O(g(n))$ represent the condition that $|f(n)|\le Cg(n)$ for some constant $C>0$ when n is sufficiently large; we use $f(n)\gtrsim g(n)$ to denote $f(n)\ge C|g(n)|$ for some constant $C>0$ when n is sufficiently large; and we let $f(n)\asymp g(n)$ indicate that $f(n)\lesssim g(n)$ and $f(n)\gtrsim g(n)$ hold simultaneously. The notation $f(n)\gg g(n)$ (resp. $f(n)\ll g(n)$) means that there exists some sufficiently large (resp. small) constant $c_1>0$ (resp. $c_2>0$) such that $f(n)\ge c_1 g(n)$ (resp. $f(n)\le c_2 g(n)$). We also let $f(n)=o(g(n))$ indicate that $\lim_{n\to\infty} f(n)/g(n)=0$. For any real number $a,b\in\mathbb{R}$, we shall define $a\wedge b:=\min\{a,b\}$ and $a\vee b:=\max\{a,b\}$.

For any matrix $\boldsymbol{M}=[M_{i,j}]_{1\le i\le n_1,1\le j\le n_2}$, we let $\boldsymbol{M}_{i,\cdot}$ and $\boldsymbol{M}_{\cdot,j}$ stand for the i-th row and the j-th column of $\boldsymbol{M}$, respectively. We shall also let $\|\boldsymbol{M}\|$, $\|\boldsymbol{M}\|_{\mathrm{F}}$, $\|\boldsymbol{M}\|_{2,\infty}$ and

$\|\boldsymbol{M}\|_\infty$ denote the spectral norm, the Frobenius norm, the $\ell_{2,\infty}$ norm (i.e., $\|\boldsymbol{M}\|_{2,\infty} \coloneqq \max_i \|\boldsymbol{M}_{i,\cdot}\|_2$), and the entrywise ℓ_∞ norm ($\|\boldsymbol{M}\|_\infty \coloneqq \max_{i,j} |M_{i,j}|$) of $\boldsymbol{M}$, respectively. For any index set Ω, the notation $\mathcal{P}_\Omega(\boldsymbol{M})$ represents the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace of matrices supported on Ω, and define $\mathcal{P}_{\Omega^c}(\boldsymbol{M}) \coloneqq \boldsymbol{M} - \mathcal{P}_\Omega(\boldsymbol{M})$ as well. In addition, we denote by $\mathcal{P}_{\mathsf{diag}}(\boldsymbol{G})$ the Euclidean projection of a square matrix $\boldsymbol{G}$ onto the subspace of matrices that vanish outside the diagonal, and define $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{G}) \coloneqq \boldsymbol{G} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{G})$. For a non-singular matrix $\boldsymbol{H} \in \mathbb{R}^{k\times k}$ with SVD $\boldsymbol{U}_H \boldsymbol{\Sigma}_H \boldsymbol{V}_H^\top$, we denote by $\mathsf{sgn}(\boldsymbol{H})$ the following orthogonal matrix

$$\mathsf{sgn}(\boldsymbol{H}) \coloneqq \boldsymbol{U}_H \boldsymbol{V}_H^\top. \tag{1.4}$$

Finally, we denote by $\mathscr{C}^d$ the set of all convex sets in $\mathbb{R}^d$. For any Lebesgue measurable set $\mathcal{A} \subseteq \mathbb{R}^d$, we adopt the shorthand notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\{\mathcal{A}\} \coloneqq \mathbb{P}(\boldsymbol{z} \in \mathcal{A})$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Throughout this paper, we let $\Phi(\cdot)$ (resp. $\phi(\cdot)$) represent the cumulative distribution function (resp. probability distribution function) of the standard Gaussian distribution. We also denote by χ^2_k the chi-square distribution with k degrees of freedom.

2. Background: the estimation algorithm HeteroPCA. In order to conduct statistical inference for PCA, the first step lies in selecting an algorithm to estimate the principal subspace and the covariance matrix of interest, which we discuss in this section. Before continuing, we introduce several useful matrix notation as follows

$$\boldsymbol{X} \coloneqq [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}, \tag{2.1a}$$

$$\boldsymbol{Y} \coloneqq \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{N}) \in \mathbb{R}^{d\times n}, \tag{2.1b}$$

where $\mathcal{P}_\Omega$ has been defined in Section 1.4, and $\boldsymbol{N} \in \mathbb{R}^{d\times n}$ represents the noise matrix such that the (l,j)-th entry of $\boldsymbol{N}$ is given by $\eta_{l,j}$. In other words, $\boldsymbol{Y}$ encapsulates all the observed data $\{y_{l,j} \mid (l,j) \in \Omega\}$, with any entry outside Ω taken to be zero. If one has full access to the noiseless data matrix $\boldsymbol{X}$, then a natural strategy to estimate $\boldsymbol{U}^\star$ would be to return the top-r eigenspace of the sample covariance matrix $n^{-1}\boldsymbol{X}\boldsymbol{X}^\top$, or equivalently, the top-r left singular subspace of $\boldsymbol{X}$. In practice, however, one needs to extract information from the corrupted and incomplete data matrix $\boldsymbol{Y}$.

A vanilla SVD-based approach. Given that $p^{-1}\boldsymbol{Y} = p^{-1}\mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{N})$ is an unbiased estimate of $\boldsymbol{X}$ (conditional on $\boldsymbol{X}$), a natural idea that comes into mind is to resort to the top-r left singular subspace of $p^{-1}\boldsymbol{Y}$ when estimating $\boldsymbol{U}^\star$. This simple procedure is summarized in Algorithm 1.

Algorithm 1 A vanilla SVD-based approach.

Input: data matrix $\boldsymbol{Y}$ (cf. (2.1b)), sampling rate p, rank r.

Compute the truncated rank-r SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ of $p^{-1}\boldsymbol{Y}/\sqrt{n}$, where $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$.

Output: $\boldsymbol{U}$ as the subspace estimate, $\boldsymbol{\Sigma}$ as an estimate of $(\boldsymbol{\Lambda}^\star)^{1/2}$, and $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$ as the covariance matrix estimate of $\boldsymbol{x}$.

An improved iterative estimator: HeteroPCA. While Algorithm 1 returns reliable estimates of $\boldsymbol{U}^\star$ and $\boldsymbol{S}^\star$ in the regime of moderate-to-high signal-to-noise ratio (SNR), it might fail to be effective if either the missing rate $1-p$ or the noise levels are too large. To offer a high-level explanation, we find it helpful to compute the expectation of a properly rescaled sample covariance matrix:

$$\frac{1}{p^2}\mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top \,\middle|\, \boldsymbol{X}\right] = \boldsymbol{X}\boldsymbol{X}^\top + \left(\frac{1}{p} - 1\right)\mathcal{P}_{\mathsf{diag}}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) + \frac{n}{p}\mathsf{diag}\left\{\left[\omega_l^{\star 2}\right]_{1\le l\le d}\right\}, \tag{2.2}$$

where for any vector $\boldsymbol{z} = [z_l]_{1\le l\le d}$ we denote by $\mathsf{diag}(\boldsymbol{z}) \in \mathbb{R}^{d\times d}$ a diagonal matrix whose (l,l)-th entry equals z_l. Here, we rescale the sample covariance matrix by p^{-2} on the left-hand side, given that $p^{-1}\boldsymbol{Y}$ is an unbiased estimate for $\boldsymbol{X}$ and therefore we expect $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$ to be close to $\boldsymbol{X}\boldsymbol{X}^\top$. If the sampling rate p is overly small and/or if the noise is of large size but heteroskedastic, then the second and the third terms on the right-hand side of (2.2) might result in significant bias on the diagonal of the matrix $\mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top \,\middle|\, \boldsymbol{X}\right]$, thus hampering the statistical accuracy of the eigenspace of $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$ (or equivalently, the left singular space of $p^{-1}\boldsymbol{Y}$) when employed to estimate $\boldsymbol{U}^\star$. Viewed in this light, a more effective estimator would include procedures that properly handle the diagonal components of $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$.

To remedy this issue, several previous works (e.g., Florescu and Perkins (2016); Cai et al. (2021)) adopted a spectral method with diagonal deletion, which essentially discards any diagonal entry of $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$ before computing its top-r eigenspace. However, diagonal deletion comes at a price: while this operation mitigates the significant bias due to heteroskedasticity and missing data, it introduces another type of bias that might be non-negligible if the goal is to enable efficient fine-grained inference. To address this bias issue, Zhang, Cai and Wu (2022) proposed an iterative refinement scheme — termed HeteroPCA — that copes with the diagonal entries in a more refined manner. Informally, HeteroPCA starts by computing the rank-r eigenspace of the diagonal-deleted version of $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$, and then alternates between imputing the diagonal entries of $\boldsymbol{X}\boldsymbol{X}^\top$ and estimating the eigenspace of $p^{-2}\boldsymbol{Y}\boldsymbol{Y}^\top$ with the aid of the imputed diagonal. A precise description of this procedure is summarized in Algorithm 2; here, $\mathcal{P}_{\mathsf{off\text{-}diag}}$ and $\mathcal{P}_{\mathsf{diag}}$ have been defined in Section 1.4.

Algorithm 2 HeteroPCA (by Zhang, Cai and Wu (2022)).

Input: data matrix $\boldsymbol{Y}$ (cf. (2.1b)), sampling rate p, rank r, maximum number of iterations t_0.

Initialization: set $\boldsymbol{G}^0 = \frac{1}{np^2}\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$.

Updates: **for** $t = 0, 1, \ldots, t_0$ **do**

$$\left(\boldsymbol{U}^t, \boldsymbol{\Lambda}^t\right) = \mathsf{eigs}\left(\boldsymbol{G}^t, r\right);$$

$$\boldsymbol{G}^{t+1} = \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{G}^t\right) + \mathcal{P}_{\mathsf{diag}}\left(\boldsymbol{U}^t\boldsymbol{\Lambda}^t\boldsymbol{U}^{t\top}\right) = \frac{1}{np^2}\mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{Y}\boldsymbol{Y}^\top\right) + \mathcal{P}_{\mathsf{diag}}\left(\boldsymbol{U}^t\boldsymbol{\Lambda}^t\boldsymbol{U}^{t\top}\right).$$

Here, for any symmetric matrix $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ and $1 \le r \le d$, $\mathsf{eigs}(\boldsymbol{G}, r)$ returns $(\boldsymbol{U}, \boldsymbol{\Lambda})$, where $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ is the top-r eigen-decomposition of $\boldsymbol{G}$.

Output: $\boldsymbol{U} = \boldsymbol{U}^{t_0}$ as the subspace estimate, $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}^{t_0})^{1/2}$ as an estimate of $(\boldsymbol{\Lambda}^\star)^{1/2}$, and $\boldsymbol{S} = \boldsymbol{U}^{t_0}\boldsymbol{\Lambda}^{t_0}\boldsymbol{U}^{t_0\top}$ as the covariance matrix estimate.

3. Distributional theory and inference procedures. In this section, we augment the HeteroPCA estimator introduced in Section 2 by a suite of distributional theory, and demonstrate how to employ our distributional characterizations to perform inference on both the principal subspace represented by $\boldsymbol{U}^\star$ and the covariance matrix $\boldsymbol{S}^\star$.

3.1. *Key quantities and assumptions.* Before continuing, we introduce several additional notation and assumptions that play a key role in our theoretical development. Recall that the eigen-decomposition of the covariance matrix $\boldsymbol{S}^\star \in \mathbb{R}^{d\times d}$ (see (1.2)) is assumed to be $\boldsymbol{U}^\star\boldsymbol{\Lambda}^\star\boldsymbol{U}^{\star\top}$. We assume the diagonal matrix $\boldsymbol{\Lambda}^\star$ to be $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_r^\star\}$, where the diagonal entries are given by the non-zero eigenvalues of $\boldsymbol{S}^\star$ obeying

$$\lambda_1^\star \ge \cdots \ge \lambda_r^\star > 0.$$

The condition number of $\boldsymbol{S}^\star$ is denoted by

$$\kappa := \lambda_1^\star / \lambda_r^\star. \tag{3.1}$$

We also find it helpful to introduce the square root of $\boldsymbol{\Lambda}^\star$ as follows

$$\boldsymbol{\Sigma}^\star = \mathsf{diag}\{\sigma_1^\star, \ldots, \sigma_r^\star\} = (\boldsymbol{\Lambda}^\star)^{1/2}, \qquad \text{where } \sigma_i^\star = (\lambda_i^\star)^{1/2},\ 1 \le i \le r. \tag{3.2}$$

Furthermore, we introduce an incoherence parameter commonly employed in prior literature (Candès, 2014; Chi, Lu and Chen, 2019).

Definition 1 (Incoherence). *The rank-r matrix $\boldsymbol{S}^\star \in \mathbb{R}^{d\times d}$ defined in (1.2) is said to be μ-incoherent if the following condition holds:*

$$\|\boldsymbol{U}^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{d}}. \tag{3.3}$$

Here, we recall that $\|\boldsymbol{U}^\star\|_{2,\infty}$ denotes the largest ℓ_2 norm of all rows of the matrix $\boldsymbol{U}^\star$.

Remark 1. When μ is small (e.g., $\mu \asymp 1$), this condition essentially ensures that the energy of $\boldsymbol{U}^\star$ is nearly evenly dispersed across all of its rows. As a worthy note, the theory developed herein allows the incoherence parameter μ to grow with the problem dimension.

In light of a global rotational ambiguity issue (i.e., for any $r \times r$ rotation matrix $\boldsymbol{R}$, the matrices $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{U}\boldsymbol{R} \in \mathbb{R}^{d\times r}$ share the same column space), in general we can only hope to estimate $\boldsymbol{U}^\star$ up to global rotation (unless additional eigenvalue separation conditions are imposed). Consequently, our theoretical development focuses on characterizing the error distribution $\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star$ of an estimator $\boldsymbol{U}$ when accounting for a proper rotation matrix $\boldsymbol{R}$. In particular, we shall pay particular attention to a specific way of rotation as follows

$$\boldsymbol{U}\mathsf{sgn}\left(\boldsymbol{U}^\top \boldsymbol{U}^\star\right) - \boldsymbol{U}^\star,$$

where we recall that for any non-singular matrix $\boldsymbol{H} \in \mathbb{R}^{k\times k}$ with SVD $\boldsymbol{U}_H \boldsymbol{\Sigma}_H \boldsymbol{V}_H^\top$, the matrix $\mathsf{sgn}(\boldsymbol{H})$ is defined to be the rotation matrix $\boldsymbol{U}_H \boldsymbol{V}_H^\top$. This particular choice aligns $\boldsymbol{U}$ and $\boldsymbol{U}^\star$ in the following sense

$$\mathsf{sgn}\big(\boldsymbol{U}^\top \boldsymbol{U}^\star\big) = \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}},$$

where $\mathcal{O}^{r\times r}$ indicates the set of all $r \times r$ rotation matrices; see Ma et al. (2020, Appendix D.2.1). This is often referred to as Whaba's problem (Wahba, 1965) or the orthogonal Procrustes problem (Schönemann, 1966).

The last assumption is concerned with the noise levels, which are allowed to vary across different locations.

Assumption 1 (Noise levels). *The noise levels $\{\omega_i^\star\}_{1\le i\le d}$ obey*

$$\frac{\omega_{\max}^2}{\omega_{\min}^2} \le \kappa_\omega \qquad \textit{with} \quad \omega_{\max} := \max_{1\le i\le d} \omega_i^\star \quad \textit{and} \quad \omega_{\min} := \min_{1\le i\le d} \omega_i^\star. \tag{3.4}$$

3.2. *Inferential procedure and theory for HeteroPCA.* We are now positioned to investigate how to assess the uncertainty of the estimator HeteroPCA. For simplicity of presentation, we shall abuse some notation (e.g., $\boldsymbol{\Sigma}_{U,l}^\star$ and $v_{i,j}^\star$) whenever it is clear from the context.

3.2.1. *Distributional theory and inference for the principal subspace* $\boldsymbol{U}^\star$. In this subsection, we shall begin by establishing a distributional theory for the subspace estimate $\boldsymbol{U}$ returned by HeteroPCA (see Theorem 1), followed by a data-driven and provably valid method to construct fine-grained confidence regions for $\boldsymbol{U}^\star$ (see Algorithm 3 and Theorem 2). We shall also briefly discuss how our results improve upon prior estimation guarantees for HeteroPCA in the presence of missing data.

Distributional guarantees. As it turns out, the subspace estimate returned by Algorithm 2 is approximately unbiased and Gaussian under milder conditions, as posited in the following theorem. The general result beyond the case with $\kappa, \mu, r, \kappa_\omega \asymp 1$ is postponed to Theorem 11 in Appendix B.

THEOREM 1. *Assume that each column of the ground truth* $\boldsymbol{X}$ *(cf.* (2.1a)*) is independently generated from* $\mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$*, and that the sampling set* Ω *follows the random sampling model in Section 1.1. Suppose that* $p < 1-\delta$ *for some arbitrary constant* $0 < \delta < 1$ *or* $p = 1$*, and* $\kappa, \mu, r, \kappa_\omega \asymp 1$*. Assume that Assumption 1 holds and* $d \gtrsim \log^5 n$*,*

$$\frac{\omega_{\max}^2}{p\sigma_r^{\star 2}}\sqrt{\frac{d}{n}} \lesssim \frac{1}{\log^{7/2}(n+d)}, \qquad \frac{\omega_{\max}}{\sigma_r^\star}\sqrt{\frac{d}{np}} \lesssim \frac{1}{\log^3(n+d)}, \tag{3.5a}$$

$$ndp^2 \gtrsim \log^9(n+d), \qquad np \gtrsim \log^7(n+d). \tag{3.5b}$$

Suppose, in addition, that the number of iterations exceeds

$$t_0 \gtrsim \log\left[\left(\frac{\log^2(n+d)}{\sqrt{ndp}} + \frac{\omega_{\max}^2}{p\sigma_r^{\star 2}}\sqrt{\frac{d}{n}}\log(n+d) + \frac{\log(n+d)}{\sqrt{np}} + \frac{\omega_{\max}}{\sigma_r^\star}\sqrt{\frac{d\log(n+d)}{np}}\right)^{-1}\right]. \tag{3.6}$$

Let $\boldsymbol{R}$ *be the* $r \times r$ *rotation matrix* $\boldsymbol{R} = \mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star)$*. Then the estimate* $\boldsymbol{U}$ *returned by Algorithm 2 obeys: for all* $1 \le l \le d$*,*

$$\sup_{\mathcal{C}\in\mathscr{C}^r}\left|\mathbb{P}\left([\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star]_{l,\cdot} \in \mathcal{C}\right) - \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{U,l}^\star\right)\{\mathcal{C}\}\right| = o(1),$$

where $\mathscr{C}^r$ *represents the set of all convex sets in* $\mathbb{R}^r$*, and*

$$\begin{aligned}\boldsymbol{\Sigma}_{U,l}^\star \coloneqq &\left(\frac{1-p}{np}\left\|\boldsymbol{U}_{l,\cdot}^\star\boldsymbol{\Sigma}^\star\right\|_2^2 + \frac{\omega_l^{\star 2}}{np}\right)(\boldsymbol{\Sigma}^\star)^{-2} + \frac{2(1-p)}{np}\boldsymbol{U}_{l,\cdot}^{\star\top}\boldsymbol{U}_{l,\cdot}^\star \\ &+ (\boldsymbol{\Sigma}^\star)^{-2}\boldsymbol{U}^{\star\top}\mathsf{diag}\left\{\left[d_{l,i}^\star\right]_{1\le i\le d}\right\}\boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{-2}\end{aligned} \tag{3.7}$$

with

$$d_{l,i}^\star \coloneqq \frac{1}{np^2}\left[\omega_l^{\star 2} + (1-p)\left\|\boldsymbol{U}_{l,\cdot}^\star\boldsymbol{\Sigma}^\star\right\|_2^2\right]\left[\omega_i^{\star 2} + (1-p)\left\|\boldsymbol{U}_{i,\cdot}^\star\boldsymbol{\Sigma}^\star\right\|_2^2\right] + \frac{2(1-p)^2}{np^2}S_{l,i}^{\star 2}.$$

Theorem 1 asserts that each row of the estimate $\boldsymbol{U}$ returned by HeteroPCA is nearly unbiased and admits a nearly tight Gaussian approximation, whose covariance matrix can be determined via the closed-form expression (3.7). Given that $\boldsymbol{U}\boldsymbol{R}$ and $\boldsymbol{U}$ represent the same subspace, this theorem delivers a fine-grained row-wise distributional characterization for the estimator HeteroPCA.

Let us briefly mention the key error decomposition behind this theorem, which might help illuminate how Gaussian approximation emerges. Letting $\boldsymbol{E} \coloneqq n^{-1/2}(p^{-1}\boldsymbol{Y} - \boldsymbol{X})$ (which captures the randomness from both the noise and random subsampling), we can decompose

$$\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star = \underbrace{\left[\boldsymbol{E}\boldsymbol{X}^\top + \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{E}\boldsymbol{E}^\top\right)\right]\boldsymbol{U}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-2}}_{=:\,\boldsymbol{Z}\ \text{(first- and second-order approximation)}} + \underbrace{\left[\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star - \boldsymbol{Z}\right]}_{=:\,\boldsymbol{\Psi}\ \text{(residual term)}}. \tag{3.8}$$

Here, $\boldsymbol{Z}$ contains not only a linear mapping of $\boldsymbol{E}$ but also a certain quadratic mapping, the latter of which is crucial when coping with the regime $n \gg d$. As a consequence of the central limit theorem (which will be solidified in the proof), $\boldsymbol{Z}$ admits the following Gaussian approximation

$$\boldsymbol{Z}_{l,\cdot} \overset{\mathrm{d}}{\approx} \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}^\star_{U,l}\big), \qquad 1 \le l \le d. \tag{3.9}$$

At the same time, the ℓ_2 norm of the residual term $\boldsymbol{\Psi}_{l,\cdot}$ is well controlled and provably negligible compared to the corresponding component in $\boldsymbol{Z}_{l,\cdot}$, thus ascertaining the tightness of the advertised Gaussian approximation.

REMARK 2. The decomposition (3.8) also sheds light on why our current theory assumes a finite κ_ω (cf. (3.4)) when conducting statistical inference (which is unnecessary for the task of estimation). Consider, for example, a simple case when (i) there is no missing data ($p = 1$), and (ii) for some $1 \le l \le d$, one has $\omega_l^\star = 0$ (and hence $\kappa_\omega = \infty$) and $\|\boldsymbol{U}^\star_{l,\cdot}\|_2 > 0$. In this case, $\boldsymbol{Z}_{l,\cdot} = \boldsymbol{0}$ since $\boldsymbol{\Sigma}^\star_{U,l} = \boldsymbol{0}$, although $[\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star]_{l,\cdot}$ is in general non-zero. This implies that our Gaussian approximation — and the inference procedure developed based on this approximation — might fall short of efficacy when $\kappa_\omega = \infty$.

Construction of confidence regions for the principal subspace. With the above distributional theory in place, we are well-equipped to construct fine-grained confidence regions for $\boldsymbol{U}^\star$, provided that the covariance matrix $\boldsymbol{\Sigma}^\star_{U,l}$ can be estimated in a faithful manner. In Algorithm 3, we propose a procedure to estimate $\boldsymbol{\Sigma}^\star_{U,l}$, which in turn allows us to build confidence regions. As before, our estimator for $\boldsymbol{\Sigma}^\star_{U,l}$ can be viewed as a sort of "plug-in" method in accordance with the expression (3.7).

The following theorem confirms the validity of the proposed inference procedure when $\kappa, \mu, r, \kappa_\omega \asymp 1$. The more general case will be studied in Theorem 12 in Appendix B.

THEOREM 2. *Suppose that the conditions of Theorem 1 hold. Then there exists a $r \times r$ rotation matrix $\boldsymbol{R} = \mathsf{sgn}\left(\boldsymbol{U}^\top\boldsymbol{U}^\star\right)$ such that the confidence regions $\mathsf{CR}^{1-\alpha}_{U,l}$ $(1 \le l \le d)$ computed in Algorithm 3 obey*

$$\sup_{1\le l\le d}\left|\mathbb{P}\left(\boldsymbol{U}^\star_{l,\cdot}\boldsymbol{R}^\top \in \mathsf{CR}^{1-\alpha}_{U,l}\right) - (1-\alpha)\right| = o(1).$$

In words, Theorem 2 uncovers that: a valid ground-truth subspace representation is contained — in a row-wise reliable manner — within the confidence regions $\mathsf{CR}^{1-\alpha}_{U,l}$ $(1 \le l \le d)$ we construct. In the special case with $r = 1$, this result leads to valid entrywise confidence intervals for the principal component.

Interpretations and implications. We now take a moment to interpret the conditions required in Theorem 1 and Theorem 2, and discuss some appealing attributes of our methods. As before, the discussion below focuses on the scenario where $\mu, \kappa, r, \kappa_\omega \asymp 1$ for the sake of simplicity.

Algorithm 3 Confidence regions for $\boldsymbol{U}^{\star}_{l,\cdot}$ $(1 \le l \le d)$ based on HeteroPCA.

Input: output $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{S})$ of Algorithm 2, sampling rate p, coverage level $1-\alpha$.
Compute estimates of the noise levels $\{\omega_l^{\star}\}_{1\le l\le d}$ as follows

$$\omega_l^2 := \frac{\sum_{j=1}^n y_{l,j}^2 \mathbb{1}_{(l,j)\in\Omega}}{\sum_{j=1}^n \mathbb{1}_{(l,j)\in\Omega}} - S_{l,l} \qquad \text{for all} \quad 1 \le l \le d.$$

Compute an estimate of $\boldsymbol{\Sigma}^{\star}_{U,l}$ (cf. (3.7)) as follows:

$$\boldsymbol{\Sigma}_{U,l} := \left(\frac{1-p}{np}\left\|\boldsymbol{U}_{l,\cdot}\boldsymbol{\Sigma}\right\|_2^2 + \frac{\omega_l^2}{np}\right)\boldsymbol{\Sigma}^{-2} + \frac{2(1-p)}{np}\boldsymbol{U}_{l,\cdot}^{\top}\boldsymbol{U}_{l,\cdot} + (\boldsymbol{\Sigma})^{-2}\boldsymbol{U}^{\top}\mathsf{diag}\left\{\left[d_{l,i}\right]_{1\le i\le d}\right\}\boldsymbol{U}(\boldsymbol{\Sigma})^{-2},$$

where

$$d_{l,i} := \frac{1}{np^2}\left[\omega_l^2 + (1-p)\left\|\boldsymbol{U}_{l,\cdot}\boldsymbol{\Sigma}\right\|_2^2\right]\left[\omega_i^2 + (1-p)\left\|\boldsymbol{U}_{i,\cdot}\boldsymbol{\Sigma}\right\|_2^2\right] + \frac{2(1-p)^2}{np^2}S_{l,i}^2.$$

Compute the $(1-\alpha)$-quantile $\tau_{1-\alpha}$ of χ_r^2 and construct a Euclidean ball:

$$\mathcal{B}_{1-\alpha} := \left\{\boldsymbol{z}\in\mathbb{R}^r : \|\boldsymbol{z}\|_2^2 \le \tau_{1-\alpha}\right\}.$$

Output the $(1-\alpha)$-confidence region

$$\mathsf{CR}_{U,l}^{1-\alpha} := \boldsymbol{U}_{l,\cdot} + (\boldsymbol{\Sigma}_{U,l})^{1/2}\mathcal{B}_{1-\alpha} = \left\{\boldsymbol{U}_{l,\cdot} + (\boldsymbol{\Sigma}_{U,l})^{1/2}\boldsymbol{z} : \boldsymbol{z}\in\mathcal{B}_{1-\alpha}\right\}.$$

- *Missing data.* Both theorems accommodate the case when a large fraction of data are missing, namely, they cover the range

$$p \ge \widetilde{\Omega}\Big(\frac{1}{n\wedge\sqrt{nd}}\Big)$$

for both distributional characterizations and confidence region construction using HeteroPCA. In particular, if $n \gg d$, then the sampling rate p only needs to exceed

$$p \ge \widetilde{\Omega}\Big(\frac{1}{\sqrt{nd}}\Big);$$

this range can include some sampling rate much smaller than $1/d$ (with d the ambient dimension of each sample vector), and cannot be improved in general according to (Cai et al., 2021, Theorem 3.4).
- *Tolerable noise levels.* The noise conditon required in both Theorem 1 and Theorem 2 is given by

$$\omega_{\max}^2 \le \widetilde{O}\left(\Big(\frac{n}{d}\wedge\sqrt{\frac{n}{d}}\Big)p\sigma_r^{\star 2}\right).$$

Note that when $\kappa, \mu, r \asymp 1$, the variance obeys

$$\max_{(l,j)\in\Omega}\mathsf{var}\left(x_{l,j}\right) = \max_{l\in[d]}S_{l,l}^{\star} \asymp \max_{l\in[d]}\left\|\boldsymbol{U}_{l,\cdot}^{\star}\right\|_2^2\sigma_1^{\star 2} \asymp \frac{1}{d}\sigma_1^{\star 2}.$$

This implies that: when $p \ge \widetilde{\Omega}\big(1/(n\wedge\sqrt{nd})\big)$, our tolerable entrywise noise level $\omega_{\max}^2$ is allowed to be significantly (i.e., $\widetilde{\Omega}(np\wedge\sqrt{ndp^2})$ times) larger than the largest variance of $x_{l,j}$ for all $(l,j)\in\Omega$, thereby accommodating a wide range of noise levels.
- *Adaptivity to heteroskedasticity and unknown noise levels.* Our proposed inferential procedure is fully data-driven: it is automatically adaptive to unknown heteroskedastic noise, without requiring prior knowledge of the noise levels.

Comparison with prior estimation theory. While the main purpose of the current paper is to enable efficient statistical inference for the principal subspace, our theory (see Lemmas 18 and 19 in the appendix) also enables improved estimation guarantees compared to prior works.

- Recall that the estimation algorithm HeteroPCA was originally proposed and studied by Zhang, Cai and Wu (2022). Our results broaden the sample size range supported by their theory. More specifically, note that Zhang, Cai and Wu (2022, Theorem 6 and Remark 10) requires the sampling rate p to satisfy
$$ndp \gtrsim \max\left\{d^{1/3}n^{2/3}, d\right\} \mathsf{polylog}\,(n,d)$$
in order to guarantee consistent estimation, while our theoretical guarantees only require
$$ndp \gtrsim \max\left\{\sqrt{nd}, d\right\} \mathsf{polylog}\,(n,d)\,.$$
When $n \gg d$, the sample size requirement in Zhang, Cai and Wu (2022) is $(n/d)^{1/6}$ times more stringent than the one imposed in our theory.
- Let us discuss the advantage of HeteroPCA compared to the diagonal-deleted spectral method studied in Cai et al. (2021, Algorithms 1 and 3). Due to diagonal deletion, there is an additional bias term (see the last term $\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}r/d$ in Equation (4.16) in Cai et al. (2021)), which turns out to negatively affect our capability of performing inference. In contrast, HeteroPCA eliminates this bias term by means of successive refining, thus facilitating the subsequent inference stage.

3.2.2. *Distributional theory and inference for the covariance matrix $\boldsymbol{S}^\star$.* As it turns out, the above distributional theory for $\boldsymbol{U}^\star$ further hints at how to perform statistical inference for the covariance matrix $\boldsymbol{S}^\star$. In the sequel, we shall first develop an entrywise distributional theory for the estimate $\boldsymbol{S}$ returned by HeteroPCA (see Theorem 3), followed by a data-driven inference procedure to conduct entrywise confidence intervals for $\boldsymbol{S}^\star$ (see Algorithm 4 and Theorem 4).

Entrywise distributional guarantees. We now focus attention on characterizing the distribution of the (i,j)-th entry of $\boldsymbol{S}$ returned by Algorithm 2, which in turn suggests how to construct entrywise confidence intervals for $\boldsymbol{S}^\star$. Before proceeding, let us define a set of variance parameters $\{v^\star_{i,j}\}_{1\le i,j\le d}$ which, as we shall demonstrate momentarily, correspond to the (approximate) variance of the entries of $\boldsymbol{S}$.

- For any $1 \le i,j \le d$ obeying $i \ne j$, we define
$$\begin{aligned} v^\star_{i,j} \coloneqq{}& \frac{2-p}{np}S^\star_{i,i}S^\star_{j,j} + \frac{4-3p}{np}S^{\star 2}_{i,j} + \frac{1}{np}\left(\omega^{\star 2}_i S^\star_{j,j} + \omega^{\star 2}_j S^\star_{i,i}\right) \\ &+ \frac{1}{np^2}\sum_{k=1}^{d}\left\{\left[\omega^{\star 2}_i + (1-p)\,S^\star_{i,i}\right]\left[\omega^{\star 2}_k + (1-p)\,S^\star_{k,k}\right] + 2\,(1-p)^2\,S^{\star 2}_{i,k}\right\}\left(\boldsymbol{U}^\star_{k,\cdot}\boldsymbol{U}^{\star\top}_{j,\cdot}\right)^2 \end{aligned} \tag{3.10}$$
$$+ \frac{1}{np^2}\sum_{k=1}^{d}\left\{\left[\omega^{\star 2}_j + (1-p)\,S^\star_{j,j}\right]\left[\omega^{\star 2}_k + (1-p)\,S^\star_{k,k}\right] + 2\,(1-p)^2\,S^{\star 2}_{j,k}\right\}\left(\boldsymbol{U}^\star_{k,\cdot}\boldsymbol{U}^{\star\top}_{i,\cdot}\right)^2.$$
- For any $1 \le i \le d$, we set
$$v^\star_{i,i} \coloneqq \frac{12-9p}{np}S^{\star 2}_{i,i} + \frac{4}{np}\omega^{\star 2}_i S^\star_{i,i}$$

(3.11)
$$+\frac{4}{np^2}\sum_{k=1}^{d}\left\{\left[\omega_i^{\star 2}+(1-p)\,S_{i,i}^{\star}\right]\left[\omega_k^{\star 2}+(1-p)\,S_{k,k}^{\star}\right]+2\,(1-p)^2\,S_{i,k}^{\star 2}\right\}\left(\boldsymbol{U}_{k,\cdot}^{\star}\boldsymbol{U}_{i,\cdot}^{\star\top}\right)^2.$$

We are now positioned to present our distributional theory for the scenario where $\kappa, \mu, r, \kappa_\omega \asymp 1$, with the more general version deferred to Theorem 13 in Appendix B. Here and throughout, $S_{i,j}$ (resp. $S_{i,j}^{\star}$) represents the (i,j)-th entry of the matrix $\boldsymbol{S}$ (resp. $\boldsymbol{S}^{\star}$).

THEOREM 3. *Suppose that $p < 1-\delta$ for some arbitrary constant $0<\delta<1$ or $p=1$, and $\kappa, \mu, r, \kappa_\omega \asymp 1$. Consider any $1 \le i, j \le d$. Assume that $\boldsymbol{U}^{\star}$ is μ-incoherent and satisfies the following condition*

(3.12)
$$\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{U}_{j,\cdot}^{\star}\right\|_2 \gtrsim \left[\frac{\omega_{\max}}{\sigma_r^{\star}}\sqrt{\frac{d\log^5(n+d)}{np}}+\frac{\omega_{\max}^2}{p\sigma_r^{\star 2}}\sqrt{\frac{d\log^5(n+d)}{n}}+\sqrt{\frac{\log^7(n+d)}{ndp^2}}\right]\sqrt{\frac{1}{d}}.$$

In addition, suppose that Assumption 1 holds, and

$$d \gtrsim \log^5 n, \qquad np \gtrsim \log^7(n+d), \qquad ndp^2 \gtrsim \log^7(n+d),$$

$$\frac{\omega_{\max}}{\sigma_r^{\star}}\sqrt{\frac{d}{np}} \lesssim \frac{1}{\log^3(n+d)}, \qquad \frac{\omega_{\max}^2}{p\sigma_r^{\star 2}}\sqrt{\frac{d}{n}} \lesssim \frac{1}{\log^{7/2}(n+d)}.$$

Assume that the number of iterations satisfies (3.6). Then the matrix $\boldsymbol{S}$ computed by Algorithm 2 obeys

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_{i,j}-S_{i,j}^{\star}}{\sqrt{v_{i,j}^{\star}}}\le t\right)-\Phi(t)\right| = o(1),$$

where $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution.

In words, the above theorem indicates that under the conditions in Theorem 1, if the sum of the ℓ_2 norm of the rows $\boldsymbol{U}_{i,\cdot}^{\star}$ and $\boldsymbol{U}_{j,\cdot}^{\star}$ are not exceedingly small, then the estimation error $S_{i,j}-S_{i,j}^{\star}$ of HeteroPCA is approximately a zero-mean Gaussian with variance $v_{i,j}^{\star}$.

REMARK 3. When it comes to inference for $S_{i,j}^{\star}$, our theorems impose the following condition (cf. (3.12)):

$$\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{U}_{j,\cdot}^{\star}\right\|_2 \ge \widetilde{\Omega}\left(\frac{1}{\sqrt{ndp^2}}+\frac{\omega_{\max}}{\sigma_r^{\star}}\sqrt{\frac{d}{np}}+\frac{\omega_{\max}^2}{p\sigma_r^{\star 2}}\sqrt{\frac{d}{n}}\right)\cdot\sqrt{\frac{1}{d}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}.$$

Note that the typical ℓ_2 norm of a row of $\boldsymbol{U}^{\star}$ is $\|\boldsymbol{U}^{\star}\|_{\mathrm{F}}/\sqrt{d}$ when the energy is uniformly spread out across all rows. This means that under our sampling rate condition, our results allow $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2+\|\boldsymbol{U}_{j,\cdot}^{\star}\|_2$ to be much smaller than its typical size. As it turns out, a lower bound on $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2+\|\boldsymbol{U}_{j,\cdot}^{\star}\|_2$ might be necessary for $S_{i,j}-S_{i,j}^{\star}$ to be approximately Gaussian. Consider, for example, the case when $\|\boldsymbol{U}_{i,\cdot}\|_2=\|\boldsymbol{U}_{j,\cdot}\|_2=0$. It can be seen from our analysis that

$$S_{i,j}-S_{i,j}^{\star} \approx \boldsymbol{Z}_{i,\cdot}\boldsymbol{\Sigma}^{\star 2}\boldsymbol{Z}_{j,\cdot}^{\top}+A_{i,j}$$

where $\boldsymbol{Z}_{i,\cdot}$, $\boldsymbol{Z}_{j,\cdot}$ and $A_{i,j}$ are all (approximately) Gaussian. This means that $S_{i,j}-S_{i,j}^{\star}$ might not follow the (approximate) Gaussian distribution claimed in Theorem 3 if $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2+\|\boldsymbol{U}_{j,\cdot}^{\star}\|_2$ is too small.

Construction of entrywise confidence intervals. The distributional characterization in Theorem 3 enables valid construction of entrywise confidence intervals for $\boldsymbol{S}^\star$, as long as we can obtain reliable estimate of the variance $v_{i,j}^\star$. In what follows, we come up with an algorithm — as summarized in Algorithm 4 — that attempts to estimate $v_{i,j}^\star$ and build confidence intervals in a data-driven manner, as confirmed by the following theorem for the scenario with $\kappa, \mu, r, \kappa_\omega \asymp 1$. The more general result is postponed to Theorem 14 in Appendix B.

Algorithm 4 Confidence intervals for $S_{i,j}^\star$ $(1 \le i, j \le d)$ based on HeteroPCA.

Input: output $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{S})$ of Algorithm 2, sampling rate p, coverage level $1-\alpha$.

Compute estimates of the noise level $\omega_l^\star$ as follows

$$\omega_l^2 := \frac{\sum_{j=1}^n y_{l,j}^2 \mathbb{1}_{(l,j)\in\Omega}}{\sum_{j=1}^n \mathbb{1}_{(l,j)\in\Omega}} - S_{l,l}.$$

Compute an estimate of $v_{i,j}^\star$ (cf. (3.10) or (3.11))as follows: if $i \neq j$ then

$$\begin{aligned} v_{i,j} :=& \frac{2-p}{np} S_{i,i} S_{j,j} + \frac{4-3p}{np} S_{i,j}^2 + \frac{1}{np}\left(\omega_i^2 S_{j,j}^\star + \omega_j^2 S_{i,i}\right) \\ &+ \frac{1}{np^2}\sum_{k=1}^d \left\{\left[\omega_i^2 + (1-p)\, S_{i,i}\right]\left[\omega_k^2 + (1-p)\, S_{k,k}\right] + 2\,(1-p)^2\, S_{i,k}^2\right\}\left(\boldsymbol{U}_{k,\cdot}\boldsymbol{U}_{j,\cdot}^\top\right)^2 \\ &+ \frac{1}{np^2}\sum_{k=1}^d \left\{\left[\omega_j^2 + (1-p)\, S_{j,j}\right]\left[\omega_k^2 + (1-p)\, S_{k,k}\right] + 2\,(1-p)^2\, S_{j,k}^2\right\}\left(\boldsymbol{U}_{k,\cdot}\boldsymbol{U}_{i,\cdot}^\top\right)^2; \end{aligned}$$

If $i = j$ then

$$\begin{aligned} v_{i,i} :=& \frac{12-9p}{np} S_{i,i}^2 + \frac{4}{np}\omega_i^2 S_{i,i} \\ &+ \frac{4}{np^2}\sum_{k=1}^d \left\{\left[\omega_i^2 + (1-p)\, S_{i,i}\right]\left[\omega_k^2 + (1-p)\, S_{k,k}\right] + 2\,(1-p)^2\, S_{i,k}^2\right\}\left(\boldsymbol{U}_{k,\cdot}\boldsymbol{U}_{i,\cdot}^\top\right)^2. \end{aligned}$$

Output the $(1-\alpha)$-confidence interval

$$\mathsf{CI}_{i,j}^{1-\alpha} := \left[S_{i,j} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{i,j}}\right].$$

THEOREM 4. *Suppose that the conditions in Theorem 3 hold. Assume that* $ndp^2 \gtrsim \log^8(n+d)$. *Then the confidence interval computed in Algorithm 4 obeys*

$$\mathbb{P}\left(S_{i,j}^\star \in \mathsf{CI}_{i,j}^{1-\alpha}\right) = 1 - \alpha + o(1).$$

Compared with Cai et al. (2021, Corollary 2), we can see that when consistent estimation is possible — namely, under the sampling rate condition $p \ge \widetilde{\Omega}((n \wedge \sqrt{nd})^{-1})$ and the noise conditions $\omega_{\max} \le \widetilde{\Omega}((\sqrt{n/d} \wedge \sqrt[4]{n/d})\sqrt{p}\sigma_r^\star)$ — it is plausible to construct fine-grained confidence interval for $S_{i,j}^\star$, provided that the size of $\|\boldsymbol{U}_{i,\cdot}^\star\|_2 + \|\boldsymbol{U}_{j,\cdot}^\star\|_2$ is not exceedingly small.

REMARK 4. Before concluding this subsection, we note that the sampling rate p might be unknown *a priori* in practice. If this is the case, then one plausible strategy is to replace p

in Algorithms 2, 3 and 4 with the following empirical estimate:

$$\widehat{p} = \frac{\sum_{l=1}^{d}\sum_{j=1}^{n} \mathbb{1}\{(l,j)\in\Omega\}}{nd}.$$

In view of the standard concentration results $\widehat{p} = (1+o(1))p$, it is straightforward to verify that all of these inference procedure and the accompanying theory remain valid. We omit the details for the sake of brevity.

3.3. *A glimpse of key technical ingredients.* Let us take a moment to highlight several technical ingredients of the current theory, which might be applicable to other high-dimensional statistical problems beyond the analysis of HeteroPCA.

Second-order perturbation theory for principal subspace. At the core of our analysis lies a "second-order" perturbation theory tailored to general subspace estimation problems, to be presented in Section 6. More concretely, we establish a second-order expansion of the subspace perturbation error (see Theorem 5) that makes explicit the following two parts: (i) nearly tight first- and second-order terms, which can be expressed succinctly as linear and quadratic mappings of the perturbation matrix; (ii) the remaining higher-order terms that are provably negligible. Given that HeteroPCA is an iterative algorithm, developing such a refined perturbation theory for HeteroPCA becomes substantially more challenging than the vanilla SVD-based approach. Our refined perturbation theory allows us to tighten prior estimation theory (e.g., the Davis-Kahan $\sin\Theta$ Theorem (Davis and Kahan, 1970) or recent $\ell_{2,\infty}$-type perturbation bounds (Abbe et al., 2020; Cai et al., 2021; Chen et al., 2021a)), the latter of which focused mainly on providing orderwise estimation error bounds.

Fine-grained distributional characterizations for the principal subspace $\boldsymbol{U}^\star$. As alluded to previously, we establish the distributional characterization for the principal subspace (i.e., Theorem 1) based on a key error decomposition

$$\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star = \underbrace{\left[\boldsymbol{E}\boldsymbol{X}^\top + \mathcal{P}_{\text{off-diag}}\left(\boldsymbol{E}\boldsymbol{E}^\top\right)\right]\boldsymbol{U}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-2}}_{=:\boldsymbol{Z}\text{ (first- and second-order approximation)}} + \underbrace{[\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star - \boldsymbol{Z}]}_{=:\boldsymbol{\Psi}\text{ (residual term)}}, \tag{3.13}$$

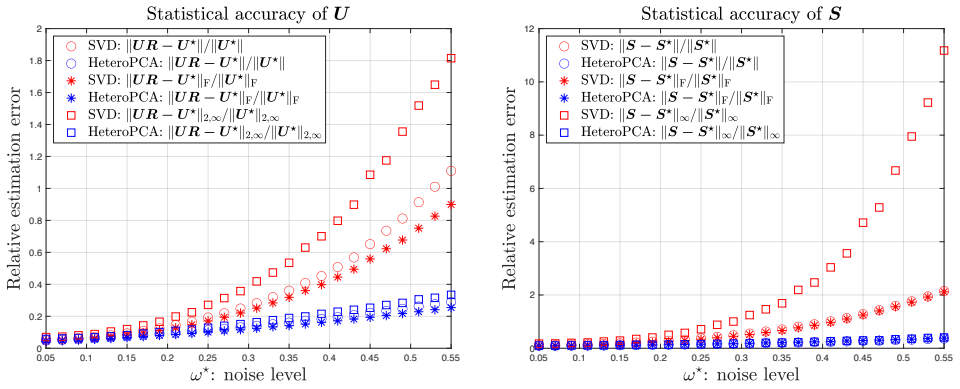
where $\boldsymbol{E} \coloneqq n^{-1/2}(p^{-1}\boldsymbol{Y} - \boldsymbol{X})$. For each $l \in [d]$, the multivariate Berry-Esseen Theorem reveals the approximate Gaussianity of $\boldsymbol{Z}_{l,\cdot}$, while at the same time, our second-order perturbation theory (cf. Theorem 5) ensures that $\boldsymbol{\Psi}_{l,\cdot}$ is stochastically dominated by $\boldsymbol{Z}_{l,\cdot}$. Additionally, rather than providing general $\ell_{2,\infty}$ bounds (as in the prior work Cai et al. (2021)), our proof relies crucially on more delicate row-dependent error control (so that the size of $\boldsymbol{\Psi}_{l,\cdot}$ is carefully bounded in accordance with the l-th row of $\boldsymbol{U}^\star$ and $\boldsymbol{N}$).

Entrywise distributional characterizations when estimating the covariance matrix $\boldsymbol{S}^\star$. Moving one step further, we derive the following key error decomposition w.r.t. the covariance matrix $\boldsymbol{S}^\star$:

$$\boldsymbol{S} - \boldsymbol{S}^\star = \underbrace{\boldsymbol{U}^\star\boldsymbol{\Sigma}^{\star 2}\boldsymbol{Z}^\top + \boldsymbol{Z}\boldsymbol{\Sigma}^{\star 2}\boldsymbol{U}^\star}_{=:\boldsymbol{W}} + \underbrace{n^{-1}\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{S}^\star}_{=:\boldsymbol{A}} + \underbrace{[\boldsymbol{S} - \boldsymbol{S}^\star - \boldsymbol{W} - \boldsymbol{A}]}_{=:\boldsymbol{\Phi}\text{ (residual term)}}, \tag{3.14}$$

where $\boldsymbol{Z}$ is defined in (3.13) and approximately Gaussian. Here, $\boldsymbol{W}$ serves as the main component as induced by the subspace estimation error, $\boldsymbol{A}$ indicates the discrepancy between the empirical covariance (using clean and fully observed data) and the true covariance, whereas $\boldsymbol{\Phi}$ is some higher-order term that is provably negligible in a strong entrywise sense. This in turn allows us to pin down tight entrywise distributional characterizations for $\boldsymbol{S} - \boldsymbol{S}^\star$.

4. Numerical experiments.



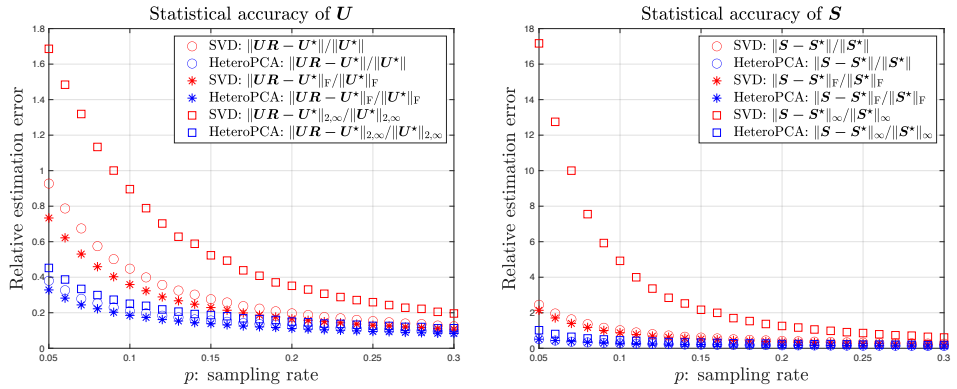
(a) (b)

FIG 1. *The relative estimation error of $\boldsymbol{U}$ and $\boldsymbol{S}$ returned by both SVD-based approach (cf. Algorithm 1) and HeteroPCA (cf. Algorithm 2) over different noise level $\omega^\star$. (a) Relative estimation errors of $\boldsymbol{UR}-\boldsymbol{U}^\star$ measured by $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{2,\infty}$ vs. the noise level $\omega^\star$; (b) Relative estimation errors of $\boldsymbol{S}-\boldsymbol{S}^\star$ measured by $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\infty}$ vs. the noise level $\omega^\star$. The results are reported over 200 independent trials for $r=3$ and $p=0.6$.*

Setup. This section conducts a series of numerical experiments to validate our distributional and inference theory developed in Section 3. Throughout this section, unless otherwise noted, we fix the dimension to be $d=100$ and the number of sample vectors to be $n=2000$, and we generate the covariance matrix as $\boldsymbol{S}^\star=\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}$ with $\boldsymbol{U}^\star\in\mathbb{R}^{n\times r}$ being a random orthonormal matrix following the Haar distribution over the Grassmann manifold $G_{d,r}$ (Vershynin, 2018, Section 5.2.6). In each Monte Carlo trial, the observed data are produced according to the model described in Section 1.1. For the purpose of introducing heteroskedasticity, we will introduce a parameter $\omega^\star$ that controls the noise level: in each independent trial, each noise level $\omega_l^\star$ $(1\le l\le d)$ is independently drawn from $\mathsf{Uniform}[0.1\omega^\star, 2\omega^\star]$; the random noise component $\eta_{l,j}$ is then drawn from $\mathcal{N}(0,\omega_l^{\star 2})$ independently for every $l\in[d]$ and $j\in[n]$.

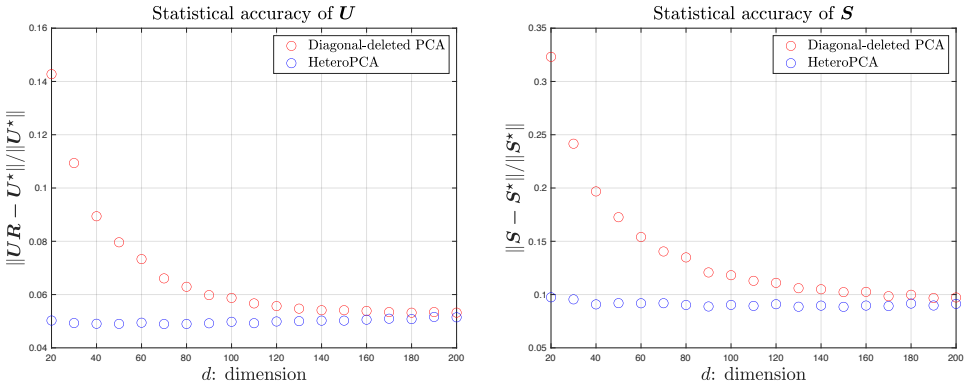
Superiority of HeteroPCA to the SVD-based approach in estimation. To begin with, we first compare the empirical estimation accuracy of the SVD approach (cf. Algorithm 1) and HeteroPCA (cf. Algorithm 2). Figure 1 displays the relative estimation errors — including the ones tailored to the principal subspace: $\|\boldsymbol{UR}-\boldsymbol{U}^\star\|/\|\boldsymbol{U}^\star\|$, $\|\boldsymbol{UR}-\boldsymbol{U}^\star\|_{\mathrm{F}}/\|\boldsymbol{U}^\star\|_{2,\infty}$, $\|\boldsymbol{UR}-\boldsymbol{U}^\star\|_{2,\infty}/\|\boldsymbol{U}^\star\|_{2,\infty}$, and the ones tailored to the covariance matrix: $\|\boldsymbol{S}-\boldsymbol{S}^\star\|/\|\boldsymbol{S}^\star\|$, $\|\boldsymbol{S}-\boldsymbol{S}^\star\|_{\mathrm{F}}/\|\boldsymbol{S}^\star\|_{\mathrm{F}}$, $\|\boldsymbol{S}-\boldsymbol{S}^\star\|_{\infty}/\|\boldsymbol{S}^\star\|_{\infty}$ — of both algorithms as the noise level $\omega^\star$ varies, with $r=3$ and $p=0.6$. Similarly, Figure 2 shows the relative numerical estimation errors of both algorithms vs. the sampling rate p, with $r=3$ and $\omega^\star=0.05$. As we shall see from both figures, HeteroPCA uniformly outperforms the SVD-based approach in all experiments, and is able to achieve appealing performance for a much wider range of noise levels and sampling rates.

Superiority of HeteroPCA to diagonal-deleted PCA in estimation. Let us also compare the empirical estimation accuracy of the diagonal-deleted spectral method (Cai et al., 2021) and HeteroPCA (cf. Algorithm 2). Recall from Section 3.2 that the main difference between the estimation error bounds of these two algorithms lies in an additional bias term due to the diagonal deletion operation (see the last term $\mu_{\mathsf{ce}}\kappa_{\mathsf{ce}}r/d$ in Equation (4.16) in Cai et al. (2021)). Figure 3 displays the relative estimation errors for estimating the principal subspace $\|\boldsymbol{UR}-\boldsymbol{U}^\star\|/\|\boldsymbol{U}^\star\|$ and for estimating the covariance matrix $\|\boldsymbol{S}-\boldsymbol{S}^\star\|/\|\boldsymbol{S}^\star\|$ as the dimension d varies, with $r=3$, $\omega^\star=0.05$ and $p=0.6$. As can be seen from the plots,



(a) (b)

FIG 2. *The relative estimation error of $\boldsymbol{U}$ and $\boldsymbol{S}$ returned by both SVD-based approach (cf. Algorithm 1) and HeteroPCA (cf. Algorithm 2) across different missing probability p. (a) Relative estimation errors of $\boldsymbol{UR}-\boldsymbol{U}^\star$ measured by $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{2,\infty}$ vs. the missing rate p; (b) Relative estimation errors of $\boldsymbol{S}-\boldsymbol{S}^\star$ measured by $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\infty}$ vs. the missing rate p. The results are reported over 200 independent trials for $r=3$ and $\omega^\star=0.05$.*

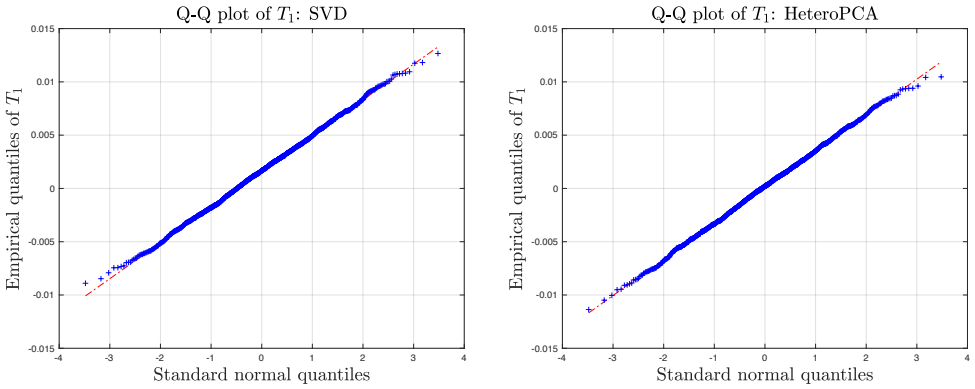


(a) (b)

FIG 3. *The relative estimation error of $\boldsymbol{U}$ and $\boldsymbol{S}$ returned by both diagonal-deleted spectral method (Cai et al., 2021) and HeteroPCA (cf. Algorithm 2). (a) Relative estimation error $\|\boldsymbol{UR}-\boldsymbol{U}^\star\|/\|\boldsymbol{U}^\star\|$ vs. dimension d; (b) Relative estimation error $\|\boldsymbol{S}-\boldsymbol{S}^\star\|/\|\boldsymbol{S}^\star\|$ vs. the dimension d. The results are reported over 200 independent trials for $r=3$, $\omega^\star=0.05$ and $p=0.6$.*

HeteroPCA uniformly outperforms the diagonal-deleted spectral method, especially when d is not too large. This numerical evidence corroborates the efficacy of the diagonal refinement scheme adopted in HeteroPCA.

Confidence regions for the principal subspace $\boldsymbol{U}^\star$. Next, we carry out a series of experiments to corroborate the practical validity of the confidence regions constructed using the SVD-based approach (Yan, Chen and Fan, 2021, Algorithm 3) and HeteroPCA (cf. Algorithm 3). To this end, we define $\widehat{\mathsf{Cov}}_U(i)$ to be the empirical probability that the constructed confidence interval $\mathsf{CR}_{U,i}^{0.95}$ covers $\boldsymbol{U}_{i,\cdot}^\star\mathsf{sgn}(\boldsymbol{U}^{\star\top}\boldsymbol{U})$ over 200 Monte Carlo trials, where $\boldsymbol{U}$ is the estimate returned by either algorithm. We also let $\mathsf{Mean}(\widehat{\mathsf{Cov}}_U)$ (resp. $\mathsf{std}(\widehat{\mathsf{Cov}}_U)$) be the empirical mean (resp. standard deviation) of $\widehat{\mathsf{Cov}}_U(i)$ over $i\in[d]$. Table 1 gathers $\mathsf{Mean}(\widehat{\mathsf{Cov}})$ and $\mathsf{std}(\widehat{\mathsf{Cov}})$ for $r=3$ and different choices of $(p,\omega^\star)$ for both algorithms. Encouragingly, the empirical coverage rates are all close to 95% for both



(a) (b)

FIG 4. *(a) Q-Q (quantile-quantile) plot of T_1 vs. the standard normal distribution for the SVD-based approach; (b) Q-Q (quantile-quantile) plot of T_1 vs. the standard normal distribution for* HeteroPCA*. The results are reported over* 2000 *independent trials for $r = 1$, $p = 0.6$ and $\omega^\star = 0.05$.*

methods when p is not too small and $\omega^\star$ is not too large. When p becomes smaller or $\omega^\star$ grows larger, HeteroPCA is still capable of performing valid statistical inference, while the SVD-based approach fails. This provides another empirical evidence on the advantage and broader applicability of HeteroPCA compared to the SVD-based approach. In addition, for the rank-1 case ($r = 1$), we define $T_i := [\boldsymbol{U} - \mathsf{sign}(\boldsymbol{U}^\top \boldsymbol{U}^\star)U^\star]_i / \sqrt{\Sigma_{U,i}}$. Figure 4 displays the Q-Q (quantile-quantile) plot of $T_1 := [U - \mathsf{sign}(\boldsymbol{U}^\top \boldsymbol{U}^\star)U^\star]_1 / \sqrt{\Sigma_{U,1}}$ vs. the standard Gaussian random variable over 2000 Monte Carlo simulations for both algorithms (when $p = 0.6$ and $\omega^\star = 0.05$); the near-Gaussian empirical distribution of T_1 also corroborates our distributional guarantees.

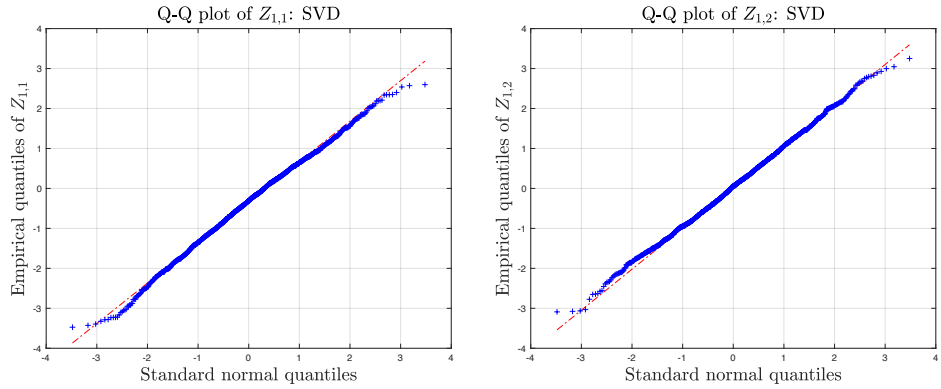
Entrywise confidence intervals for $\boldsymbol{S}^\star$. Finally, we provide numerical evidence that confirms the validity of the confidence interval constructed on the basis of the SVD-

TABLE 1
Empirical coverage rates of $\boldsymbol{U}^\star \mathsf{sgn}(\boldsymbol{U}^{\star\top} \boldsymbol{U})$ for different $(p, \omega^\star)$'s over 200 Monte Carlo trials

		The SVD-based Approach		HeteroPCA	
p	$\omega^\star$	$\mathsf{Mean}(\widehat{\mathsf{Cov}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}})$	$\mathsf{Mean}(\widehat{\mathsf{Cov}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}})$
0.6	0.05	0.9270	0.0292	0.9523	0.0157
0.6	0.1	0.8989	0.0521	0.9484	0.0154
0.4	0.05	0.8849	0.0501	0.9448	0.0184
0.4	0.1	0.8458	0.0853	0.9405	0.0182
0.2	0.05	0.7370	0.1196	0.9287	0.0204
0.2	0.1	0.6856	0.1569	0.9219	0.0204

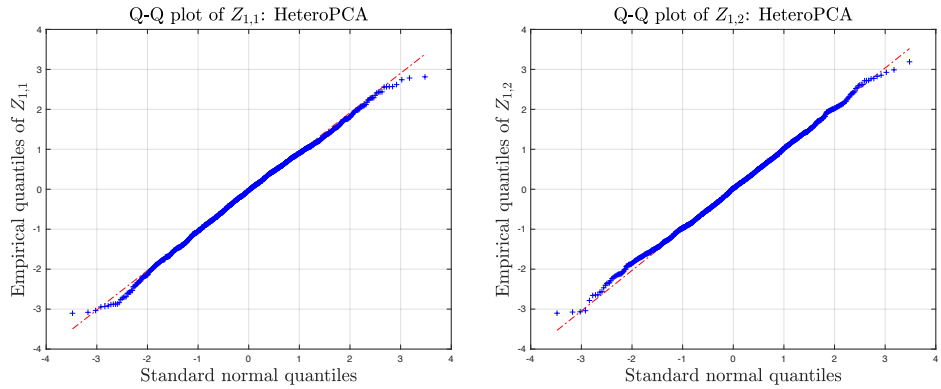
TABLE 2
Empirical coverage rates of $S^\star_{i,j}$ for different $(\omega^\star, p)$'s over 200 Monte Carlo trials

		The SVD-based Approach		HeteroPCA	
p	$\omega^\star$	$\mathsf{Mean}(\widehat{\mathsf{Cov}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}})$	$\mathsf{Mean}(\widehat{\mathsf{Cov}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}})$
0.6	0.05	0.9380	0.0244	0.9475	0.0153
0.6	0.1	0.9243	0.0425	0.9484	0.0151
0.4	0.05	0.9200	0.0509	0.9485	0.0156
0.4	0.1	0.9027	0.0713	0.9490	0.0153
0.2	0.05	0.8657	0.1031	0.9494	0.0164
0.2	0.1	0.8488	0.1186	0.9491	0.0162



(a) (b)

FIG 5. *(a) Q-Q (quantile-quantile) plot of $Z_{1,1}$ vs. the standard normal distribution for the SVD-based approach; (b) Q-Q (quantile-quantile) plot of $Z_{1,2}$ vs. a standard Gaussian distribution for the SVD-based approach. The results are reported over* 2000 *independent trials for $r = 3$, $p = 0.6$, $\omega^\star = 0.05$.*



(a) (b)

FIG 6. *(a) Q-Q (quantile-quantile) plot of $Z_{1,1}$ vs. the standard normal distribution for HeteroPCA; (b) Q-Q (quantile-quantile) plot of $Z_{1,2}$ vs. a standard Gaussian distribution for HeteroPCA. The results are reported over* 2000 *independent trials for $r = 3$, $p = 0.6$, $\omega^\star = 0.05$.*

based approach (Yan, Chen and Fan, 2021, Algorithm 4) and HeteroPCA (cf. Algorithm 4). Define $\widehat{\mathsf{Cov}}_S(i,j)$ to be the empirical probability that the 95% confidence interval $[S_{i,j} \pm 1.96\sqrt{v_{i,j}}]$ covers $S^\star_{i,j}$ over 200 Monte Carlo trials, where $S_{i,j}$ is the (i,j)-th entry of the estimate $\boldsymbol{S}$ returned by either algorithm. Let $\mathsf{Mean}(\widehat{\mathsf{Cov}}_S)$ (resp. $\mathsf{std}(\widehat{\mathsf{Cov}}_S)$) be the empirical mean (resp. standard deviation) of $\widehat{\mathsf{Cov}}_S(i,j)$ over all $i,j \in [d]$. Table 2 collects $\mathsf{Mean}(\widehat{\mathsf{Cov}})$ and $\mathsf{std}(\widehat{\mathsf{Cov}})$ for $r = 3$ and accounts for different choices of $(p, \omega^\star)$ for both algorithms. Similar to previous experiments, HeteroPCA uniformly outperforms the SVD-based approach, which again suggests that HeteroPCA is the method of choice. In addition, we define $Z_{i,j} := (S_{i,j} - S^\star_{i,j})/\sqrt{v_{i,j}}$. For both algorithms, Figure 5 and Figure 6 depict the Q-Q (quantile-quantile) plot of $Z_{1,1}$ and $Z_{1,2}$ vs. standard Gaussian distributions over 2000 Monte Carlo trials for the case with $r = 3$, $p = 0.6$ and $\omega^\star = 0.05$, which again confirm the practical validity of our distributional theory.

5. Other related works. Low-rank matrix denoising serves as a common model to study the effectiveness of spectral methods (Chen et al., 2021a), and has been the main subject

of many prior works including Cai and Zhang (2018); Chen, Cheng and Fan (2021); Ding (2020); Abbe et al. (2020); Bao, Ding and Wang (2021); Montanari, Ruan and Yan (2018); Lei (2019); Xia (2021); Cape, Tang and Priebe (2019); Agterberg, Lubberts and Priebe (2022), among others. Several recent works began to pursue a distributional theory for the eigenvector or singular vectors of the observed data matrix (Fan et al., 2020; Cheng, Wei and Chen, 2021; Bao, Ding and Wang, 2021; Xia, 2021). To name a few examples, Bao et al. (2022b) studied the limiting distribution of the inner product between an empirical singular vector and the corresponding ground truth, assuming that the associated spectral gap is sufficient large and that the noise components are homoskedastic; Xia (2021) established non-asymptotic Gaussian approximation for certain projection distance in the presence of i.i.d. Gaussian noise. Furthermore, the presence of missing data forms another source of technical challenges, leading to a problem often dubbed as noisy low-rank matrix completion (Candès and Plan, 2010; Negahban and Wainwright, 2012; Chen et al., 2020).

Spectral methods have been successfully applied to tackle noisy matrix completion (Keshavan, Montanari and Oh, 2010b; Sun and Luo, 2016; Chen and Wainwright, 2015; Zheng and Lafferty, 2016; Ma et al., 2020; Chen, Liu and Li, 2020; Cho, Kim and Rohe, 2017), which commonly serve as an effective initialization scheme for nonconvex optimization methods (Chi, Lu and Chen, 2019). While statistical inference for noisy matrix completion has been investigated recently (Chen et al., 2019a; Xia and Yuan, 2021; Chernozhukov et al., 2023), these prior works focused on performing inference based on optimization-based estimators. How to construct fine-grained confidence intervals based on spectral methods remains previously out of reach for noisy matrix completion. It is also noteworthy that the inferential procedures proposed in Chen et al. (2019a); Xia and Yuan (2021) (for noisy matrix completion) were developed for the regime where reliable estimation of the full low-rank matrix is feasible. This, however, falls short of covering the most challenging regime considered herein (where one might only be able to estimate the column subspace but not the row subspace). This crucial difference in the regimes of interest leads to substantial challenges unaddressed by these prior works.

Additionally, the recent work (Xia, 2019) tackled the confidence regions for spectral estimators tailored to the low-rank matrix regression problem, without accommodating the noisy matrix completion context. Most importantly, while the SVD-based vanilla spectral method often works well for the balanced case (such that the column dimension and the row dimension are on the same order), sub-optimality has been well recognized when estimating the column subspace of interest in the highly unbalanced case (so that the column dimension far exceeds the row dimension); this issue is also present when it comes to existing optimization-based methods like nuclear norm minimization. As a result, all prior schemes mentioned in this paragraph failed to tackle the highly balanced case in an statistically efficient manner.

Turning to PCA or subspace estimation, there has been an enormous literature dedicated to this topic; see Johnstone and Paul (2018); Balzano, Chi and Lu (2018) for an overview of prior development. Noteworthily, the need to handle the diagonals of the sample covariance matrix in the presence of heteroskedastic noise and/or missing data has been pointed out in many prior works, e.g., Loh and Wainwright (2012); Lounici (2014); Montanari and Sun (2018); Florescu and Perkins (2016); Cai et al. (2021). The iterative refinement scheme proposed by Zhang, Cai and Wu (2022) turns out to be among the most effective and adaptive schemes in handling the diagonals. Aimed at designing fine-grained estimators for the principal components, Koltchinskii, Löffler and Nickl (2020); Li et al. (2021) proposed statistically efficient de-biased estimators for linear functionals of principal components, and moreover, the estimator proposed in Koltchinskii, Löffler and Nickl (2020) has also been shown to exhibit asymptotic normality in the presence of i.i.d. Gaussian noise. Bloemendal

et al. (2016) also pinned down the asymptotic distributions of certain principal components under a spiked covariance model. However, these papers fell short of presenting valid and data-driven uncertainty quantification methods for the proposed estimators, and their results operates under the assumptions of homoskedastic noise without any missing data, a scenario that is remarkably more restricted than ours. Under the spiked covariance model, Bao et al. (2022b) studied the limiting distribution of the angle between the eigenvectors of the sample covariance matrix and any fixed vector, under the "balanced" scenario where the aspect ratio n/d is a constant. In addition, recent years have witnessed much activity in high-dimensional PCA in the face of missing data (Zhang, Cai and Wu, 2022; Zhu, Wang and Samworth, 2022; Cai et al., 2021; Pavez and Ortega, 2020); these works, however, focused primarily on developing estimation guarantees, which did not provide either distributional guarantees for the estimators or concrete procedures that allow for confidence region construction. Additionally, the HeteroPCA algorithm has been further extended by two follow-up works Zhou and Chen (2023a,b) to accommodate the scenario with large condition numbers as well as tensor clustering in the presence of heteroskedastic noise

From a technical viewpoint, it is worth mentioning that the ℓ_∞ and $\ell_{2,\infty}$ perturbation theory has been an active research direction in recent years (Fan, Wang and Zhong, 2018; Cape, Tang and Priebe, 2019; Chen, Cheng and Fan, 2021; Eldridge, Belkin and Wang, 2018; Agterberg, Lubberts and Priebe, 2022; Xie, 2021). Among multiple existing technical frameworks, the leave-one-out analysis idea — which has been applied to a variety of statistical estimation problems (El Karoui et al., 2013; El Karoui, 2015; Zhong and Boumal, 2018; Chen et al., 2019b, 2021b; Cai et al., 2022; Ling, 2022; Cai, Poor and Chen, 2022; Chen, Gao and Zhang, 2022) — provides a powerful and flexible framework that enables ℓ_∞ and $\ell_{2,\infty}$ statistical guarantees for spectral methods (Chen et al., 2019c; Abbe et al., 2020; Cai et al., 2021); see (Chen et al., 2021a, Chapter 4) for an accessible introduction of this powerful framework. Our analysis for the HeteroPCA approach is influenced by the one in Cai et al. (2021). Note, however, that Cai et al. (2021) didn't come with any distributional guarantees for spectral methods, which we seek to accomplish in this paper.

It is important to note that although the current version of this paper focuses primarily on the HeteroPCA method, a preliminary version available on arXiv (Yan, Chen and Fan, 2021) includes a discussion on distributional theory and inferential procedures for PCA using the SVD-based approach (cf. Algorithm 1). This content was subsequently omitted during the revision phase based on editiorial suggestions. Interested readers are referred to Yan, Chen and Fan (2021) for a set of inferential results developed for the SVD-based approach, in parallel to Theorems 1 to 4 in this paper.

Finally, we note in passing that constructing confidence intervals for sparse regression (based on, say, the Lasso estimator or other sparsity-promoting estimator), has attracted a flurry of research activity in the past few years (Celentano, Montanari and Wei, 2023; Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Cai and Guo, 2017; Ning and Liu, 2017; Ren et al., 2015). The methods derived therein, however, are not directly applicable to perform statistical inference for PCA and/or other low-rank models.

6. A detour: subspace estimation. We now take a detour to look at an intimately related problem, which we shall refer to as *subspace estimation* and will play a crucial role in understanding the HeteroPCA approach. We will set out to develop a fine-grained statistical theory for HeteroPCA when applied to this subspace estimation setting. The resulting theory will be invoked in Appendix D to analyze the PCA context.

6.1. *Model and algorithm.*

Model and assumptions. Suppose that we are interested in a rank-r matrix $\boldsymbol{M}^\natural \in \mathbb{R}^{n_1 \times n_2}$, whose SVD is given by

$$\boldsymbol{M}^\natural = \sum_{i=1}^{r} \sigma_i^\natural \boldsymbol{u}_i^\natural \boldsymbol{v}_i^{\natural\top} = \boldsymbol{U}^\natural \boldsymbol{\Sigma}^\natural \boldsymbol{V}^{\natural\top} \in \mathbb{R}^{n_1 \times n_2} \tag{6.1}$$

Here, $\boldsymbol{U}^\natural = [\boldsymbol{u}_1^\natural, \ldots, \boldsymbol{u}_r^\natural]$ (resp. $\boldsymbol{V}^\natural = [\boldsymbol{v}_1^\natural, \ldots, \boldsymbol{v}_r^\natural]$) consists of orthonormal columns that correspond to the left (resp. right) singular vectors of $\boldsymbol{M}^\natural$, and $\boldsymbol{\Sigma}^\natural = \mathsf{diag}\{\sigma_1^\natural, \ldots, \sigma_r^\natural\}$ is a diagonal matrix consisting of the singular values of $\boldsymbol{M}^\natural$. Without loss of generality, we assume that

$$n = \max\{n_1, n_2\}.$$

It is assumed that the singular values are sorted (in magnitude) in descending order, namely,

$$\sigma_1^\natural \geq \cdots \geq \sigma_r^\natural \geq 0, \tag{6.2}$$

with the condition number denoted by

$$\kappa^\natural := \sigma_1^\natural / \sigma_r^\natural. \tag{6.3}$$

What we have observed is a noisy copy of $\boldsymbol{M}^\natural$, namely,

$$\boldsymbol{M} = \boldsymbol{M}^\natural + \boldsymbol{E}, \tag{6.4}$$

where $\boldsymbol{E} = [E_{i,j}]_{1 \leq i,j \leq n}$ stands for a noise matrix. We focus on *estimating the column subspace* represented by $\boldsymbol{U}^\natural$ and the singular values encapsulated in $\boldsymbol{\Sigma}^\natural$, but not the row space $\boldsymbol{V}^\natural$. An important special scenario one should bear in mind is the highly unbalanced case where the column dimension n_2 far exceeds the row dimension n_1; in this case, it is common to encounter situations where reliable estimation of $\boldsymbol{M}^\natural$ and $\boldsymbol{V}^\natural$ is infeasible but that of $\boldsymbol{U}^\natural$ shows promise. For this reason, we refer to this setting as *subspace estimation* in order to differentiate it from matrix denoising, emphasizing that we are only interested in column subspace estimation.

With the new aim in mind, we shall modify our incoherence and noise assumptions accordingly. Here, we abuse the notation with the understanding that the following set of assumptions will be used only when analyzing the approach based on HeteroPCA. We shall also denote $n := \max\{n_1, n_2\}$.

ASSUMPTION 2 (Incoherence). *The rank-r matrix $\boldsymbol{M}^\natural \in \mathbb{R}^{n_1 \times n_2}$ defined in (6.1) is said to be $\mu^\natural$-incoherent if the following holds:*

$$\left\|\boldsymbol{U}^\natural\right\|_{2,\infty} \leq \sqrt{\frac{\mu^\natural r}{n_1}}, \quad \left\|\boldsymbol{V}^\natural\right\|_{2,\infty} \leq \sqrt{\frac{\mu^\natural r}{n_2}}, \quad \text{and} \quad \left\|\boldsymbol{M}^\natural\right\|_\infty \leq \sqrt{\frac{\mu^\natural}{n_1 n_2}} \left\|\boldsymbol{M}^\natural\right\|_{\mathrm{F}}.$$

ASSUMPTION 3 (Heteroskedastic random noise). *Assume that the $E_{i,j}$'s are independently generated, and suppose that there exist non-negative quantities $\{\sigma_i\}_{i=1}^{n_1}$, $\{B_i\}_{i=1}^{n_1}$, σ and B obeying*

$$\forall (i,j) \in [n_1] \times [n_2]: \qquad \mathbb{E}\left[E_{i,j}\right] = 0, \qquad \mathsf{var}\left(E_{i,j}^2\right) = \sigma_{i,j}^2 \leq \sigma_i^2 \leq \sigma^2, \qquad |E_{i,j}| \leq B_i \leq B,$$

where for all $i \in [n_1]$

$$B_i \lesssim \frac{\sigma_i \min\left\{\sqrt{n_2}, \sqrt[4]{n_1 n_2}\right\}}{\sqrt{\log n}}, \qquad \text{and} \qquad B \lesssim \frac{\sigma \min\left\{\sqrt{n_2}, \sqrt[4]{n_1 n_2}\right\}}{\sqrt{\log n}}. \tag{6.5}$$

Algorithm: HeteroPCA for subspace estimation. The paradigm HeteroPCA can naturally be applied to tackle the above subspace estimation task. Let us introduce the ground-truth gram matrix as follows

$$\boldsymbol{G}^{\natural} := \boldsymbol{M}^{\natural}\boldsymbol{M}^{\natural\top}. \tag{6.6}$$

Given that $\boldsymbol{M} = \boldsymbol{M}^{\natural} + \boldsymbol{E}$ is an unbiased estimate of $\boldsymbol{M}^{\natural}$, one might naturally attempt to estimate the column space of $\boldsymbol{M}$ by looking at the eigenspace of the sample Gram matrix $\boldsymbol{M}\boldsymbol{M}^{\top}$. It can be easily seen that

$$\boldsymbol{E}\left[\boldsymbol{M}\boldsymbol{M}^{\top}\right] = \boldsymbol{M}^{\natural}\boldsymbol{M}^{\natural\top} + \mathsf{diag}\left\{\left[\sum_{j=1}^{n_2}\sigma_{i,j}^2\right]_{1\le i\le n_1}\right\}, \tag{6.7}$$

where the diagonal term on the right-hand side of (6.7) might incur significant bias in the most challenging regime. The HeteroPCA algorithm seeks to handle the diagonal part in an iterative manner, alternating between imputing the values of the diagonal entries and eigen-decomposition of $\boldsymbol{M}\boldsymbol{M}^{\top}$ with the diagonal replaced by the imputed values. The procedure is summarized in Algorithm 5.

Algorithm 5 HeteroPCA for general subspace estimation (HeteroPCA).

Initialization: set $\boldsymbol{G}^0 = \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{M}\boldsymbol{M}^{\top}\right)$.

Updates: **for** $t = 0, 1, \ldots, t_0$ **do**

$$\left(\boldsymbol{U}^t, \boldsymbol{\Lambda}^t\right) = \mathsf{eigs}\left(\boldsymbol{G}^t, r\right); \tag{6.8a}$$

$$\boldsymbol{G}^{t+1} = \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{M}\boldsymbol{M}^{\top}\right) + \mathcal{P}_{\mathsf{diag}}\left(\boldsymbol{U}^t\boldsymbol{\Lambda}^t\boldsymbol{U}^{t\top}\right). \tag{6.8b}$$

Here, $\mathsf{eigs}(\boldsymbol{G}, r)$ returns $(\boldsymbol{U}, \boldsymbol{\Lambda})$ where $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$ is the top-r eigen-decomposition of $\boldsymbol{G}$.

Output: $\boldsymbol{U} = \boldsymbol{U}^{t_0}$, $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{t_0}$, $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}^{t_0})^{1/2}$, $\boldsymbol{S} = \boldsymbol{U}^{t_0}\boldsymbol{\Lambda}^{t_0}\boldsymbol{U}^{t_0\top}$.

6.2. *Fine-grained statistical guarantees for HeteroPCA.* We now move on to present our theoretical guarantees for Algorithm 5. In order to account for the potential global rotational ambiguity, we introduce the following rotation matrix as before

$$\boldsymbol{R}_{\boldsymbol{U}} := \arg\min_{\boldsymbol{O}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}\boldsymbol{O} - \boldsymbol{U}^{\natural}\right\|_{\mathrm{F}}^2, \tag{6.9}$$

where we recall that $\mathcal{O}^{r\times r}$ represents the set of $r\times r$ orthonormal matrices. It is also helpful to define the following quantities: for all $m \in [n_1]$,

$$\zeta_{\mathsf{op}} := \sigma^2\sqrt{n_1 n_2}\log n + \sigma\sigma_1^{\natural}\sqrt{n_1\log n}, \tag{6.10a}$$

$$\zeta_{\mathsf{op},m} := \sigma\sigma_m\sqrt{n_1 n_2}\log n + \sigma_m\sigma_1^{\natural}\sqrt{n_1\log n}, \tag{6.10b}$$

Our result is as follows, with the proof postponed to Appendix C.

THEOREM 5. *Suppose that Assumptions 2-3 hold. Assume that*

$$n_1 \gtrsim \kappa^{\natural 4}\mu^{\natural}r + \mu^{\natural 2}r\log^2 n, \qquad n_2 \gtrsim r\log^4 n, \qquad \textit{and} \qquad \zeta_{\mathsf{op}} \ll \frac{\sigma_r^{\natural 2}}{\kappa^{\natural 2}}, \tag{6.11}$$

and that the algorithm is run for $t_0 \geq \log\left(\frac{\sigma_1^{\natural 2}}{\zeta_{\mathsf{op}}}\right)$ *iterations. With probability exceeding* $1 - O(n^{-10})$*, there exist two matrices* $\boldsymbol{Z}$ *and* $\boldsymbol{\Psi}$ *such that the estimates returned by* HeteroPCA *obey*

$$\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^{\natural} = \boldsymbol{Z} + \boldsymbol{\Psi}, \tag{6.12}$$

where

$$\boldsymbol{Z} \coloneqq \boldsymbol{E}\boldsymbol{V}^{\natural}\left(\boldsymbol{\Sigma}^{\natural}\right)^{-1} + \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{E}\boldsymbol{E}^{\top}\right)\boldsymbol{U}^{\natural}\left(\boldsymbol{\Sigma}^{\natural}\right)^{-2}, \tag{6.13a}$$

$$\|\boldsymbol{\Psi}\|_{2,\infty} \lesssim \kappa^{\natural 2}\frac{\mu^{\natural} r}{n_1}\frac{\zeta_{\mathsf{op}}}{\sigma_r^{\natural 2}} + \kappa^{\natural 2}\frac{\zeta_{\mathsf{op}}^2}{\sigma_r^{\natural 4}}\sqrt{\frac{\mu^{\natural} r}{n_1}}. \tag{6.13b}$$

In fact, for each $m \in [n_1]$*, we further have*

$$\|\boldsymbol{Z}_{m,\cdot}\|_2 \lesssim \frac{\zeta_{\mathsf{op},m}}{\sigma_r^{\natural 2}}\sqrt{\frac{\mu^{\natural} r}{n_1}} + \left\|\boldsymbol{U}_{m,\cdot}^{\natural}\right\|_2\left(\kappa^{\natural 2}\sqrt{\frac{\mu^{\natural} r}{n_1}}\frac{\zeta_{\mathsf{op}}}{\sigma_r^{\natural 2}} + \kappa^{\natural 2}\frac{\zeta_{\mathsf{op}}^2}{\sigma_r^{\natural 4}}\right), \tag{6.13c}$$

$$\|\boldsymbol{\Psi}_{m,\cdot}\|_2 \lesssim \kappa^{\natural 2}\frac{\zeta_{\mathsf{op}}\zeta_{\mathsf{op},m}}{\sigma_r^{\natural 4}}\sqrt{\frac{\mu^{\natural} r}{n_1}} + \left\|\boldsymbol{U}_{m,\cdot}^{\natural}\right\|_2\left(\kappa^{\natural 2}\sqrt{\frac{\mu^{\natural} r}{n_1}}\frac{\zeta_{\mathsf{op}}}{\sigma_r^{\natural 2}} + \kappa^{\natural 2}\frac{\zeta_{\mathsf{op}}^2}{\sigma_r^{\natural 4}}\right). \tag{6.13d}$$

REMARK 5. The interested reader might wonder why Theorem 5 is not valid for small n_1 (e.g., (6.11) does not hold when $n_1 = 2$), and we provide some intuition here. Recall from (6.7) that the diagonal of the sample Gram matrix can be significantly biased, and the HeteroPCA algorithm uses the off-diagonal information to iteratively estimate and refine the diagonal. When n_1 is small, the (untrustworthy) diagonal entries account for a non-negligible fraction of all entries of the entire sample Gram matrix, and as a result, we cannot hope to debias the diagonal reliably by HeteroPCA using only off-diagonal observations.

The expressions (6.12) and (6.13) make apparent a key decomposition of the estimation error. As we shall see, the term $\boldsymbol{Z}$ is often the dominant term, which captures both the first-order and second-order approximation (w.r.t. the noise matrix $\boldsymbol{E}$) of the estimation error. Unless the noise level σ is very small, we cannot simply ignore the second-order term $\mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{E}\boldsymbol{E}^{\top}\right)\boldsymbol{U}^{\natural}\left(\boldsymbol{\Sigma}^{\natural}\right)^{-2}$, as it is not necessarily dominated in size by the linear mapping term $\boldsymbol{E}\boldsymbol{V}^{\natural}\left(\boldsymbol{\Sigma}^{\natural}\right)^{-1}$. The simple and closed-form expression of $\boldsymbol{Z}$ — in conjunction with the fact that $\boldsymbol{\Psi}$ is well-controlled — plays a crucial role when developing a non-asymptotic distributional theory.

While Theorem 5 is established mainly to help derive distributional characterizations for PCA, we remark that our analysis also delivers $\ell_{2,\infty}$ statistical guarantees in terms of estimating $\boldsymbol{U}^{\natural}$ (see Lemma 6 in the appendix). More specifically, our analysis asserts that

$$\left\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^{\natural}\right\|_{2,\infty} \lesssim \frac{\zeta_{\mathsf{op}}}{\sigma_r^{\natural 2}}\sqrt{\frac{\mu^{\natural} r}{n_1}} \tag{6.14}$$

with high probability, under the conditions of Theorem 5. It is perhaps helpful to compare (6.14) with prior $\ell_{2,\infty}$ theory concerning estimation of $\boldsymbol{U}^{\natural}$.

- We first compare Theorem 5 with the recent work (Agterberg, Lubberts and Priebe, 2022, Theorem 2), which focused on the regime $n_2 \gtrsim n_1$ and showed that

$$\inf_{\boldsymbol{O}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{U}\boldsymbol{O} - \boldsymbol{U}^{\natural}\right\|_{2,\infty} \lesssim \left(\frac{\sigma^2}{\sigma_r^{\natural 2}}\sqrt{rn_1n_2}\log n + \kappa^{\natural}\frac{\sigma}{\sigma_r^{\natural}}\sqrt{rn_1\log n}\right)\sqrt{\frac{\mu^{\natural} r}{n_1}} \asymp \frac{\zeta_{\mathsf{op}}}{\sigma_r^{\natural 2}}\sqrt{\frac{\mu^{\natural} r^2}{n_1}}$$

under the noise condition $\sigma\sqrt{n_2} \ll \sigma_r^\natural/(\kappa^\natural\sqrt{r\log n})$ (in addition to a few other conditions omitted here). Note that when $\kappa^\natural, \mu^\natural, r \asymp 1$, their $\ell_{2,\infty}$ error bound resembles (6.14), but the condition $\sigma\sqrt{n_2} \ll \sigma_r^\natural/\sqrt{\log n}$ required therein is much stronger than the noise condition $\zeta_{\mathsf{op}} \ll \sigma_r^{\natural 2}$ — which is equivalent to $\sigma\sqrt[4]{n_1 n_2} \ll \sigma_r^\natural/\sqrt{\log n}$ when $n_2 \gtrsim n_1$ — imposed by our theory (see (6.11)). It is also worth emphasizing that the theory of Agterberg, Lubberts and Priebe (2022) is capable of accommodating dependent data (i.e. they only require the rows of $\boldsymbol{E}$ to be independent and allow dependence within rows), which is beyond the scope of the present paper.

- Compared with the $\ell_{2,\infty}$ estimation error guarantees for the diagonal-deleted spectral method in Cai et al. (2021, Theorem 1), our bound (6.14) is able to get rid of the bias term incurred by diagonal deletion (see Cai et al. (2021, Equation (17))), thus improving upon this prior result.

It should be noted that fine-grained perturbation results akin to Theorem 5 were also developed for the SVD algorithm in an earlier version of this paper on arXiv, as detailed in Yan, Chen and Fan (2021, Section 6.1). Subsequently, Yan and Wainwright (2024) presented more refined results for cases where the entries of the noise matrix $\boldsymbol{E}$ follow a sub-Gaussian distribution, with further information available in Appendix F therein.

Before concluding this section, it is natural to ask whether Theorem 5 can be used to conduct subspace inference when every entry of $\boldsymbol{E}$ is allowed to have completely difference variance. To begin with, for a broad class of $\boldsymbol{E}$ with independent and heteroskedastic components, we can readily apply Theorem 5 to obtain a distributional theory for HeteroPCA when estimating $\boldsymbol{U}^\natural$. Caution needs to be exercised, however, when it comes to confidence interval construction. On closer inspection, evaluating $\boldsymbol{Z}$ (i.e., the first- and second-order approximation of the subspace estimation error) in Theorem 5 requires knowledge about the right singular subspace $\boldsymbol{V}^\natural$ of $\boldsymbol{M}^\natural$, which might sometimes be difficult or even infeasible to estimate in the unbalanced regime where $n_2 \gg n_1$. As a result, our theory is not guaranteed to deliver useful inferential methods for such cases, unless additional information about $\boldsymbol{V}^\natural$ is available.

7. Discussion. In this paper, we have developed a suite of statistical inference procedures to construct confidence regions for PCA in the presence of missing data and heterogeneous corruption, which should be easy-to-use in practice due to their data-driven nature. Compared to other prior algorithms like the SVD-based approach and the diagonal-deleted spectral method, the solution developed based on HeteroPCA enjoys a broadened applicability range without compromising statistical efficiency. The fine-grained distributional characterizations we have developed are non-asymptotic, which naturally lend themselves to high-dimensional settings.

Moving forward, there are a variety of directions that are worthy of further investigation.

- *Improved dependency on κ, μ, r and κ_ω.* In our general theorems (see Theorems 11-14 in the appendix), we allow κ, μ, r and κ_ω to grow. However, our theoretical results scale suboptimally with these problem parameters. It remains unclear how to sharpen the dependency on these parameters, which might require developing more refined analysis techniques.
- *Approximate low-rank structure.* Our results assume exact low-rank structure of the spiked component $\boldsymbol{S}^\star$ of the covariance matrix. In reality, there is no shortage of applications where $\boldsymbol{S}^\star$ is at best approximately low-rank. How to develop trustworthy inference procedures in the presence of approximate low-rank structure? Unfortunately, our current leave-one-out analysis framework relies heavily on the exact rank-r structure (unless $\sigma_{r+1}^\star$ is extremely small); new analysis ideas are needed in order to tackle approximate low-rank structure.

- *General missing pattern.* Uncertainty quantification in the face of heterogeneous missing patterns is another important topic of practical value. Consider, for example, the case where the entries in the same row of $\boldsymbol{X}$ are sampled with the same rate (i.e., the (l, j)-th entry of $\boldsymbol{X}$ is observed with probability p_l). Then by constructing the following data matrix via inverse probability weighting

$$\Big[\mathsf{diag}\big(p_1, p_2, \cdots, p_d\big)\Big]^{-1}\boldsymbol{Y},$$

  we obtain an unbiased estimate of $\boldsymbol{X}$, and the theory developed can be readily extended to perform valid inference. Note that we can also replace $\{p_l\}$ via their empirical estimates in the inference procedures. Nevertheless, in the more general case where the sampling rates are allowed to vary across all locations, it is unclear how to construct an unbiased estimate of $\boldsymbol{X}$ without knowing the per-entry sampling rates in advance; hence, our theory fails to accommodate this general scenario. Extending our current results to such general sampling patterns might call for new analysis tools.
- *Inference for individual principal components.* Moving beyond inference and uncertainty quantification for the principal subspace and the spiked covariance matrix, it is interesting to investigate how to conduct valid inference on individual principal components, particularly when the associated eigengap is vanishingly small (Li et al., 2021).
- *Extension to unknown mean, dependent or adversarial noise.* If the observed data are inherently biased with *a priori* unknown means, how to properly compensate for the bias? What if the noise components are inter-dependent, and what if the observed data samples are further corrupted by a non-negligible fraction of adversarial outliers?
- *Minimax-optimal estimation and inference.* As recognized in the matrix completion literature (Keshavan, Montanari and Oh, 2010b; Ma et al., 2020; Chen, Liu and Li, 2020), spectral methods alone are in general unable to yield minimax-optimal statistical accuracy in the presence of missing data, given that spectral methods inherently treat the missingness effect as some sort of "noise". The same message — namely, sub-optimality of HeteroPCA in the face of missing data — carries over to the PCA setting considered herein. We conjecture that a subsequent refinement procedure (e.g., gradient descent tailored to compute the maximum likelihood estimate) is needed in order to reach minimax optimality, and we leave this for future investigation.
- *Applications in financial econometrics*. In addition to applications to the uncertainty quantification in the matrix completion problems in recommender system, the inferential procedure and analysis tools we have developed in this paper have applications in finance and econometrics. For example, our analysis and results for principal subspace are useful in testing factor structures in famous Fama-French factor models, and can also be used in sector/industry clustering using stock returns (Porter et al., 1998); our results on uncertainty quantification for the spiked covariance matrix could also shed light on how to better quantify the risk in portfolio optimization that takes into account on the uncertainty in the risk estimation.

Acknowledgements. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR YIP award FA9550-19-1-0030, by the ONR grant N00014-19-1-2120, by the ARO grant W911NF-20-1-0097, and by the NSF grants CCF-1907661, IIS-2218713, IIS-2218773, DMS-2014279, CCF-2221009. J. Fan is supported in part by the ONR grants N00014-19-1-2120, N00014-22-1-2340, by the NSF grants DMS-1662139, DMS-1712591, DMS-2052926, DMS-2053832, DMS-2210833, and by the NIH grant 2R01-GM072611-15. Y. Yan is supported in part by the Charlotte Elizabeth Procter Honorific Fellowship from Princeton University and the Norbert Wiener Postdoctoral Fellowship from MIT.

REFERENCES

ABBE, E., FAN, J., WANG, K., ZHONG, Y. et al. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics* **48** 1452–1474.

AGTERBERG, J., LUBBERTS, Z. and PRIEBE, C. (2022). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*.

BAI, J. and WANG, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics* **8** 53–80.

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase Transition of the Largest Eigenvalue for Nonnull Complex Sample Covariance Matrices. *Annals of Probability* 1643–1697.

BALZANO, L., CHI, Y. and LU, Y. M. (2018). Streaming PCA and subspace tracking: The missing data case. *Proceedings of the IEEE* **106** 1293–1310.

BAO, Z., DING, X. and WANG, K. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics* **49** 370–392.

BAO, Z., DING, X., WANG, J. and WANG, K. (2022a). Statistical inference for principal components of spiked covariance matrices. *The Annals of Statistics* **50** 1144–1169.

BAO, Z., DING, X., WANG, J. and WANG, K. (2022b). Statistical inference for principal components of spiked covariance matrices. *The Annals of Statistics* **50** 1144–1169.

BLOEMENDAL, A., KNOWLES, A., YAU, H.-T. and YIN, J. (2016). On the principal components of sample covariance matrices. *Probability theory and related fields* **164** 459–552.

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics* **45** 615–646.

CAI, C., POOR, H. V. and CHEN, Y. (2022). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *IEEE Transactions on Information Theory* **69** 407–452.

CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* **46** 60–89.

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. *The Annals of Statistics* **49** 944–967.

CAI, C., LI, G., POOR, H. V. and CHEN, Y. (2022). Nonconvex Low-Rank Tensor Completion from Noisy Data. *Operations Research* **70** 1219–1237.

CANDÈS, E. (2014). Mathematics of sparsity (and a few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea* **123**.

CANDÈS, E. and PLAN, Y. (2010). Matrix Completion With Noise. *Proceedings of the IEEE* **98** 925 -936.

CANDÈS, E. and RECHT, B. (2009). Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics* **9** 717-772.

CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics* **47** 2405–2439.

CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics* **51** 2194–2220.

CHEN, Y., CHENG, C. and FAN, J. (2021). Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *Annals of Statistics* **49** 435–458.

CHEN, P., GAO, C. and ZHANG, A. Y. (2022). Partial recovery for top-k ranking: optimality of MLE and suboptimality of the spectral method. *The Annals of Statistics* **50** 1618–1652.

CHEN, J., LIU, D. and LI, X. (2020). Nonconvex Rectangular Matrix Completion via Gradient Descent without $\ell_{2,\infty}$ Regularization. *IEEE Transactions on Information Theory*.

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

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019a). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* **116** 22931–22937.

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2019b). Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval. *Mathematical Programming* **176** 5-37.

CHEN, Y., FAN, J., MA, C. and WANG, K. (2019c). Spectral Method and Regularized MLE Are Both Optimal for Top-K Ranking. *Annals of Statistics* **47** 2204-2235.

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 Journal on Optimization* **30** 3098–3121.

CHEN, Y., CHI, Y., FAN, J., MA, C. et al. (2021a). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning* **14** 566–806.

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2021b). Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data. *The Annals of Statistics* **49** 2948–2971.

CHENG, C., WEI, Y. and CHEN, Y. (2021). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory* **67** 7380–7419.

CHERNOZHUKOV, V., HANSEN, C., LIAO, Y. and ZHU, Y. (2023). Inference for low-rank models. *The Annals of statistics* **51** 1309–1330.

CHI, Y., LU, Y. M. and CHEN, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing* **67** 5239 - 5269.

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

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.

DING, X. (2020). High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli* **26** 387–417.

DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics* **46** 1742.

EL KAROUI, N. (2015). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields* 1–81.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* **110** 14557–14562.

ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory* 321–358.

FAN, J., LI, K. and LIAO, Y. (2021). Recent developments on factor models and its applications in econometric learning. *Annual Review of Financial Economics* **13** to appear.

FAN, J., WANG, W. and ZHONG, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research* **18** 1–42.

FAN, J. and YAO, Q. (2017). *The elements of financial econometrics*. Cambridge University Press.

FAN, J., FAN, Y., HAN, X. and LV, J. (2020). Asymptotic Theory of Eigenvectors for Random Matrices with Diverging Spikes. *Journal of the American Statistical Association* 1–14.

FAN, J., WANG, K., ZHONG, Y. and ZHU, Z. (2021). Robust high-dimensional factor models with applications to statistical machine learning. *Statistical Science* **36** 303–327.

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

GAGLIARDINI, P., OSSOLA, E. and SCAILLET, O. (2019). Estimation of large dimensional conditional factor models in finance. *Swiss Finance Institute Research Paper* **19-46**.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15** 2869–2909.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics* **29** 295–327.

JOHNSTONE, I. M. and PAUL, D. (2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE* **106** 1277–1292.

JOLLIFFE, I. T. (1986). Principal components in regression analysis. In *Principal component analysis* 129–155. Springer.

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

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010b). Matrix Completion From a Few Entries. *IEEE Transactions on Information Theory* **56** 2980 -2998.

KOLTCHINSKII, V., LÖFFLER, M. and NICKL, R. (2020). Efficient estimation of linear functionals of principal components. *The Annals of Statistics* **48** 464–490.

LEI, L. (2019). Unified $\ell_{2,\infty}$ Eigenspace Perturbation Theory for Symmetric Random Matrices. *arXiv preprint arXiv:1909.04798*.

LI, G., CAI, C., GU, Y., POOR, H. V. and CHEN, Y. (2021). Minimax Estimation of Linear Functions of Eigenvectors in the Face of Small Eigen-Gaps. *arXiv preprint arXiv:2104.03298*.

LING, S. (2022). Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods. *Applied and Computational Harmonic Analysis* **60** 20–52.

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40** 1637–1664.

LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058.

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. *Foundations of Computational Mathematics* **20** 451–632.

MONTANARI, A., RUAN, F. and YAN, J. (2018). Adapting to unknown noise distribution in matrix denoising. *arXiv preprint arXiv:1810.02954*.

MONTANARI, A. and SUN, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics* **71** 2381–2425.

NADLER, B. (2008). Finite Sample Approximation Results for Principal Component Analysis: A Matrix Perturbation Approach. *The Annals of Statistics* 2791–2817.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research* 1665–1697.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45** 158–195.

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

PAVEZ, E. and ORTEGA, A. (2020). Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory*.

PORTER, M. E. et al. (1998). *Clusters and the new economics of competition* **76**. Harvard Business Review Boston.

REN, Z., SUN, T., ZHANG, C. and ZHOU, H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics* **43** 991–1026.

SCHÖNEMANN, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31** 1–10.

SUN, R. and LUO, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory* **62** 6535–6579.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202.

VASWANI, N., CHI, Y. and BOUWMANS, T. (2018). Rethinking PCA for modern data sets: Theory, algorithms, and applications. *Proceedings of the IEEE* **106** 1274–1276.

VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.

WAHBA, G. (1965). A least squares estimate of satellite attitude. *SIAM review* **7** 409–409.

XIA, D. (2019). Confidence Region of singular subspaces for Low-Rank matrix regression. *IEEE Transactions on Information Theory* **65** 7437–7459.

XIA, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics* **15** 3798–3851.

XIA, D. and YUAN, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 58–77.

XIE, F. (2021). Entrywise limit theorems of eigenvectors and their one-step refinement for sparse random graphs. *arXiv preprint arXiv:2106.09840*.

YAN, Y., CHEN, Y. and FAN, J. (2021). Inference for heteroskedastic PCA with missing data (full version). *arXiv preprint arXiv:2107.12365*.

YAN, Y. and WAINWRIGHT, M. J. (2024). Entrywise Inference for Causal Panel Data: A Simple and Instance-Optimal Approach. *arXiv preprint arXiv:2401.13665*.

ZHANG, A. R., CAI, T. T. and WU, Y. (2022). Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics* **50** 53–80.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76** 217–242.

ZHENG, Q. and LAFFERTY, J. (2016). Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv:1605.07051*.

ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bound for phase synchronization. *SIAM Journal on Optimization*.

ZHOU, Y. and CHEN, Y. (2023a). Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA. *arXiv preprint arXiv:2303.06198*.

ZHOU, Y. and CHEN, Y. (2023b). Heteroskedastic tensor clustering. *arXiv preprint arXiv:2311.02306* **3**.

ZHU, Z., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84** 2000–2031.