The Annals of Statistics
2018, Vol. 46, No. 1, 60–89
https://doi.org/10.1214/17-AOS1541

RATE-OPTIMAL PERTURBATION BOUNDS FOR SINGULAR SUBSPACES WITH APPLICATIONS TO HIGH-DIMENSIONAL STATISTICS

BY T. TONY CAI[1,*] AND ANRU ZHANG[†]

University of Pennsylvania[*] and University of Wisconsin–Madison[†]

Perturbation bounds for singular spaces, in particular Wedin's $\sin\Theta$ theorem, are a fundamental tool in many fields including high-dimensional statistics, machine learning and applied mathematics. In this paper, we establish separate perturbation bounds, measured in both spectral and Frobenius $\sin\Theta$ distances, for the left and right singular subspaces. Lower bounds, which show that the individual perturbation bounds are rate-optimal, are also given.

The new perturbation bounds are applicable to a wide range of problems. In this paper, we consider in detail applications to low-rank matrix denoising and singular space estimation, high-dimensional clustering and canonical correlation analysis (CCA). In particular, separate matching upper and lower bounds are obtained for estimating the left and right singular spaces. To the best of our knowledge, this is the first result that gives different optimal rates for the left and right singular spaces under the same perturbation.

1. Introduction. Singular value decomposition (SVD) and spectral methods have been widely used in statistics, probability, machine learning and applied mathematics as well as many applications. Examples include low-rank matrix denoising [Donoho and Gavish (2014), Shabalin and Nobel (2013), Yang, Ma and Buja (2016)], matrix completion [Candès and Recht (2009), Candès and Tao (2010), Chatterjee (2014), Gross (2011), Keshavan, Montanari and Oh (2010)], principle component analysis [Anderson (2003), Cai, Ma and Wu (2013, 2015b), Johnstone and Lu (2009)], canonical correlation analysis [Gao, Ma and Zhou (2014), Gao et al. (2015), Hardoon, Szedmak and Shawe-Taylor (2004), Hotelling (1936)], community detection [Balakrishnan et al. (2011), Lei and Rinaldo (2015), Rohe, Chatterjee and Yu (2011), von Luxburg, Belkin and Bousquet (2008)]. Specific applications include collaborative filtering (the Netflix problem) [Goldberg et al. (1992)], multi-task learning [Argyriou, Evgeniou and Pontil (2008)], system identification [Liu and Vandenberghe (2009)] and sensor localization [Singer and Cucuringu (2010), Candes and Plan (2010)], among many others. In addition, the

Received May 2016; revised November 2016.

[1]Supported in part by NSF Grant DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

MSC2010 subject classifications. Primary 62H12, 62C20; secondary 62H25.

Key words and phrases. Canonical correlation analysis, clustering, high-dimensional statistics, low-rank matrix denoising, perturbation bound, singular value decomposition, $\sin\Theta$ distances, spectral method.

SVD is often used to find a "warm start" for more delicate iterative algorithms; see, for example, Cai, Li and Ma (2016), Sun and Luo (2015).

Perturbation bounds, which concern how the spectrum changes after a small perturbation to a matrix, often play a critical role in the analysis of the SVD and spectral methods. To be more specific, for an approximately low-rank matrix X and a perturbation matrix Z, it is crucial in many applications to understand how much the left or right singular spaces of X and $X + Z$ differ from each other. This problem has been widely studied in the literature [Davis and Kahan (1970), Stewart (1991, 2006), Wedin (1972), Weyl (1912), Yu, Wang and Samworth (2015)]. Among these results, the $\sin\Theta$ theorems, established by Davis and Kahan (1970) and Wedin (1972), have become fundamental tools and are commonly used in applications. While Davis and Kahan (1970) focused on eigenvectors of symmetric matrices, Wedin's $\sin\Theta$ theorem studies the more general singular vectors for asymmetric matrices and provides a uniform perturbation bound for both the left and right singular spaces in terms of the singular value gap and perturbation level.

Several generalizations and extensions have been made in different settings after the seminal work of Wedin (1972). For example, Vu (2011), Shabalin and Nobel (2013), O'Rourke, Vu and Wang (2013), Wang (2015) considered the rotations of singular vectors after random perturbations; Fan, Wang and Zhong (2016) gave an ℓ_∞ eigenvector perturbation bound and used the result for robust covariance estimation. See also Dopico (2000), Stewart (2006).

Despite its wide applicability, Wedin's perturbation bound is not sufficiently precise for some analyses, as the bound is uniform for both the left and right singular spaces. It clearly leads to suboptimal result if the left and right singular spaces change in different orders of magnitude after the perturbation. In a range of applications, especially when the row and column dimensions of the matrix differ significantly, it is even possible that one side of the singular space can be accurately recovered, while the other side cannot. The numerical experiment given in Section 2.3 provides a good illustration for this point. It can be seen from the experiment that the left and right singular perturbation bounds behave distinctly when the row and column dimensions are significantly different. Furthermore, for a range of applications, the primary interest only lies in one of the singular spaces. For example, in the analysis of bipartite network data, such as the Facebook user-public-page-subscription network, the interest is often focused on grouping the public pages (or grouping the users). This is the case for many clustering problems. See Section 4 for further discussions.

In this paper, we establish separate perturbation bounds for the left and right singular subspaces. The bounds are measured in both the spectral and Frobenius $\sin\Theta$ distances, which are equivalent to several widely used losses in the literature. We also derive lower bounds that are within a constant factor of the corresponding upper bounds. These results together show that the obtained perturbation bounds are rate-optimal.

The newly established perturbation bounds are applicable to a wide range of problems in high-dimensional statistics. In this paper, we discuss in detail the applications of the perturbation bounds to the following high-dimensional statistical problems:

1. *Low-rank matrix denoising and singular space estimation*: Suppose one observes a low-rank matrix with random additive noise and wishes to estimate the mean matrix or its left or right singular spaces. Such a problem arises in many applications. We apply the obtained perturbation bounds to study this problem. Separate matching upper and lower bounds are given for estimating the left and right singular spaces. These results together establish the optimal rates of convergence. Our analysis shows an interesting phenomenon that in some settings it is possible to accurately estimate the left singular space but not the right one and vice versa. To the best of our knowledge, this is the first result that gives different optimal rates for the left and right singular spaces under the same perturbation. Another fact we observe is that in certain class of low-rank matrices, one can stably recover the original matrix if and only if one can accurately recover both its left and right singular spaces.
2. *High-dimensional clustering*: Unsupervised learning is an important problem in statistics and machine learning with a wide range of applications. We apply the perturbation bounds to the analysis of clustering for high-dimensional Gaussian mixtures. Particularly in a high-dimensional two-class clustering setting, we propose a simple PCA-based clustering method and use the obtained perturbation bounds to prove matching upper and lower bounds for the misclassification rates.
3. *Canonical correlation analysis* (*CCA*): CCA is a commonly used tools in multivariate analysis to identify and measure the associations among two sets of random variables. The perturbation bounds are also applied to analyze CCA. Specifically, we develop sharper upper bounds for estimating the left and right canonical correlation directions. To the best of our knowledge, this is the first result that captures the phenomenon that in some settings it is possible to accurately estimate one side of canonical correlation directions but not the other side.

In addition to these applications, the perturbation bounds can also be applied to the analysis of *community detection in bipartite networks, multidimensional scaling, cross-covariance matrix estimation, and singular space estimation for matrix completion* and other problems to yield better results than what are known in the literature. These applications demonstrate the usefulness of the newly established perturbation bounds.

The rest of the paper is organized as follows. In Section 2, after basic notation and definitions are introduced, the perturbation bounds are presented separately for the left and right singular subspaces. Both the upper bounds and lower bounds are provided. We then apply the newly established perturbation bounds to low-rank

matrix denoising and singular space estimation, high-dimensional clustering and canonical correlation analysis in Sections 3–5. Section 6 presents some numerical results and other potential applications are briefly discussed in Section 7. The main theorems are proved in Section 8 and the proofs of some additional technical results are given in the supplementary material [Cai and Zhang (2017)].

2. Rate-optimal perturbation bounds for singular subspaces. We establish in this section rate-optimal perturbation bounds for singular subspaces. We begin with basic notation and definitions that will be used in the rest of the paper.

2.1. *Notation and definitions.* For $a, b \in \mathbb{R}$, let $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$. Let $\mathbb{O}_{p,r} = \{V \in \mathbb{R}^{p\times r} : V^\intercal V = I_r\}$ be the set of all $p \times r$ orthonormal columns and write $\mathbb{O}_p$ for $\mathbb{O}_{p,p}$, the set of p-dimensional orthogonal matrices. For a matrix $A \in \mathbb{R}^{p_1\times p_2}$, write the SVD as $A = U\Sigma V^\intercal$, where $\Sigma = \operatorname{diag}\{\sigma_1(A), \sigma_2(A), \ldots\}$ with the singular values $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq 0$ in descending order. In particular, we use $\sigma_{\min}(A) = \sigma_{\min(p_1,p_2)}(A)$, $\sigma_{\max}(A) = \sigma_1(A)$ as the smallest and largest nontrivial singular values of A. Several matrix norms will be used in the paper: $\|A\| = \sigma_1(A)$ is the spectral norm; $\|A\|_F = \sqrt{\sum_i \sigma_i^2(A)}$ is the Frobenius norm; and $\|A\|_* = \sum_i \sigma_i(A)$ is the nuclear norm. We denote $\mathbb{P}_A \in \mathbb{R}^{p_1\times p_1}$ as the projection operator onto the column space of A, which can be written as $\mathbb{P}_A = A(A^\intercal A)^\dagger A^\intercal$. Here, $(\cdot)^\dagger$ represents the Moore–Penrose pseudoinverse. Given the SVD $A = U\Sigma V^\intercal$ with Σ nonsingular, a simpler form for $\mathbb{P}_A$ is $\mathbb{P}_A = UU^\intercal$. We adopt the R convention to denote the submatrix: $A_{[a:b,c:d]}$ represents the a-to-bth row, c-to-dth column of matrix A; we also use $A_{[a:b,:]}$ and $A_{[:,c:d]}$ to represent a-to-bth full rows of A and c-to-dth full columns of A, respectively. We use $C, C_0, c, c_0, \ldots$ to denote generic constants, whose actual values may vary from time to time.

We use the $\sin\Theta$ *distance* to measure the difference between two $p \times r$ orthogonal columns V and $\hat{V}$. Suppose the singular values of $V^\intercal\hat{V}$ are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. Then we call

$$\Theta(V, \hat{V}) = \operatorname{diag}\big(\cos^{-1}(\sigma_1), \cos^{-1}(\sigma_2), \ldots, \cos^{-1}(\sigma_r)\big)$$

as the principle angles. A quantitative measure of distance between the column spaces of V and $\hat{V}$ is then $\|\sin\Theta(\hat{V}, V)\|$ or $\|\sin\Theta(\hat{V}, V)\|_F$. Some more convenient characterizations and properties of the $\sin\Theta$ distances will be given in Lemma 1 in Section 8.1.

2.2. *Perturbation upper bounds and lower bounds.* We are now ready to present the perturbation bounds for the singular subspaces. Let $X \in \mathbb{R}^{p_1\times p_2}$ be an approximately low-rank matrix and let $Z \in \mathbb{R}^{p_1\times p_2}$ be a "smal" perturbation matrix. Our goal is to provide separate and rate-sharp bounds for the $\sin\Theta$ distances between the left singular subspaces of X and $X + Z$ and between the right singular subspaces of X and $X + Z$.

Suppose X is approximately rank-r with the SVD $X = U\Sigma V^\intercal$, where a significant gap exists between $\sigma_r(X)$ and $\sigma_{r+1}(X)$. The leading r left and right singular vectors of X are of particular interest. We decompose X as follows:

$$X = \begin{bmatrix} U & U_\perp \end{bmatrix} \cdot \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \cdot \begin{bmatrix} V^\intercal \\ V_\perp^\intercal \end{bmatrix}, \tag{2.1}$$

where $U \in \mathbb{O}_{p_1,r}, V \in \mathbb{O}_{p_2,r}$, $\Sigma_1 = \operatorname{diag}(\sigma_1(X), \ldots, \sigma_r(X)) \in \mathbb{R}^{r\times r}$, $\Sigma_2 = \operatorname{diag}(\sigma_{r+1}(X), \ldots) \in \mathbb{R}^{(p_1-r)\times(p_2-r)}$, $[U\ U_\perp] \in \mathbb{O}_{p_1}, [V\ V_\perp] \in \mathbb{O}_{p_2}$ are orthogonal matrices.

Let Z be a perturbation matrix and let $\hat{X} = X + Z$. Partition the SVD of $\hat{X}$ in the same way as in (2.1),

$$\hat{X} = X + Z = \begin{bmatrix} \hat{U} & \hat{U}_\perp \end{bmatrix} \cdot \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \cdot \begin{bmatrix} \hat{V}^\intercal \\ \hat{V}_\perp^\intercal \end{bmatrix}, \tag{2.2}$$

while $\hat{U}, \hat{U}_\perp, \hat{\Sigma}_1, \hat{\Sigma}_2, \hat{V}$ and $\hat{V}_\perp$ have the same structures as $U, U_\perp, \Sigma_1, \Sigma_2, V$ and $V_\perp$. Decompose the perturbation Z into four blocks:

$$Z = Z_{11} + Z_{12} + Z_{21} + Z_{22}, \tag{2.3}$$

where

$$Z_{11} = \mathbb{P}_U Z \mathbb{P}_V, \qquad Z_{21} = \mathbb{P}_{U_\perp} Z \mathbb{P}_V,$$
$$Z_{12} = \mathbb{P}_U Z \mathbb{P}_{V_\perp}, \qquad Z_{22} = \mathbb{P}_{U_\perp} Z \mathbb{P}_{V_\perp}.$$

Define

$$z_{ij} := \|Z_{ij}\| \qquad \text{for } i, j = 1, 2.$$

Theorem 1 below provides separate perturbation bounds for the left and right singular subspaces in terms of both spectral and Frobenius $\sin\Theta$ distances.

THEOREM 1 (Perturbation bounds for singular subspaces). *Let X, $\hat{X}$ and Z be given as (2.1)–(2.3). Denote*

$$\alpha := \sigma_{\min}\big(U^\intercal \hat{X} V\big) \quad \textit{and} \quad \beta := \big\|U_\perp^\intercal \hat{X} V_\perp\big\|.$$

If $\alpha^2 > \beta^2 + z_{12}^2 \wedge z_{21}^2$, then

$$\begin{aligned} \big\|\sin\Theta(V, \hat{V})\big\| &\le \frac{\alpha z_{12} + \beta z_{21}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge 1, \\ \big\|\sin\Theta(V, \hat{V})\big\|_F &\le \frac{\alpha \|Z_{12}\|_F + \beta \|Z_{21}\|_F}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge \sqrt{r}. \end{aligned} \tag{2.4}$$

$$\begin{aligned} \big\|\sin\Theta(U, \hat{U})\big\| &\le \frac{\alpha z_{21} + \beta z_{12}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge 1, \\ \big\|\sin\Theta(U, \hat{U})\big\|_F &\le \frac{\alpha \|Z_{21}\|_F + \beta \|Z_{12}\|_F}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge \sqrt{r}. \end{aligned} \tag{2.5}$$

One can see the respective effects of the perturbation on the left and right singular spaces. In particular, if $z_{12} \geq z_{21}$ (which is typically the case when $p_2 \gg p_1$), then Theorem 3 gives a smaller bound for $\|\sin\Theta(U, \hat{U})\|$ than for $\|\sin\Theta(V, \hat{V})\|$.

REMARK 1. The assumption $\alpha^2 > \beta^2 + z_{12}^2 \wedge z_{21}^2$ in Theorem 1 ensures that the amplitude of $U^\top \hat{X} V = \Sigma_1 + U^\top Z V$ dominates those of $U_\perp^\top \hat{X} V_\perp = \Sigma_2 + U_\perp^\top Z V_\perp$, $U^\top Z V_\perp$ and $U_\perp^\top Z V$, so that $\hat{U}$ and $\hat{V}$ can be close to U and V, respectively. This assumption essentially means that there exists significant gap between the rth and $(r+1)$st singular values of X and the perturbation term Z is bounded. We will show in Theorem 2 that $\hat{U}$ and $\hat{V}$ might be inconsistent when this condition fails to hold.

REMARK 2. Consider the setting where $X \in \mathbb{R}^{p_1 \times p_2}$ is a fixed rank-r matrix with $r \leq p_1 \ll p_2$, and $Z \in \mathbb{R}^{p_1 \times p_2}$ is a random matrix with i.i.d. standard normal entries. In this case, Z_{11}, Z_{12}, Z_{21}, and Z_{22} are all i.i.d. standard normal matrices of dimensions $r \times r$, $r \times (p_2 - r)$, $(p_1 - r) \times r$, and $(p_1 - r) \times (p_2 - r)$, respectively. By random matrix theory [see, e.g., Tao (2012), Vershynin (2012)], $\alpha \geq \sigma_r(X) - \|Z_{11}\| \geq \sigma_r(X) - C(\sqrt{p_1} + \sqrt{p_2})$, $\beta \leq C(\sqrt{p_1} + \sqrt{p_2})$, $z_{12} \leq C\sqrt{p_2}$ and $z_{21} \leq C\sqrt{p_1}$ for some constant $C > 0$ with high probability. When $\sigma_r(X) \geq C_{\text{gap}} p_2 / \sqrt{p_1}$ for some large constant C_{gap}, Theorem 3 immediately implies

$$\|\sin\Theta(V, \hat{V})\| \leq \frac{C\sqrt{p_2}}{\sigma_r(X)},$$

$$\|\sin\Theta(U, \hat{U})\| \leq \frac{C\sqrt{p_1}}{\sigma_r(X)}.$$

Further discussions on perturbation bounds for general sub-Gaussian perturbation matrix with matching lower bounds with be given in Section 3.

Theorem 1 gives upper bounds for the perturbation effects. We now establish lower bounds for the differences as measured by the $\sin\Theta$ distances. Theorem 2 first states that $\hat{U}$ and $\hat{V}$ might be inconsistent when the condition $\alpha^2 > \beta^2 + z_{12}^2 \wedge z_{21}^2$ fails to hold, and then provides the lower bounds that match those in (2.11) and (2.12), proving that the results given in Theorem 1 is essentially sharp. Theorem 2 also provides the worst-case matrix pair (X, Z) that nearly achieves the supremum in (2.9) and (2.7). The matrix pair shows where the lower bound is "close" to the upper bound, which is useful in understanding the fundamentals about singular subspace perturbations.

Before stating the lower bounds, we define the following class of (X, Z) pairs of $p_1 \times p_2$ matrices and perturbations:

$$\begin{aligned} \mathcal{F}_{r,\alpha,\beta,z_{21},z_{12}} = \big\{ & (X, Z) : \hat{X}, U, V \text{ are given as (2.1) and (2.2)}, \\ & \sigma_{\min}(U^\top \hat{X} V) \geq \alpha, \|U_\perp^\top \hat{X} V_\perp\| \leq \beta, \\ & \|Z_{12}\| \leq z_{12}, \|Z_{21}\| \leq z_{21} \big\}. \end{aligned} \tag{2.6}$$

In addition, we also define

$$\begin{aligned}(2.7)\qquad &\mathcal{G}_{\alpha,\beta,z_{21},z_{12},\tilde{z}_{21},\tilde{z}_{12}}\\ &\quad=\big\{(X,Z):\|Z_{21}\|_F\le\tilde{z}_{21},\|Z_{12}\|_F\le\tilde{z}_{12},(X,Z)\in\mathcal{F}_{r,\alpha,\beta,z_{21},z_{12}}\big\}.\end{aligned}$$

THEOREM 2 (Perturbation lower bound). *If $\alpha^2\le\beta^2+z_{12}^2\wedge z_{21}^2$ and $r\le\frac{p_1\wedge p_2}{2}$, then*

$$\inf_{\tilde{V}}\sup_{(X,Z)\in\mathcal{F}}\big\|\sin\Theta(V,\tilde{V})\big\|\ge\frac{1}{2\sqrt{2}}.\tag{2.8}$$

- *Provided that $\alpha^2>\beta^2+z_{12}^2+z_{21}^2$, $r\le\frac{p_1\wedge p_2}{2}$ we have the following lower bound for all estimate $\tilde{V}\in O_{p_2\times r}$ based on the observations $\hat{X}$:*

$$\inf_{\tilde{V}}\sup_{(X,Z)\in\mathcal{F}}\big\|\sin\Theta(V,\tilde{V})\big\|\ge\frac{1}{8\sqrt{10}}\bigg(\frac{\alpha z_{12}+\beta z_{21}}{\alpha^2-\beta^2-z_{12}^2\wedge z_{21}^2}\wedge 1\bigg).\tag{2.9}$$

In particular, if $X=\alpha UV^\intercal+\beta U_\perp V_\perp^\intercal$ and $Z=z_{12}UV_\perp^\intercal+z_{21}U_\perp V^\intercal$, then $(X,Z)\in\mathcal{F}$ and

$$\begin{aligned}\frac{1}{\sqrt{10}}\bigg(\frac{\alpha z_{12}+\beta z_{21}}{\alpha^2-\beta^2-z_{21}^2\wedge z_{12}^2}\wedge 1\bigg)&\le\big\|\sin\Theta(V,\hat{V})\big\|\\ &\le\bigg(\frac{\alpha z_{12}+\beta z_{21}}{\alpha^2-\beta^2-z_{21}^2\wedge z_{12}^2}\wedge 1\bigg),\end{aligned}$$

when $\hat{U},\hat{V}$ are the leading r left and right singular vectors of $\hat{X}$.

- *Provided that $\alpha^2>\beta^2+z_{12}^2+z_{21}^2$, $\tilde{z}_{21}^2\le rz_{21}^2$, $\tilde{z}_{12}^2\le rz_{12}^2$, $r\le\frac{p_1\wedge p_2}{2}$, we have the following lower bound for all estimator $\tilde{V}_1\in\mathbb{O}_{p_2\times r}$ based on the observations $\hat{X}$:*

$$\inf_{\tilde{V}_1}\sup_{(X,Z)\in\mathcal{G}}\big\|\sin\Theta(V,\tilde{V})\big\|_F\ge\frac{1}{8\sqrt{10}}\bigg(\frac{\alpha\tilde{z}_{12}+\beta\tilde{z}_{21}}{\alpha^2-\beta^2-z_{12}^2\wedge z_{21}^2}\wedge\sqrt{r}\bigg).\tag{2.10}$$

In particular, if $X=\alpha UV^\intercal+\beta U_\perp V_\perp^\intercal$, $Z=\tilde{z}_{12}UV_\perp^\intercal+\tilde{z}_{21}U_\perp V^\intercal$, then $(X,Z)\in\mathcal{G}$ and

$$\begin{aligned}\frac{1}{\sqrt{10}}\bigg(\frac{\alpha\tilde{z}_{12}+\beta\tilde{z}_{21}}{\alpha^2-\beta^2-z_{21}^2\wedge z_{12}^2}\wedge\sqrt{r}\bigg)&\le\big\|\sin\Theta(V,\hat{V})\big\|\\ &\le\bigg(\frac{\alpha\tilde{z}_{12}+\beta\tilde{z}_{21}}{\alpha^2-\beta^2-z_{21}^2\wedge z_{12}^2}\wedge\sqrt{r}\bigg),\end{aligned}$$

where $\hat{U},\hat{V}$ are respectively the leading r left and right singular vectors of $\hat{X}$.

The following Proposition 1, which provides upper bounds for the $\sin\Theta$ distances between leading singular vectors of a matrix A and arbitrary orthogonal columns W, can be viewed as another version of Theorem 1. For some applications, applying Proposition 1 might be more convenient than using Theorem 1 directly.

PROPOSITION 1. *Suppose $A \in \mathbb{R}^{p_1 \times p_2}$, $\tilde{V} = [V\ V_\perp] \in \mathbb{O}_{p_2}$ are right singular vectors of A, $V \in \mathbb{O}_{p_2,r}$, $V_\perp \in \mathbb{O}_{p_2,p_2-r}$ correspond to the first r and last $(p_2 - r)$ singular vectors, respectively. $\tilde{W} = [W\ W_\perp] \in \mathbb{O}_{p_2}$ is any orthogonal matrix with $W \in \mathbb{O}_{p_2,r}$, $W_\perp \in \mathbb{O}_{p_2,p_2-r}$. Given that $\sigma_r(AW) > \sigma_{r+1}(A)$, we have*

$$\|\sin\Theta(V, W)\| \le \frac{\sigma_r(AW)\|\mathbb{P}_{(AW)}AW_\perp\|}{\sigma_r^2(AW) - \sigma_{r+1}^2(A)} \wedge 1, \tag{2.11}$$

$$\|\sin\Theta(V, W)\|_F \le \frac{\sigma_r(AW)\|\mathbb{P}_{(AW)}AW_\perp\|_F}{\sigma_r^2(AW) - \sigma_{r+1}^2(A)} \wedge \sqrt{r}. \tag{2.12}$$

It is also of practical interest to provide perturbation bounds for a given subset of singular vectors and in particular for a given singular vector. The following Corollary 1 provides the one-sided perturbation bound for $\hat{U}_{[:,i:j]}$ and $\hat{V}_{[:,i:j]}$ when there are significant gaps between the $(i-1)$st and ith and between the jth and $(j+1)$st singular values and the perturbation is bounded. Particularly when $i = j$, Corollary 1 provides the upper bound for the perturbation of the ith left and right singular vectors of $\hat{X}$, $\hat{u}_i$ and $\hat{v}_i$.

COROLLARY 1 (Perturbation bounds for individual singular vectors). *Suppose $X, \hat{X}$ and Z are given as (2.1)–(2.3). For any $k \ge 1$, let $U_{(k)} = U_{[:,1:k]} \in \mathbb{O}_{p_1,k}$, $V_{(k)} = V_{[:,1:k]} \in \mathbb{O}_{p_2,k}$, and $U_{(k)\perp} \in \mathbb{O}_{p_1,p_1-k}$, $V_{(k)\perp} \in \mathbb{O}_{p_2,p_2-k}$ be the orthogonal complements. Denote*

$$\alpha^{(k)} = \sigma_{\min}\big(U_{(k)}^\intercal \hat{X} V_{(k)}\big), \qquad \beta^{(k)} = \|U_{(k)\perp}^\intercal \hat{X} V_{(k)\perp}\|,$$
$$z_{12}^{(k)} = \|U_{(k)}^\intercal Z V_{(k)\perp}\|, \qquad z_{21}^{(k)} = \|U_{(k)\perp}^\intercal Z V_{(k)}\|,$$

for $k = 1, \ldots, p_1 \wedge p_2$. We further define $\alpha^{(0)} = \infty$, $\beta^{(0)} = \|\hat{X}\|$, $z_{12}^{(0)} = z_{21}^{(0)} = 0$. For $1 \le i \le j \le p_1 \wedge p_2$, provided that $(\alpha^{(i-1)})^2 > (\beta^{(i-1)})^2 + (z_{12}^{(i-1)})^2 \wedge (z_{21}^{(i-1)})^2$ and $(\alpha^{(j)})^2 > (\beta^{(j)})^2 + (z_{12}^{(j)})^2 \wedge (z_{21}^{(j)})^2$, we have

$$\|\sin\Theta(\hat{V}_{[:,i:j]}, V_{[:,i:j]})\|$$
$$\le \Big\{ \sum_{k\in\{i-1,j\}} \Big(\frac{(\alpha^{(k)} z_{12}^{(k)} + \beta^{(k)} z_{21}^{(k)})}{(\alpha^{(k)})^2 - (\beta^{(k)})^2 - (z_{21}^{(k)})^2 \wedge (z_{12}^{(k)})^2} \Big)^2 \Big\}^{1/2} \wedge 1$$

and

$$\begin{aligned}&\big\|\sin\Theta(\hat{U}_{[:,i:j]}, U_{[:,i:j]})\big\|\\&\quad\le\left\{\sum_{k\in\{i-1,j\}}\left(\frac{(\alpha^{(k)}z_{21}^{(k)}+\beta^{(k)}z_{12}^{(k)})}{(\alpha^{(k)})^2-(\beta^{(k)})^2-(z_{21}^{(k)})^2\wedge(z_{12}^{(k)})^2}\right)^2\right\}^{1/2}\wedge 1.\end{aligned}$$

In particular, for any integer $1\le i\le p_1\wedge p_2$, *if* $(\alpha^{(i-1)})^2>(\beta^{(i-1)})^2+(z_{12}^{(i-1)})^2\wedge(z_{21}^{(i-1)})^2$ *and* $(\alpha^{(i)})^2>(\beta^{(i)})^2+(z_{12}^{(i)})^2\wedge(z_{21}^{(i)})^2$, $u_i, \hat{u}_i, v_i, \hat{v}_i$, *that is, the* i*th singular vectors of* X *and* $\hat{X}$, *are different by*

$$\sqrt{1-(v_i^{\intercal}\hat{v}_i)^2}\le\left\{\sum_{k=i-1}^{i}\left(\frac{(\alpha^{(k)}z_{12}^{(k)}+\beta^{(k)}z_{21}^{(k)})}{(\alpha^{(k)})^2-(\beta^{(k)})^2-(z_{21}^{(k)})^2\wedge(z_{12}^{(k)})^2}\right)^2\right\}^{1/2}\wedge 1,$$

$$\sqrt{1-(u_i^{\intercal}\hat{u}_i)^2}\le\left\{\sum_{k=i-1}^{i}\left(\frac{(\alpha^{(k)}z_{21}^{(k)}+\beta^{(k)}z_{12}^{(k)})}{(\alpha^{(k)})^2-(\beta^{(k)})^2-(z_{21}^{(k)})^2\wedge(z_{12}^{(k)})^2}\right)^2\right\}^{1/2}\wedge 1.$$

REMARK 3. The upper bound given in Corollary 1 is rate-optimal over the following set of (X, Z) pairs:

$$\begin{aligned}&\mathcal{H}_{\alpha^{(i-1)},\beta^{(i-1)},z_{12}^{(i-1)},z_{21}^{(i-1)},\alpha^{(j)},\beta^{(j)},z_{12}^{(j)},z_{21}^{(j)}}\\&\quad=\left\{(X,Z):\begin{array}{l}\sigma_{\min}\big(U_{(k)}^{\intercal}\hat{X}V_{(k)}\big)\ge\alpha^{(k)},\ \big\|U_{(k)\perp}^{\intercal}\hat{X}V_{(k)\perp}\big\|\le\beta^{(k)},\\ \big\|U_{(k)}^{\intercal}ZV_{(k)\perp}\big\|\le z_{12}^{(k)},\ \big\|U_{(k)\perp}^{\intercal}ZV_{(k)}\big\|\le z_{21}^{(k)},\end{array}\ k\in\{i-1,j\}\right\}.\end{aligned}$$

The detailed analysis can be carried out similarly to the one for Theorem 2.

2.3. *Comparisons with Wedin's* $\sin\Theta$ *theorem.* Theorems 1 and 2 together establish separate rate-optimal perturbation bounds for the left and right singular subspaces. We now compare the results with the well-known Wedin's $\sin\Theta$ theorem, which gives uniform upper bounds for the singular subspaces on both sides. Specifically, using the same notation as in Section 2.2, Wedin's $\sin\Theta$ theorem states that if $\sigma_{\min}(\hat{\Sigma}_1)-\sigma_{\max}(\Sigma_2)=\delta>0$, then

$$\max\big\{\big\|\sin\Theta(V,\hat{V})\big\|,\big\|\sin\Theta(U,\hat{U})\big\|\big\}\le\frac{\max\{\|Z\hat{V}\|,\|\hat{U}^{\intercal}Z\|\}}{\delta},$$

$$\max\big\{\big\|\sin\Theta(V,\hat{V})\big\|_F,\|\sin\Theta(U,\hat{U})\|_F\big\}\le\frac{\max\{\|Z\hat{V}\|_F,\|\hat{U}^{\intercal}Z\|_F\}}{\delta}.$$

When X are Z are symmetric, Theorem 1, Proposition 1 and Wedin's $\sin\Theta$ theorem provide similar upper bound for singular subspace perturbation.

As mentioned in the Introduction, the uniform bound on both left and right singular subspaces in Wedin's $\sin\Theta$ theorem might be suboptimal in some cases

when X or Z are asymmetric. For example, in the setting discussed in Remark 2, applying Wedin's theorem leads to

$$\max\{\|\sin\Theta(V,\hat{V})\|,\|\sin\Theta(U,\hat{U})\|\} \le \frac{C\max\{\sqrt{p_1},\sqrt{p_2}\}}{\sigma_r(X)},$$

which is suboptimal for $\|\sin\Theta(U,\hat{U})\|$ if $p_2 \gg p_1$.

3. Low-rank matrix denoising and singular-space estimation. In this section, we apply the perturbation bounds given in Theorem 1 for low-rank matrix denoising. It can be seen that the new perturbation bounds are particularly powerful when the matrix dimensions differ significantly. We also establish a matching lower bound for low-rank matrix denoising which shows that the results are rate-optimal.

As mentioned in the Introduction, accurate recovery of a low-rank matrix based on noisy observations has a wide range of applications, including magnetic resonance imaging (MRI) and relaxometry; see, for example, Candès, Sing-Long and Trzasko (2013), Shabalin and Nobel (2013) and the reference therein. This problem is also important in the context of dimensional reduction. Suppose one observes a low-rank matrix with additive noise,

$$Y = X + Z,$$

where $X = U\Sigma V^{\mathsf{T}} \in \mathbb{R}^{p_1\times p_2}$ is a low-rank matrix with $U \in \mathbb{O}_{p_1,r}$, $V \in \mathbb{O}_{p_2,r}$, and $\Sigma = \operatorname{diag}\{\sigma_1(X),\dots,\sigma_r(X)\} \in \mathbb{R}^{r\times r}$, and $Z \in \mathbb{R}^{p_1\times p_2}$ is an i.i.d. mean-zero sub-Gaussian matrix. The goal is to estimate the underlying low-rank matrix X or its singular values or singular vectors.

This problem has been actively studied. For example, Benaych-Georges and Nadakuditi (2012), Bura and Pfeiffer (2008), Capitaine, Donati-Martin and Féral (2009), Shabalin and Nobel (2013) focused on the asymptotic distributions of single singular value and vector when p_1, p_2 and the singular values grows proportionally. Vu (2011) discussed the squared matrix perturbed by i.i.d. Bernoulli matrix and derived an upper bound on the rotation angle of singular vectors. O'Rourke, Vu and Wang (2013) further generalized the results in Vu (2011) and proposed a trio-concentrated random matrix perturbation setting. Recently, Wang (2015) provides the ℓ_∞ distance under relatively complicated settings when matrix is perturbed by i.i.d. Gaussian noise. Candès, Sing-Long and Trzasko (2013), Donoho and Gavish (2014), Gavish and Donoho (2014) studied the algorithm for recovering X, where singular value thresholding (SVT) and hard singular value thresholding (HSVT), stated as

$$\begin{aligned}\operatorname{SVT}(Y)_\lambda &= \arg\min_X\left\{\frac{1}{2}\|Y-X\|_F^2 + \lambda\|X\|_*\right\},\\ \operatorname{HSVT}(Y)_\lambda &= \arg\min_X\left\{\frac{1}{2}\|Y-X\|_F^2 + \lambda\operatorname{rank}(X)\right\}\end{aligned} \tag{3.1}$$

were proposed. The optimal choice of thresholding level λ^* was further discussed in Donoho and Gavish (2014) and Gavish and Donoho (2014). Especially, Donoho and Gavish (2014) proves that

$$\inf_{\hat{X}} \sup_{\substack{X\in\mathbb{R}^{p_1\times p_2}\\ \operatorname{rank}(X)\le r}} \mathbb{E}\|\hat{X}-X\|_F^2 \asymp r(p_1+p_2),$$

when Z is i.i.d. standard normal random matrix. If one defines the class of rank-r matrices, $\mathcal{F}_{r,t}=\{X\in\mathbb{R}^{p_1\times p_2}:\sigma_r(X)\ge t\}$, the following upper bound for the relative error is an immediate consequence of our results

$$\sup_{X\in\mathcal{F}_{r,t}} \mathbb{E}\frac{\|\hat{X}-X\|_F^2}{\|X\|_F^2} \le \frac{C(p_1+p_2)}{t^2}\wedge 1, \tag{3.2}$$

where

$$\hat{X}=\begin{cases}\operatorname{SVT}(Y)_{\lambda^*} & \text{if } t^2\ge C(p_1+p_2),\\ 0 & \text{if } t^2<C(p_1+p_2).\end{cases}$$

In the following discussion, we assume that the entries of $Z=(Z_{ij})$ have unit variance (which can be simply achieved by normalization). To be more precise, we define the class of distributions $\mathcal{G}_\tau$ for some $\tau>0$ as follows:

$$\text{If } Z\sim\mathcal{G}_\tau \qquad \text{then } \mathbb{E}Z=0, \operatorname{Var}(Z)=1, \mathbb{E}\exp(tZ)\le\exp(\tau t), \forall t\in\mathbb{R}. \tag{3.3}$$

The distribution of the entries of Z, Z_{ij}, is assumed to satisfy

$$Z_{ij}\overset{\text{i.i.d.}}{\sim}\mathcal{G}_\tau, \qquad 1\le i\le p_1, 1\le j\le p_2.$$

Suppose $\hat{U}$ and $\hat{V}$ are respectively the first r left and right singular vectors of Y. We use $\hat{U}$ and $\hat{V}$ as the estimators of U and V, respectively. Then the perturbation bounds for singular spaces yield the following results.

THEOREM 3 (Upper bound). *Suppose $X=U\Sigma V^\intercal\in\mathbb{R}^{p_1\times p_2}$ is of rank-r. There exists constants $C>0$ that only depends on τ such that*

$$\mathbb{E}\big\|\sin\Theta(V,\hat{V})\big\|^2 \le \frac{Cp_2(\sigma_r^2(X)+p_1)}{\sigma_r^4(X)}\wedge 1,$$

$$\mathbb{E}\big\|\sin\Theta(V,\hat{V})\big\|_F^2 \le \frac{Cp_2r(\sigma_r^2(X)+p_1)}{\sigma_r^4(X)}\wedge r.$$

$$\mathbb{E}\big\|\sin\Theta(U,\hat{U})\big\|^2 \le \frac{Cp_1(\sigma_r^2(X)+p_2)}{\sigma_r^4(X)}\wedge 1,$$

$$\mathbb{E}\big\|\sin\Theta(U,\hat{U})\big\|_F^2 \le \frac{Cp_1r(\sigma_r^2(X)+p_2)}{\sigma_r^4(X)}\wedge r.$$

Theorem 3 provides a nontrivial perturbation upper bound for $\sin\Theta(V, \hat{V})$ [or $\sin\Theta(U, \hat{U})$] if there exists a constant $C_{\mathrm{gap}} > 0$ such that

$$\sigma_r^2 \geq C_{\mathrm{gap}}\big((p_1 p_2)^{\frac{1}{2}} + p_2\big)$$

[or $\sigma_r^2 \geq C_{\mathrm{gap}}((p_1 p_2)^{\frac{1}{2}} + p_1)$]. In contrast, Wedin's $\sin\Theta$ theorem requires the singular value gap $\sigma_r^2(X) \geq C_{\mathrm{gap}}(p_1 + p_2)$, which shows the power of the proposed unilateral perturbation bound.

Furthermore, the upper bounds in Theorem 3 are rate-sharp in the sense that the following matching lower bounds hold. To the best of our knowledge, this is the first result that gives different optimal rates for the left and right singular spaces under the same perturbation.

THEOREM 4 (Lower bound). *Define the following class of low-rank matrices*:

$$\mathcal{F}_{r,t} = \big\{X \in \mathbb{R}^{p_1 \times p_2} : \sigma_r(X) \geq t\big\}. \tag{3.4}$$

If $r \leq \frac{p_1}{16} \wedge \frac{p_2}{2}$, *then*

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\big\|\sin\Theta(V, \tilde{V})\big\|^2 \geq c\bigg(\frac{p_2(t^2 + p_1)}{t^4} \wedge 1\bigg),$$

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\big\|\sin\Theta(V, \tilde{V})\big\|_F^2 \geq c\bigg(\frac{p_2 r(t^2 + p_1)}{t^4} \wedge r\bigg).$$

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\big\|\sin\Theta(U, \tilde{U})\big\|^2 \geq c\bigg(\frac{p_1(t^2 + p_2)}{t^4} \wedge 1\bigg),$$

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\big\|\sin\Theta(U, \tilde{U})\big\|_F^2 \geq c\bigg(\frac{p_1 r(t^2 + p_2)}{t^4} \wedge r\bigg).$$

REMARK 4. Using similar technical arguments, we can also obtain the following lower bound for estimating the low-rank matrix X over $\mathcal{F}_{r,t}$ under a relative error loss:

$$\inf_{\tilde{X}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\frac{\|\tilde{X} - X\|_F^2}{\|X\|_F^2} \geq c\bigg(\frac{p_1 + p_2}{t^2} \wedge 1\bigg). \tag{3.5}$$

Combining equations (3.2) and (3.5) yields the minimax optimal rate for relative error in matrix denoising:

$$\inf_{\tilde{X}} \sup_{X \in \mathcal{F}_{r,t}} \mathbb{E}\frac{\|\tilde{X} - X\|_F^2}{\|X\|_F^2} \asymp c\bigg(\frac{p_1 + p_2}{t^2} \wedge 1\bigg).$$

An interesting fact is that

$$c\left(\frac{p_1+p_2}{t^2}\wedge 1\right)\asymp c\left(\frac{p_2(t^2+p_1)}{t^4}\wedge 1\right)+c\left(\frac{p_1(t^2+p_2)}{t^4}\wedge 1\right),$$

which yields directly

$$\inf_{\tilde U}\sup_{X\in\mathcal{F}_{r,t}}\mathbb{E}\left\|\sin\Theta(U,\tilde U)\right\|^2+\inf_{\tilde V}\sup_{X\in\mathcal{F}_{r,t}}\mathbb{E}\left\|\sin\Theta(V,\tilde V)\right\|^2\asymp\inf_{\tilde X}\sup_{X\in\mathcal{F}_{r,t}}\mathbb{E}\frac{\|\tilde X-X\|_F^2}{\|X\|_F}.$$

Hence, for the class of $\mathcal{F}_{r,t}$, one can stably recover X in relative Frobenius norm loss if and only if one can stably recover both U and V in spectral $\sin\Theta$ norm.

Another interesting aspect of Theorems 3 and 4 is that, when $p_1\gg p_2$, $(p_1p_2)^{\frac{1}{2}}\ll t^2\ll p_1$, there is no stable algorithm for recovery of either the left singular space U or whole matrix X in the sense that there exists uniform constant $c>0$ such that

$$\inf_{\tilde U}\sup_{X\in\mathcal{F}_{r,t}}\mathbb{E}\left\|\sin\Theta(U,\tilde U)\right\|^2\geq c,\qquad \inf_{\tilde X}\sup_{X\in\mathcal{F}_{r,t}}\mathbb{E}\frac{\|\tilde X-X\|_F^2}{\|X\|_F^2}\geq c.$$

In fact, for $X=tUV^\top\in\mathcal{F}_{r,t}$, if we simply apply SVT or HSVT algorithms with optimal choice of λ as proposed in Donoho and Gavish (2014) and Gavish and Donoho (2014), with high probability, $\mathrm{SVT}_\lambda(\hat X)=\mathrm{HSVT}_\lambda(\hat X)=0$. On the other hand, the spectral method does provide a consistent recovery of the right singular-space according to Theorem 3:

$$\mathbb{E}\left\|\sin\Theta(V,\hat V)\right\|^2\to 0.$$

This phenomenon is well demonstrated by the simulation result (Table 1) provided in Section 1.

4. High-dimensional clustering. Unsupervised learning or clustering is an ubiquitous problem in statistics and machine learning [Hastie, Tibshirani and Friedman (2009)]. The perturbation bounds given in Theorem 1 as well as the results in Theorems 3 and 4 have a direct implication in high-dimensional clustering. Suppose the locations of n points, $X=[X_1\cdots X_n]\in\mathbb{R}^{p\times n}$, which lie in a certain r-dimensional subspace $\mathcal{S}$ in $\mathbb{R}^p$, are observed with noise

$$Y_i=X_i+\varepsilon_i,\qquad i=1,\ldots,n.$$

Here, $X_i\in\mathcal{S}\subseteq\mathbb{R}^p$ are fixed coordinates, $\varepsilon_i\in\mathbb{R}^p$ are random noises. The goal is to cluster the observations Y. Let the SVD of X be given by $X=U\Sigma V^\top$, where $U\in\mathbb{O}_{p,r}$, $V\in\mathbb{O}_{n,r}$ and $\Sigma\in\mathbb{R}^{r\times r}$. When $p\gg n$, applying the standard algorithms (e.g., k-means) directly to the coordinates Y may lead to suboptimal results with expensive computational costs due to the high-dimensionality. A better approach is to first perform dimension reduction by computing the SVD of Y directly or

on its random projections, and then carry out clustering based on the first r right singular vectors $\hat{V} \in \mathbb{O}_{n,r}$; see, for example, Feldman, Schmidt and Sohler (2013) and Boutsidis et al. (2015), and the references therein. It is important to note that the left singular space U are not directly used in the clustering procedure. Thus, Theorem 3 is more suitable for the analysis of the clustering method than Wedin's $\sin\Theta$ theorem as the method main depends on the accuracy of $\hat{V}$ as an estimate of V.

Let us consider the following two-class clustering problem in more detail [see Azizyan, Singh and Wasserman (2013), Hastie, Tibshirani and Friedman (2009), Jin, Ke and Wang (2015), Jin and Wang (2016)]. Suppose $l_i \in \{-1, 1\}$, $i = 1, \dots, n$, are indicators representing the class label of the nth nodes and let $\mu \in \mathbb{R}^p$ be a fixed vector. Suppose one observes $Y = [Y_1, \dots, Y_n]$ where

$$Y_i = l_i\mu + Z_i, \qquad Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_p), 1 \le i \le n,$$

where neither the labels l_i nor the mean vector μ are observable. The goal is to cluster the data into two groups. The accuracy of any clustering algorithm is measured by the misclassification rate

$$\mathcal{M}(l, \hat{l}) := \frac{1}{n} \min_{\pi} \big|\{i : l_i \neq \pi(\hat{l}_i)\}\big|. \tag{4.1}$$

Here, π is any permutations on $\{-1, 1\}$, as any permutation of the labels $\{-1, 1\}$ does not change the clustering outcome.

In this case, EY_i is either μ or $-\mu$, which lies on a straight line. A simple PCA-based clustering method is to set

$$\hat{l} = \operatorname{sgn}(\hat{v}), \tag{4.2}$$

where $\hat{v} \in \mathbb{R}^n$ is the first right singular vector of Y. We now apply the $\sin\Theta$ upper bound in Theorem 3 to analyze the performance guarantees of this clustering method. We are particularly interested in the high-dimensional case where $p \ge n$. The case where $p < n$ can be handled similarly.

THEOREM 5. *Suppose $p \ge n$, π is any permutation on $\{-1, 1\}$. When $\|\mu\|_2 \ge C_{\text{gap}}(p/n)^{\frac{1}{4}}$ for some large constant $C_{\text{gap}} > 0$, then for some other constant $C > 0$ the misclassification rate for the PCA-based clustering method $\hat{l}$ given in (4.2) satisfies*

$$\mathbb{E}\mathcal{M}(\hat{l}, l) \le C\frac{n\|\mu\|_2^2 + p}{n\|\mu\|_2^4}.$$

It is intuitively clear that the clustering accuracy depends on the signal strength $\|\mu\|_2$. The stronger the signal, the easier to cluster. In particular, Theorem 5 requires the minimum signal strength condition $\|\mu\|_2 \ge C_{\text{gap}}(p/n)^{\frac{1}{4}}$. The following

lower bound result shows that this condition is necessary for consistent clustering: When the condition $\|\mu\|_2 \geq C_{\text{gap}}(p/n)^{\frac{1}{4}}$ does not hold, it is not possible to essentially do better than random guessing.

THEOREM 6. *Suppose $p \geq n$, there exists $c_{\text{gap}}, C_n > 0$ such that if $n \geq C_n$,*

$$\inf_{\tilde{l}} \sup_{\substack{\mu:\|\mu\|_2 \leq c_{\text{gap}}(p/n)^{\frac{1}{4}} \\ l\in\{-1,1\}^n}} \mathbb{E}\mathcal{M}(\tilde{l}, l) \geq \frac{1}{4}.$$

REMARK 5. Azizyan, Singh and Wasserman (2013) considered a similar setting when $n \geq p$, l_i's are i.i.d. Rademacher variables and derived rates of convergence for both the upper and lower bounds with a logarithmic gap between the upper and lower bounds. In contrast, with the help of the newly obtained perturbation bounds, we are able to establish the optimal misclassification rate for high-dimensional setting when $n \leq p$.

Moreover, Jin and Wang (2016) and Jin, Ke and Wang (2015) considered the sparse and highly structured setting, where the contrast mean vector μ is assumed to be sparse and the nonzero coordinates are all equal. Their method is based on feature selection and PCA. Our setting is close to the "less sparse/weak signal" case in Jin, Ke and Wang (2015). In this case, they introduced a simple aggregation method with

$$\hat{l}^{(\text{sa})} = \text{sgn}(X\hat{\mu}),$$

where $\hat{\mu} = \arg\max_{\mu\in\{-1,0,1\}^p} \|X\mu\|_q$ for some $q > 0$. The statistical limit, that is, the necessary condition for obtaining correct labels for most of the points, is $\|\mu\|_2 > C$ in their setting, which is smaller than the boundary $\|\mu\|_2 > C(p/n)^{\frac{1}{4}}$ in Theorem 5. As shown in Theorem 6, the bound $\|\mu\|_2 > C(p/n)^{\frac{1}{4}}$ is necessary. The reason for this difference is that they focused on highly structured contrast mean vector μ which only takes two values $\{0, \nu\}$. In contrast, we considered the general $\mu \in \mathbb{R}^p$, which leads to stronger condition and larger statistical limit. Moreover, the simple aggregation algorithm is computational difficult for a general signal μ, thus the PCA-based method considered in this paper is preferred under the general dense μ setting.

5. Canonical correlation analysis. In this section, we consider an application of the perturbation bounds given in Theorem 1 to the canonical correlation analysis (CCA), which is one of the most important tools in multivariate analysis in exploring the relationship between two sets of variables [Anderson (2003), Gao, Ma and Zhou (2014), Gao et al. (2015), Hotelling (1936), Ma and Li (2016), Witten, Tibshirani and Hastie (2009)]. Given two random vectors X and Y with a certain joint distribution, the CCA first looks for the pair of vectors

$\alpha^{(1)} \in \mathbb{R}^{p_1}, \beta^{(2)} \in \mathbb{R}^{p_2}$ that maximize $\operatorname{corr}((\alpha^{(1)})^\intercal X, (\beta^{(1)})^\intercal Y)$. After obtaining the first pair of canonical directions, one can further obtain the second pair $\alpha^{(2)} \in \mathbb{R}^{p_1}, \beta^{(2)} \in \mathbb{R}^{p_2}$ such that $\operatorname{Cov}((\alpha^{(1)})^\intercal X, (\alpha^{(2)})^\intercal X) = \operatorname{Cov}((\beta^{(1)})^\intercal Y, (\beta^{(2)})^\intercal Y) = 0$, and $\operatorname{Corr}((\alpha^{(2)})^\intercal X, (\beta^{(2)})^\intercal Y)$ is maximized. The higher order canonical directions can be obtained by repeating this process. If (X, Y) is further assumed to have joint covariance, say

$$\operatorname{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix},$$

the population canonical correlation directions can be inductively defined as the following optimization problem. For $k = 1, 2, \ldots,$

$$\begin{aligned} (\alpha^{(k)}, \beta^{(k)}) = \underset{a \in \mathbb{R}^{p_1}, b \in \mathbb{R}^{p_2}}{\arg\max} \quad & a^\intercal \Sigma_{XY} b, \\ \text{subject to} \quad & a^\intercal \Sigma_X a = b^\intercal \Sigma_Y b = 1, \\ & a^\intercal \Sigma_X \alpha^{(l)} = b^\intercal \Sigma_Y \beta^{(l)} = 0, \qquad \forall 1 \le l \le k-1. \end{aligned}$$

A more explicit form for the canonical correlation directions is given in Hotelling (1936): $(\Sigma_X^{\frac{1}{2}} \alpha^{(k)}, \Sigma_Y^{\frac{1}{2}} \beta^{(k)})$ is the kth pair of singular vectors of $\Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}}$. We combine the leading r population canonical correlation directions and write

$$A = [\alpha^{(1)} \cdots \alpha^{(r)}], \qquad B = [\beta^{(1)} \cdots \beta^{(r)}].$$

Suppose one observes i.i.d. samples $(X_i^\intercal, Y_i^\intercal)^\intercal \sim N(0, \Sigma)$. Then the sample covariance and cross-covariance for X and Y can be calculated as

$$\hat{\Sigma}_X = \frac{1}{n} \sum_{i=1}^n X_i X_i^\intercal, \qquad \hat{\Sigma}_Y = \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\intercal, \qquad \hat{\Sigma}_{XY} = \frac{1}{n} \sum_{i=1}^n X_i Y_i^\intercal.$$

The standard approach to estimate the canonical correlation directions $\alpha^{(k)}, \beta^{(k)}$ is via the SVD of $\hat{\Sigma}_X^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_Y^{-\frac{1}{2}}$

$$\hat{\Sigma}_X^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_Y^{-\frac{1}{2}} = \hat{U} \hat{S} \hat{\mathrm{V}} = \sum_{k=1}^{p_1 \wedge p_2} \hat{U}_{[:,k]} \hat{S}_{kk} \hat{U}_{[:,k]}^\intercal.$$

Then the leading r sample canonical correlation directions can be calculated as

$$\begin{aligned} \hat{A} &= \hat{\Sigma}_X^{-\frac{1}{2}} \hat{U}_{[:,1:r]}, \qquad \hat{A} = [\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \ldots, \hat{\alpha}^{(r)}], \\ \hat{B} &= \hat{\Sigma}_Y^{-\frac{1}{2}} \hat{V}_{[:,1:r]}, \qquad \hat{B} = [\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \ldots, \hat{\beta}^{(r)}]. \end{aligned} \tag{5.1}$$

$\hat{A}, \hat{B}$ are consistent estimators for the first r left and right canonical directions in the classical fixed dimension case.

Let $X^* \in \mathbb{R}^{p_1}$ be an independent copy of the original sample X, we define the following two losses to measure the accuracy of the estimator of the canonical correlation directions

$$L_{\mathrm{sp}}(\hat{A}, A) = \max_{\substack{O \in \mathbb{O}_r \\ v \in \mathbb{R}^r, \|v\|_2 = 1}} \mathbb{E}_{X^*}\big((\hat{A}Ov)^\intercal X^* - (Av)^\intercal X^*\big)^2, \tag{5.2}$$

$$L_F(\hat{A}, A) = \max_{O \in \mathbb{O}_r} \mathbb{E}_{X^*}\big\|(\hat{A}O)^\intercal X^* - A^\intercal X^*\big\|_2^2. \tag{5.3}$$

These two losses quantify how well the estimator $(\hat{A}O)^\intercal X^*$ can predict the values of the canonical variables $A^\intercal X^*$, where $O \in \mathbb{O}_r$ is a rotation matrix as the objects of interest here are the directions.

The following theorem gives the upper bound for one side of the canonical correlation directions. The main technical tool is the perturbation bounds given in Section 2.

THEOREM 7. *Suppose* $(X_i, Y_i) \sim N(0, \Sigma), i = 1, \ldots, n$, *where* $S = \Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}}$ *is of rank-r. Suppose* $\hat{A} \in \mathbb{R}^{p_1 \times r}$ *is given by* (5.1). *Then there exist uniform constants* $C_{\mathrm{gap}}, C, c > 0$ *such that whenever* $\sigma_r(S)^2 \geq \frac{C_{\mathrm{gap}}((p_1 p_2)^{\frac{1}{2}} + p_1 + p_2^{3/2} n^{-\frac{1}{2}})}{n}$

$$\mathbb{P}\bigg(L_{\mathrm{sp}}(\hat{A}, A) \leq \frac{C p_1 (n\sigma_r^2(S) + p_2)}{n^2 \sigma_r^4(S)}\bigg) \geq 1 - C\exp(-cp_1 \wedge p_2),$$

$$\mathbb{P}\bigg(L_F(\hat{A}, A) \leq \frac{C p_1 r (n\sigma_r^2(S) + p_2)}{n^2 \sigma_r^4(S)}\bigg) \geq 1 - C\exp(-cp_1 \wedge p_2).$$

The results for $\hat{B}$ *can be stated similarly.*

REMARK 6. Chen et al. (2013) and Gao, Ma and Zhou (2014), Gao et al. (2015) considered sparse CCA, where the canonical correlation directions A and B are assumed to be jointly sparse. In particular, Chen et al. (2013) and Gao et al. (2015) proposed estimators under different settings and provided a unified rate-optimal bound for jointly estimating left and right canonical correlations. Gao, Ma and Zhou (2014) proposed another computationally feasible estimators $\hat{A}^*$ and $\hat{B}^*$ and provided a minimax rate-optimal bound for $L_F(\hat{A}^*, A)$ under regularity conditions that can also be used to prove the consistency of $\hat{B}^*$.

Now consider the setting where $p_2 \gg p_1$, $\frac{p_2}{n} \gg \sigma_r^2(S) = t^2 \gg \frac{(p_1 p_2)^{\frac{1}{2}}}{n}$. The lower bound result in Theorem 3.3 by Gao, Ma and Zhou (2014) implies that there is no consistent estimator for the right canonical correlation directions B. While Theorem 7 given above shows that the left sample canonical correlation directions $\hat{A}$ are a consistent estimator of A. This interesting phenomena again shows the merit of our proposed unilateral perturbation bound.

It is also interesting to develop the lower bounds for $\hat{A}$ and $\hat{B}$. The best known result, given in Theorem 3.2 in Gao, Ma and Zhou (2014), is the following two-sided lower bound for both $\hat{A}$ and $\hat{B}$ in Frobenius norm loss:

$$\inf_{\hat{A},\hat{B}} \sup_{A,B} \mathbb{P}\left\{\max\{L_F(\hat{A}, A), L_F(\hat{B}, B)\} \geq c\left(\frac{r(p_1+p_2)}{n\sigma_{\min}^2(S)} \wedge 1\right)\right\} \geq 0.8.$$

Establishing the matching one-sided lower bound for Theorem 7 is technical challenging. We leave it for future research.

6. Simulations. In this section, we carry out numerical experiments to further illustrate the advantages of the separate bounds for the left and right singular subspaces over the uniform bounds. As mentioned earlier, in a range of cases, especially when the numbers of rows and columns of the matrix differ significantly, it is even possible that the singular space on one side can be stably recovered, while the other side cannot. To illustrates this point, we specifically perform simulation studies in matrix denoising, high-dimensional clustering and canonical correlation analysis.

We first consider the matrix denoising model discussed in Section 3. Let $X = tUV^\intercal \in \mathbb{R}^{p_1 \times p_2}$, where $t \in \mathbb{R}$, U and V are $p_1 \times r$ and $p_2 \times r$ random uniform orthonormal columns with respect to the Haar measure. Let the perturbation $Z = (Z_{ij})_{p_1 \times p_2}$ be randomly generated with $Z_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$. We calculate the SVD of $X + Z$ and form the first r left and right singular vectors as $\hat{U}$ and $\hat{V}$. The average losses in Frobenius and spectral $\sin\Theta$ distances for both the left and right singular space estimates with 1,000 repetitions are given in Table 1 for various values of (p_1, p_2, r, t). It can be easily seen from this experiment that the left and right singular perturbation bounds behave very distinctly when $p_1 \gg p_2$.

We then consider the high-dimensional clustering model studied in Section 4. Let $\tilde{\mu} \sim N(0, I_p)$ and $\mu = t(p/n)^{1/4} \cdot \tilde{\mu}/\|\tilde{\mu}\|_2 \in \mathbb{R}^p$, where $t = \|\mu\|_2$ essentially

TABLE 1
Average losses in Frobenius and spectral $\sin\Theta$ *distances for both the left and right singular space changes after Gaussian noise perturbations*

(p_1, p_2, r, t)	$\|sin\Theta(\hat{U}, U)\|^2$	$\|\sin\Theta(\hat{V}, V)\|^2$	$\|\sin\Theta(\hat{U}, U)\|_F^2$	$\|\sin\Theta(\hat{V}, V)\|_F^2$
(100, 10, 2, 15)	0.3512	0.0669	0.6252	0.0934
(100, 10, 2, 30)	0.1120	0.0139	0.1984	0.0196
(100, 20, 5, 20)	0.2711	0.0930	0.9993	0.2347
(100, 20, 5, 40)	0.0770	0.0195	0.2835	0.0508
(1000, 20, 5, 30)	0.5838	0.0699	2.6693	0.1786
(1000, 20, 10, 100)	0.1060	0.0036	0.9007	0.0109
(1000, 200, 10, 50)	0.3456	0.0797	2.9430	0.4863
(1000, 200, 50, 100)	0.1289	0.0205	4.3614	0.2731

TABLE 2
Average misclassification rate for different settings

	n					
(p, t, ρ)	**5**	**10**	**20**	**50**	**100**	**200**
(100, 1, 1/2)	0.2100	0.1485	0.0690	0.0494	0.0440	0.0333
(100, 1, 3/4)	0.2150	0.1590	0.0680	0.0468	0.0422	0.0290
(100, 3, 1/2)	0.0019	0.0005	0.0000	0.0000	0.0000	0.0000
(100, 3, 3/4)	0.0020	0.0005	0.0000	0.0000	0.0000	0.0000
(1000, 1, 1/2)	0.3260	0.3510	0.3594	0.2855	0.2691	0.1364
(1000, 1, 3/4)	0.3610	0.3610	0.3462	0.3057	0.2696	0.1410
(1000, 3, 1/2)	0.1370	0.0485	0.0066	0.0019	0.0013	0.0003
(1000, 3, 3/4)	0.1160	0.0425	0.0046	0.0019	0.0018	0.0006

represents the signal strength. The group label $l \in \mathbb{R}^n$ is randomly generated as

$$l_i \overset{\text{i.i.d.}}{\sim} \begin{cases} 1 & \text{with probability } \rho, \\ -1 & \text{with probability } 1-\rho. \end{cases}$$

Based on n i.i.d. observations: $Y_i = l_i\mu + Z_i$, $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_p)$, $i = 1, \ldots, n$, we apply the proposed estimator (4.2) to estimate l. The results for different values of (n, p, t, ρ) are provided in Table 2. It can be seen that the numerical results match our theoretical analysis—the proposed $\hat{l}$ achieves good performance roughly when $t \geq C(p/n)^{1/4}$.

We finally investigate the numerical performance of canonical correlation analysis particularly when the dimensions of two samples differ significantly. Suppose $\Sigma_X = I_{p_1} + \frac{1}{2\|Z_{p_1}+Z_{p_1}^\intercal\|}(Z_{p_1} + Z_{p_1}^\intercal)$, $\Sigma_Y = I_{p_2} + \frac{1}{2\|Z_{p_2}+Z_{p_2}^\intercal\|}(Z_{p_2} + Z_{p_2}^\intercal)$, $\Sigma_{XY} = \Sigma_X^{1/2} \cdot (tUV^\intercal)\Sigma_Y^{1/2}$, where Z_{p_1} and Z_{p_2} are i.i.d. Gaussian matrices; $U \in \mathbb{O}_{p_1,r}$, $V \in \mathbb{O}_{p_2,r}$ are random orthogonal matrices. With n pairs of observations,

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N(0, \Sigma), \qquad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\intercal & \Sigma_Y \end{bmatrix}, i = 1, \ldots, n,$$

we apply the procedure discussed in Section 5 to obtain $\hat{A}$ and $\hat{B}$, that is, the estimates for left and right canonical correlation directions. Since the exact losses in $L_{\text{sp}}(\cdot,\cdot)$, $L_F(\cdot,\cdot)$ metrics (5.2) involves difficult optimization, we instead measure the losses in

$$\|\sin\Theta(\hat{U}, U)\|, \qquad \|\sin\Theta(\hat{U}, U)\|_F, \qquad \|\sin\Theta(\hat{V}, V)\|$$

and

$$\|\sin\Theta(\hat{V}, V)\|_F.$$

TABLE 3
Average losses in $L_{\mathrm{sp}}(\cdot,\cdot)$ and $L_F(\cdot,\cdot)$ metrics for the left and right canonical directions

(p_1, p_2, r, t)	$\|\sin\Theta(\hat{U}_S, U_S)\|$	$\|\sin\Theta(\hat{U}_S, U_S)\|_F$	$\|\sin\Theta(\hat{V}_S, V_S)\|$	$\|\sin\Theta(\hat{V}_S, V_S)\|_F$
(30, 10, 100, 0.8)	0.3194	0.6609	0.1571	0.2530
(30, 10, 200, 0.5)	0.5348	1.1111	0.3343	0.5256
(100, 10, 200, 0.8)	0.4103	1.0145	0.1120	0.1825
(100, 10, 500, 0.5)	0.5183	1.2821	0.1614	0.2606
(200, 20, 500, 0.8)	0.3239	0.8428	0.0746	0.1442
(200, 20, 800, 0.5)	0.5834	1.5155	0.2423	0.4605
(500, 50, 1000, 0.8)	0.3875	1.0515	0.1091	0.2472
(500, 50, 2000, 0.5)	0.5677	1.5467	0.2216	0.4910

Here, $U, V, \hat{U}, \hat{V}$ are the first r left and right singular vectors of $\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}$ and $\hat{\Sigma}_X^{-1/2}\hat{\Sigma}_{XY}\hat{\Sigma}_Y^{-1/2}$, respectively. It is shown in Step 1 of the proof for Theorem 7 that these measures are equivalent to L_{sp} and L_F. The results under various choices of (p_1, p_2, n, t) are collected in Table 3. It can be easily seen that the performance of the right canonical direction estimation is much better than the left ones when p_1 is much larger than p_2, which is consistent with the theoretical results in Theorem 7 and illustrates the power of the newly proposed perturbation bound results.

7. Discussions. We have established in the present paper new and rate-optimal perturbation bounds, measured in both spectral and Frobenius $\sin\Theta$ distances, for the left and right singular subspaces separately. These perturbation bounds are widely applicable to the analysis of many high-dimensional problems. In particular, we applied the perturbation bounds to study three important problems in high-dimensional statistics: low-rank matrix denoising and singular space estimation, high-dimensional clustering and CCA. As mentioned in the Introduction, in addition to these problems and possible extensions discussed in the previous sections, the obtained perturbation bounds can be used in a range of other applications including *community detection in bipartite networks, multidimensional scaling, cross-covariance matrix estimation and singular space estimation for matrix completion*. We briefly discuss these problems here.

An interesting application of the perturbation bounds given in Section 2 is *community detection in bipartite graphs*. Community detection in networks has attracted much recent attention. The focus of the current community detection literature has been mainly on unipartite graph (i.e., there is only one type of nodes). However, in some applications, the nodes can be divided into different types and only the interactions between the different types of nodes are available or of interest, such as people versus committees, Facebook users versus public pages [see

Alzahrani and Horadam (2016), Melamed (2014)]. The observations on the connectivity of the network between two types of nodes can be described by an adjacency matrix A, where $A_{ij} = 1$ if the ith Type 1 node and jth Type 2 node are connected, and $A_{ij} = 0$ otherwise. The spectral method is one of the most commonly used approaches in the literature with theoretical guarantees [Lei and Rinaldo (2015), Rohe, Chatterjee and Yu (2011)]. In a bipartite network, the left and right singular subspaces could behave very differently from each other. Our perturbation bounds can be used for community detection in bipartite graph and potentially lead to sharper results in some settings.

Another possible application lies in *multidimensional scaling* (*MDS*) with distance matrix between two sets of points. MDS is a popular method of visualizing the data points embedded in low-dimensional space based on the distance matrices [Borg and Groenen (2005)]. Traditionally MDS deals with unipartite distance matrix, where all distances between any pairs of points are observed. In some applications, the data points are from two groups and one is only able to observe its biparitite distance matrix formed by the pairwise distances between points from different groups. As the SVD is a commonly used technique for dimension reduction in MDS, the perturbation bounds developed in this paper can be potentially used for the analysis of MDS with bipartite distance matrix.

In some applications, the *cross-covariance matrix*, not the overall covariance matrix, is of particular interest. Cai et al. (2015a) considered multiple testing of cross-covariances in the context of the phenome-wide association studies (PheWAS). Suppose $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$ are jointly distributed with covariance matrix Σ. Given n i.i.d. samples (X_i, Y_i), $i = 1, \ldots, n$, from the joint distribution, one wishes to make statistical inference for the cross-covariance matrix Σ_{XY}. If Σ_{XY} has low-rank structure, the perturbation bounds established in Section 2 could be potentially applied to make statistical inference for Σ_{XY}.

Matrix completion, whose central goal is to recover a large low-rank matrix based on a limited number of observable entries, has been widely studied in the last decade. Among various methods for matrix completion, spectral method is fast, easy to implement and achieves good performance [Chatterjee (2014), Cho, Kim and Rohe (2015), Keshavan, Montanari and Oh (2010)]. The new perturbation bounds can be potentially used for singular space estimation under the matrix completion setting to yield better results.

In addition to the aforementioned problems, *high-dimensional clustering with correlated features* is an important extension of the problem of clustering with independent features considered in the present paper. Specifically, based on n observations $Y_i = l_i\mu + Z_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, where $Z_i \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, one aims to recover the unknown labels $\{l_i\}_{i=1}^n$. When Σ is known or can be well estimated, one can transform $\tilde{Y}_i = \Sigma^{-1/2} Y_i$, $i = 1, \ldots, n$ and perform the spectral method on $\{\tilde{Y}_i\}_{i=1}^n$. It would be an interesting and challenging problem to consider the general setting where Σ is unknown. We leave this for future research.

8. Proofs. We prove the main results in Sections 2, 3 and 4 in this section. The proofs for CCA and the additional technical results are given in the supplementary material [Cai and Zhang (2017)].

8.1. *Proofs of general unilateral perturbation bounds.* Some technical tools are needed to prove Theorems 1, 2 and Proposition 1. In particular, we need a few useful properties of $\sin\Theta$ distances given below in Lemma 1. Specifically, Lemma 1 provides some more convenient expressions than the definitions for the $\sin\Theta$ distances. It also shows that they are indeed distances as they satisfy triangle inequality. Some other widely used metrics for orthogonal spaces, including

$$D_{\mathrm{sp}}(\hat{V}, V) = \inf_{O\in\mathbb{O}_r} \|\hat{V} - VO\|, \qquad D_F(\hat{V}, V) = \inf_{O\in\mathbb{O}_r} \|\hat{V} - VO\|_F, \tag{8.1}$$

$$\|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\|, \qquad \|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\|_F \tag{8.2}$$

are shown to be equivalent to the $\sin\Theta$ distances.

LEMMA 1 (Properties of the $\sin\Theta$ distances). *The following properties hold for the* $\sin\Theta$ *distances*:

1. (*Equivalent expressions.*) *Suppose* $V, \hat{V} \in \mathbb{O}_{p,r}$. *If* $V_\perp$ *is an orthogonal extension of* V, *namely* $[V\ V_\perp] \in \mathbb{O}_p$, *we have the following equivalent forms for* $\|\sin\Theta(\hat{V}, V)\|$ *and* $\|\sin\Theta(\hat{V}, V)\|_F$:

$$\|\sin\Theta(\hat{V}, V)\| = \sqrt{1 - \sigma^2_{\min}(\hat{V}^{\mathsf{T}}V)} = \|\hat{V}^{\mathsf{T}}V_\perp\|, \tag{8.3}$$

$$\|\sin\Theta(\hat{V}, V)\|_F = \sqrt{r - \|V^{\mathsf{T}}\hat{V}\|_F^2} = \|\hat{V}^{\mathsf{T}}V_\perp\|_F. \tag{8.4}$$

2. (*Triangle inequality.*) *For any* $V_1, V_2, V_3 \in \mathbb{O}_{p,r}$,

$$\|\sin\Theta(V_2, V_3)\| \le \|\sin\Theta(V_1, V_2)\| + \|\sin\Theta(V_1, V_3)\|, \tag{8.5}$$

$$\|\sin\Theta(V_2, V_3)\|_F \le \|\sin\Theta(V_1, V_2)\|_F + \|\sin\Theta(V_1, V_3)\|_F. \tag{8.6}$$

3. (*Equivalence with other metrics.*) *The metrics defined as* (8.1) *and* (8.2) *are equivalent to* $\sin\Theta$ *distances as the following inequalities hold*:

$$\begin{aligned}
\|\sin\Theta(\hat{V}, V)\| &\le D_{\mathrm{sp}}(\hat{V}, V) \le \sqrt{2}\|\sin\Theta(\hat{V}, V)\|,\\
\|\sin\Theta(\hat{V}, V)\|_F &\le D_F(\hat{V}, V) \le \sqrt{2}\|\sin\Theta(\hat{V}, V)\|_F,\\
\|\sin\Theta(\hat{V}, V)\| &\le \|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\| \le 2\|\sin\Theta(\hat{V}, V)\|,\\
\|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\|_F &= \sqrt{2}\|\sin\Theta(\hat{V}, V)\|_F.
\end{aligned}$$

PROOF OF PROPOSITION 1. First, we can rotate the right singular space by right multiplying the whole matrices $A, V^\intercal, W^\intercal$ by $[W\ W_\perp]$ without changing the singular values and left singular vectors. Thus, without loss of generality, we assume that

$$[W\ W_\perp] = I_{p_2}.$$

Next, we further calculate the SVD: $AW = A_{[:,1:r]} := \bar{U}\bar{\Sigma}\bar{V}^\intercal$, where $\bar{U} \in \mathbb{O}_{p_1,r}$, $\bar{\Sigma} \in \mathbb{R}^{r\times r}$, $\bar{V} \in \mathbb{O}_r$, and rotate the left singular space by left multiplying the whole matrix A by $[\bar{U}\ \bar{U}_\perp]^\intercal$, then rotate the right singular space by right multiplying $A_{[:,1:r]}$ by $\bar{V}$. After this rotation, the singular structure of A, AW are unchanged. Again without loss of generality, we can assume that $[\bar{U}\ \bar{U}_\perp]^\intercal = I_{p_1}$, $\bar{V} = I_r$. After these two steps of rotations, the formation of A is much simplified,

$$A = \begin{array}{c} \\ r \\ p_1 - r \end{array} \begin{array}{c} \begin{array}{ccc} r & & p_2 - r \end{array} \\ \begin{bmatrix} \sigma_1(AW) & & \\ & \ddots & \\ & & \sigma_r(AW) \\ 0 & & \end{bmatrix} \begin{array}{c} \\ \\ \bar{U}^\intercal AW_\perp \\ \\ \bar{U}_\perp^\intercal AW_\perp \end{array} \end{array}, \tag{8.7}$$

while the problem we are considering is still without loss of generality. For convenience, denote

$$(\bar{U}^\intercal AW_\perp)^\intercal = \left[y^{(1)}\ y^{(2)} \cdots y^{(r)}\right], \qquad y^{(1)}, \ldots, y^{(r)} \in \mathbb{R}^{p_2 - r}. \tag{8.8}$$

We can further compute that

$$A^\intercal A = \begin{array}{c} \\ r \\ p_2 - r \end{array} \begin{array}{c} \begin{array}{ccc} r & & p_2 - r \end{array} \\ \begin{bmatrix} \sigma_1^2(AW) & & \\ & \ddots & \\ & & \sigma_r^2(AW) \\ \sigma_1(AW)y^{(1)} & \cdots & \sigma_r(AW)y^{(r)} \end{bmatrix} \begin{array}{c} \sigma_1(AW)y^{(1)\intercal} \\ \vdots \\ \sigma_r(AW)y^{(r)\intercal} \\ (AW_\perp)^\intercal AW_\perp \end{array} \end{array}. \tag{8.9}$$

By basic theory in algebra, the ith eigenvalue of $A^\intercal A$ is equal to $\sigma_i^2(A)$, and the ith eigenvector of $A^\intercal A$ is equal to the ith right singular vector of A (up-to-sign). Suppose the singular vectors of A are $\tilde{V} = [v^{(1)}, v^{(2)}, \ldots, v^{(p_2)}]$, where the singular values can be further decomposed into two parts as

$$v^{(k)} = \begin{array}{c} r \\ p_2 - r \end{array} \begin{bmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{bmatrix}, \qquad \text{or equivalently,}$$
$$\alpha^{(k)} = W^\intercal v^{(k)}, \qquad \beta^{(k)} = W_\perp^\intercal v^{(k)}. \tag{8.10}$$

By observing the ith entry of $A^\intercal A v^{(k)} = \sigma_k^2(A) v^{(k)}$, we know for $1 \le i \le r$, $r+1 \le k \le p_2$,

$$\begin{aligned} &\big(\sigma_i^2(AW) - \sigma_k^2(A)\big)\alpha_i^{(k)} + \sigma_i(AW) y^{(i)\intercal}\beta^{(k)} = 0, \\ &\Rightarrow \quad \alpha_i^{(k)} = \frac{-\sigma_i(AW)}{\sigma_i^2(AW) - \sigma_k^2(A)} y^{(i)\intercal}\beta^{(k)}. \end{aligned} \tag{8.11}$$

Recall the assumption that

$$\sigma_1(AW) \ge \cdots \ge \sigma_r(AW) > \sigma_{r+1}(A) \ge \cdots \sigma_{p_2}(A) \ge 0. \tag{8.12}$$

Also $\frac{x}{x^2-y^2} = \frac{1}{x-y^2/x}$ is a decreasing function for x and a increasing function for y when $x > y \ge 0$, so

$$\begin{aligned} &\frac{\sigma_i(AW)}{\sigma_i^2(AW) - \sigma_k^2(A)} \\ &\qquad \le \frac{\sigma_r(AW)}{\sigma_r^2(AW) - \sigma_{r+1}^2(A)}, \qquad 1 \le i \le r, r+1 \le k \le p_2. \end{aligned} \tag{8.13}$$

Since $[\beta^{(r+1)} \ \cdots \ \beta^{(p_2)}]$ is the submatrix of the orthogonal matrix V,

$$\big\|[\beta^{(r+1)} \cdots \ \beta^{(p_2)}]\big\| \le 1. \tag{8.14}$$

Now we can give an upper bound for the Frobenius norm of $[\alpha^{(r+1)} \cdots \ \alpha^{(p_2)}]$

$$\begin{aligned} &\big\|[\alpha^{(r+1)} \cdots \alpha^{(p_2)}]\big\|_F^2 \\ &\quad = \sum_{i=1}^{r} \sum_{k=r+1}^{p_2} \big(\alpha_i^{(k)}\big)^2 \\ &\quad \overset{(8.13)}{\le} \frac{\sigma_r^2(AW)}{(\sigma_r^2(AW) - \sigma_{r+1}^2(A))^2} \sum_{i=1}^{r} \sum_{k=r+1}^{p_2} \big(y^{(i)\intercal}\beta^{(k)}\big)^2 \\ &\quad \le \frac{\sigma_r^2(AW)}{(\sigma_r^2(AW) - \sigma_{r+1}^2(A))^2} \big\|[y_1 \cdots y_r]^\intercal\big\|_F^2 \big\|[\beta^{(r+1)} \cdots \ \beta^{(p_2)}]\big\|^2 \\ &\quad \overset{(8.8)(8.14)}{\le} \frac{\sigma_r^2(AW)}{(\sigma_r^2(AW) - \sigma_{r+1}^2(A))^2} \big\|\bar{U}^\intercal A W_\perp\big\|_F^2. \end{aligned}$$

It is more complicated to give a upper bound for the spectral norm of $[\alpha^{(r+1)} \cdots \alpha^{(p_2)}]$. Suppose $s = (s_{r+1}, \ldots, s_{p_2}) \in \mathbb{R}^{p_2 - r}$ is any vector with $\|s\|_2 = 1$. Based

on (8.11),

$$\begin{aligned}\sum_{k=r+1}^{p_2} s_k\alpha_i^{(k)} &= \sum_{k=r+1}^{p_2} \frac{-s_k\sigma_i(AW)y^{(i)\intercal}\beta^{(k)}}{\sigma_i^2(AW)-\sigma_k^2(A)}\\ &= \sum_{k=r+1}^{p_2} \frac{-s_k}{\sigma_i(AW)}\frac{1}{1-\sigma_k^2(A)/\sigma_i(AW)^2}y^{(i)\intercal}\beta^{(k)}\\ &\overset{(8.12)}{=} \sum_{k=r+1}^{p_2}\sum_{l=0}^{\infty}\frac{-s_k\sigma_k^{2l}(A)}{\sigma_i^{2l+1}(AW)}y^{(i)\intercal}\beta^{(k)}\\ &= \sum_{l=0}^{\infty}\frac{-y^{(i)\intercal}}{\sigma_i^{2l+1}(AW)}\left(\sum_{k=r+1}^{p_2} s_k\sigma_k^{2l}(A)\beta^{(k)}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\sum_{k=r+1}^{p_2} s_k\alpha^{(k)}\right\|_2 &\le \sum_{l=0}^{\infty}\left\|\begin{bmatrix} y^{(1)\intercal}/\sigma_1^{2l+1}(AW)\\ \vdots\\ y^{(r)\intercal}/\sigma_r^{2l+1}(AW)\end{bmatrix}\right.\\ &\qquad\left.\cdot\left(\sum_{k=r+1}^{p_2} s_k\sigma_k^{2l}(A)\beta^{(k)}\right)\right\|_2\\ &\le \sum_{l=0}^{\infty}\frac{\|[y^{(1)}\ y^{(2)}\cdots y^{(r)}]\|}{\sigma_r^{2l+1}(AW)}\\ &\qquad\cdot\|[\beta^{(r+1)}\ \beta^{(r+2)}\cdots\beta^{(p_2)}]\|\\ &\qquad\cdot\|(s_{r+1}\sigma_{r+1}^{2l}(A),\ldots,s_{p_2}\sigma_{p_2}^{2l}(A))\|_2\\ &\overset{(8.8)(8.14)(8.12)}{\le} \sum_{l=0}^{\infty}\frac{\|\tilde{U}^{\intercal}AW_{\perp}\|}{\sigma_r^{2l+1}(AW)}\cdot\sigma_{r+1}^{2l}(A)\|s\|_2\\ &= \frac{\|\tilde{U}^{\intercal}AW_{\perp}\|\sigma_r(AW)}{\sigma_r^2(AW)-\sigma_{r+1}^2(A)},\end{aligned}$$

which implies

$$\|[\alpha^{(r+1)}\ \cdots\ \alpha^{(p_2)}]\| \le \frac{\|\tilde{U}^{\intercal}AW_{\perp}\|\sigma_r(AW)}{\sigma_r^2(AW)-\sigma_{r+1}^2(A)}.$$

Note the definition of $\alpha^{(i)}$ in (8.10), we know

$$[\alpha^{(r+1)}\alpha^{(r+2)}\cdots\alpha^{(p_2)}] = \tilde{V}_{[1:r,(r+1):p_2]} = (V_{\perp})_{[1:r,:]}.$$

Thus,

$$
\begin{aligned}
\|\sin\Theta(V,W)\| &\overset{(8.3)}{=} \|W^\top V_\perp\| \\
&= \|[\alpha^{(r+1)}\cdots\alpha^{(p_2)}]\| \\
&\le \frac{\|\tilde{U}^\top A W_\perp\|\sigma_r(AW)}{\sigma_r^2(AW)-\sigma_{r+1}^2(A)}, \\
\|\sin\Theta(V,W)\|_F^2 &\overset{(8.4)}{=} \|W^\top V_\perp\|_F^2 \\
&= \|[\alpha^{(r+1)}\ \cdots\ \alpha^{(p_2)}]\|_F^2 \\
&\le \frac{\|\tilde{U}^\top A W_\perp\|_F^2\sigma_r^2(AW)}{(\sigma_r^2(AW)-\sigma_{r+1}^2(A))^2}.
\end{aligned} \tag{8.15}
$$

Finally, since $\bar{U}$ is the left singular vectors of AW,

$$
\|\bar{U}^\top A W_\perp\| = \|\mathbb{P}_{(AW)} A W_\perp\|, \qquad \|\bar{U}^\top A W_\perp\|_F = \|\mathbb{P}_{(AW)} A W_\perp\|. \tag{8.16}
$$

The upper bounds 1 in (2.11) and $\sqrt{r}$ on (2.12) are trivial. Therefore, we have finished the proof of Proposition 1. □

PROOF OF THEOREM 1. Before proving this theorem, we introduce the following lemma on the inequalities of the singular values in the perturbed matrix.

LEMMA 2. *Suppose* $X \in \mathbb{R}^{p\times n}$, $Y \in \mathbb{R}^{p\times n}$, $\operatorname{rank}(X) = a$, $\operatorname{rank}(Y) = b$:

1. $\sigma_{a+b+1-r}(X+Y) \le \min(\sigma_{a+1-r}(X), \sigma_{b+1-r}(Y))$ *for* $r \ge 1$;
2. *if we further have* $X^\top Y = 0$ *or* $XY^\top = 0$, *we must have* $a + b \le n \wedge p$, *and*

$$
\sigma_r^2(X+Y) \ge \max\bigl(\sigma_r^2(X), \sigma_r^2(Y)\bigr)
$$

for any $r \ge 1$. *Also*,

$$
\sigma_1^2(X+Y) \le \sigma_1^2(X) + \sigma_1^2(Y).
$$

The proof of Lemma 2 is provided in the supplementary materials [Cai and Zhang (2017)]. Applying Lemma 2, we get

$$
\begin{aligned}
\sigma_{\min}^2(\hat{X}V) = \sigma_r^2(\hat{X}V) &= \sigma_r^2\bigl(UU^\top\hat{X}V + U_\perp U_\perp^\top \hat{X}V\bigr) \\
&\ge \sigma_r^2\bigl(UU^\top\hat{X}V\bigr) = \alpha^2 \qquad \text{(by Lemma 2, part 2).}
\end{aligned} \tag{8.17}
$$

Since U, V have r columns, $\operatorname{rank}(\hat{X}VV^{\intercal}), \operatorname{rank}(UU^{\intercal}\hat{X}) \le r$. Also since $\hat{X} = U_{\perp}U_{\perp}^{\intercal}\hat{X} + UU^{\intercal}\hat{X} = \hat{X}V_{\perp}V_{\perp}^{\intercal} + \hat{X}VV^{\intercal}$, we have

$$\begin{aligned}
\sigma_{r+1}^2(\hat{X}) &\le \min\{\sigma_1^2(U_{\perp}U_{\perp}^{\intercal}\hat{X}), \sigma_1^2(\hat{X}V_{\perp}V_{\perp}^{\intercal})\} \qquad \text{(by Lemma 2, part 1)} \\
&= \min\{\sigma_1^2(Z_{21} + U_{\perp}^{\intercal}\hat{X}V_{\perp}), \sigma_1^2(Z_{12} + U_{\perp}^{\intercal}\hat{X}V_{\perp})\} \\
&\le (\beta^2 + z_{12}^2) \wedge (\beta^2 + z_{21}^2) \qquad \text{(by Lemma 2, part 2)} \\
&= \beta^2 + z_{12}^2 \wedge z_{21}^2.
\end{aligned}$$

We shall also note the fact that for any matrix $A \in \mathbb{R}^{p\times r}$ with $r \le p$, denote the SVD as $A = U_A\Sigma_A V_A^{\intercal}$, then

$$\|A(A^{\intercal}A)^{\dagger}\| = \|U_A\Sigma_A V_A^{\intercal}(V_A\Sigma_A^2 V_A^{\intercal})^{\dagger}\| = \|U_A\Sigma_A^{\dagger}V_A^{\intercal}\| \le \sigma_{\min}^{-1}(A). \tag{8.18}$$

Thus,

$$\begin{aligned}
\|\mathbb{P}_{(\hat{X}V)}\hat{X}V_{\perp}\| &= \|\mathbb{P}_{(\hat{X}V)}\mathbb{P}_U\hat{X}V_{\perp} + \mathbb{P}_{(\hat{X}V)}\mathbb{P}_{U_{\perp}}\hat{X}V_{\perp}\| \\
&\le \|\mathbb{P}_{(\hat{X}V)}UU^{\intercal}\hat{X}V_{\perp}\| + \|\mathbb{P}_{(\hat{X}V)}U_{\perp}U_{\perp}^{\intercal}\hat{X}V_{\perp}\| \\
&\le \|U^{\intercal}\hat{X}V_{\perp}\| + \|\hat{X}V[(\hat{X}V)^{\intercal}(\hat{X}V)]^{-1}(\hat{X}V)^{\intercal}U_{\perp}U_{\perp}^{\intercal}\hat{X}V_{\perp}\| \\
&\le \|U^{\intercal}\hat{X}V_{\perp}\| + \|\hat{X}V[(\hat{X}V)^{\intercal}(\hat{X}V)]^{-1}\| \cdot \|U_{\perp}^{\intercal}\hat{X}V\| \cdot \|U_{\perp}^{\intercal}\hat{X}V_{\perp}\| \\
&\overset{(8.18)}{\le} \|U^{\intercal}ZV_{\perp}\| + \frac{1}{\sigma_{\min}(\hat{X}V)}\|U_{\perp}^{\intercal}ZV\| \cdot \|U_{\perp}^{\intercal}\hat{X}V_{\perp}\| \\
&\overset{(8.17)}{\le} z_{12} + \frac{\beta}{\alpha}z_{21} = \frac{\alpha z_{12} + \beta z_{21}}{\alpha}.
\end{aligned}$$

Similarly,

$$\|P_{(\hat{X}V)}\hat{X}V_{\perp}\|_F \le \frac{\alpha\|Z_{12}\|_F + \beta\|Z_{21}\|_F}{\alpha}.$$

Next, applying Proposition 1 by setting $A = \hat{X}$, $\tilde{W} = [V\ V_{\perp}]$, $\tilde{V} = [\hat{V}\ \hat{V}_{\perp}]$, we could obtain (2.4). □

SUPPLEMENTARY MATERIAL

Supplement to "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics" (DOI: 10.1214/17-AOS1541SUPP; .pdf). The supplementary material includes the proofs for Theorem 2, Corollary 1, matrix denoising, high-dimensional clustering, canonical correlation analysis and all the technical lemmas.

REFERENCES

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

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ.

ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2008). Convex multi-task feature learning. *Mach*. *Learn*. **73** 243–272.

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems* 2139–2147.

BALAKRISHNAN, S., XU, M., KRISHNAMURTHY, A. and SINGH, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems* 954–962.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J*. *Multivariate Anal*. **111** 120–135. MR2944410

BORG, I. and GROENEN, P. J. (2005). *Modern Multidimensional Scaling*: *Theory and Applications*. Springer, Berlin.

BOUTSIDIS, C., ZOUZIAS, A., MAHONEY, M. W. and DRINEAS, P. (2015). Randomized dimensionality reduction for k-means clustering. *IEEE Trans*. *Inform*. *Theory* **61** 1045–1062.

BURA, E. and PFEIFFER, R. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statist*. *Probab*. *Lett*. **78** 2275–2280.

CAI, T. T., LI, X. and MA, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *Ann*. *Statist*. **44** 2221–2251. MR3546449

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann*. *Statist*. **41** 3074–3110. MR3161458

CAI, T. T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab*. *Theory Related Fields* **161** 781–815.

CAI, T. T and ZHANG, A. (2017). Supplement to "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics." DOI:10.1214/17-AOS1541SUPP.

CAI, T., CAI, T. T., LIAO, K. and LIU, W. (2015). Large-scale simultaneous testing of cross-covariance matrix with applications to PheWAS. Technical report.

CANDES, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc*. *IEEE* **98** 925–936.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found*. *Comput*. *Math*. **9** 717–772. MR2565240

CANDÈS, E. J., SING-LONG, C. A. and TRZASKO, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans*. *Signal Process*. **61** 4643–4657. MR3105401

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans*. *Inform*. *Theory* **56** 2053–2080.

CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann*. *Probab*. **37** 1–47. MR2489158

CHATTERJEE, S. (2014). Matrix estimation by universal singular value thresholding. *Ann*. *Statist*. **43** 177–214. MR3285604

CHEN, M., GAO, C., REN, Z. and ZHOU, H. H. (2013). Sparse CCA via precision adjusted iterative thresholding. Preprint. Available at arXiv:1311.6186.

CHO, J., KIM, D. and ROHE, K. (2015). Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. Preprint. Available at arXiv:1508.05431.

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.

DONOHO, D. and GAVISH, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.* **42** 2413–2440. MR3269984

DOPICO, F. M. (2000). A note on sin Θ theorems for singular subspace variations. *BIT Numerical Mathematics* **40** 395–403.

FAN, J., WANG, W. and ZHONG, Y. (2016). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. Preprint. Available at arXiv:1603.03516.

FELDMAN, D., SCHMIDT, M. and SOHLER, C. (2013). Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* 1434–1453. SIAM, Philadelphia, PA.

GAO, C., MA, Z. and ZHOU, H. H. (2014). Sparse CCA: Adaptive estimation and computational barriers. Preprint. Available at arXiv:1409.8565.

GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982

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

GOLDBERG, D., NICHOLS, D., OKI, B. M. and TERRY, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM* **35** 61–70.

GROSS, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566.

HARDOON, D. R., SZEDMAK, S. and SHAWE-TAYLOR, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **16** 2639–2664.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

JIN, J., KE, Z. T. and WANG, W. (2015). Phase transitions for high dimensional clustering and related problems. Preprint. Available at arXiv:1502.06952.

JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering (with discussion). *Ann. Statist.* **44** 2323–2359. MR3576548

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

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

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

LIU, Z. and VANDENBERGHE, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* **31** 1235–1256. MR2558821

MA, Z. and LI, X. (2016). Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. Preprint. Available at arXiv:1605.03662.

MELAMED, D. (2014). Community structures in bipartite networks: A dual-projection approach. *PLoS ONE* **9** e97823.

O'ROURKE, S., VU, V. and WANG, K. (2013). Random perturbation of low rank matrices: Improving classical bounds. Preprint. Available at arXiv:1311.2657.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SHABALIN, A. A. and NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.* **118** 67–76.

SINGER, A. and CUCURINGU, M. (2010). Uniqueness of low-rank matrix completion by rigidity theory. *SIAM J. Matrix Anal. Appl.* **31** 1621–1641.

STEWART, G. W. (1991). Perturbation theory for the singular value decomposition. In *SVD and Signal Processing II*: *Algorithms*, *Analysis and Applications* 99–109. Elsevier, Amsterdam.

STEWART, M. (2006). Perturbation of the SVD in the presence of small singular values. *Linear Algebra Appl*. **419** 53–77.

SUN, R. and LUO, Z.-Q. (2015). Guaranteed matrix completion via nonconvex factorization. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 270–289. IEEE Computer Soc., Los Alamitos, CA. MR3473312

TAO, T. (2012). *Topics in Random Matrix Theory*. *Graduate Studies in Mathematics* **132**. Amer. Math. Soc., Providence, RI. MR2906465

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge.

VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann*. *Statist*. **36** 555–586. MR2396807

VU, V. (2011). Singular vectors under random perturbation. *Random Structures Algorithms* **39** 526–538.

WANG, R. (2015). Singular vector perturbation under Gaussian noise. *SIAM J. Matrix Anal*. *Appl*. **36** 158–177.

WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111.

WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math*. *Ann*. **71** 441–479.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 515–534.

YANG, D., MA, Z. and BUJA, A. (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. *J*. *Mach*. *Learn*. *Res*. **17** Paper No. 92, 27. MR3543498

YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~tcai/

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN–MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: anruzhang@stat.upenn.edu
URL: http://www.stat.wisc.edu/~anruzhang/