MINIMAX ESTIMATION IN SPARSE CANONICAL CORRELATION ANALYSIS

BY CHAO GAO[1,*], ZONGMING MA[2,†], ZHAO REN[1,‡]
AND HARRISON H. ZHOU[1,*]

Yale University[*], *University of Pennsylvania*[†] *and University of Pittsburgh*[‡]

Canonical correlation analysis is a widely used multivariate statistical technique for exploring the relation between two sets of variables. This paper considers the problem of estimating the leading canonical correlation directions in high-dimensional settings. Recently, under the assumption that the leading canonical correlation directions are sparse, various procedures have been proposed for many high-dimensional applications involving massive data sets. However, there has been few theoretical justification available in the literature. In this paper, we establish rate-optimal nonasymptotic minimax estimation with respect to an appropriate loss function for a wide range of model spaces. Two interesting phenomena are observed. First, the minimax rates are not affected by the presence of nuisance parameters, namely the covariance matrices of the two sets of random variables, though they need to be estimated in the canonical correlation analysis problem. Second, we allow the presence of the residual canonical correlation directions. However, they do not influence the minimax rates under a mild condition on eigengap. A generalized sin-theta theorem and an empirical process bound for Gaussian quadratic forms under rank constraint are used to establish the minimax upper bounds, which may be of independent interest.

1. Introduction. Canonical correlation analysis (CCA) [20] is one of the most classical and important tools in multivariate statistics [3, 28]. It has been widely used in various fields to explore the relation between two sets of variables measured on the same sample.

On the population level, given two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$, CCA first seeks two vectors $u_1 \in \mathbb{R}^p$ and $v_1 \in \mathbb{R}^m$ such that the correlation between the projected variables $u_1'X$ and $v_1'Y$ is maximized. More specifically, (u_1, v_1) is the solution to the following optimization problem:

$$(1) \quad \max_{u \in \mathbb{R}^p, v \in \mathbb{R}^m} \mathrm{Cov}(u'X, v'Y), \qquad \text{subject to} \quad \mathrm{Var}(u'X) = \mathrm{Var}(v'Y) = 1,$$

which is uniquely determined up to a simultaneous sign change when there is a positive eigengap. Inductively, once (u_i, v_i) is found, one can further obtain (u_{i+1}, v_{i+1}) by solving the above optimization problem repeatedly subject to the extra constraint that

$$\text{Cov}(u'X, u'_j X) = \text{Cov}(v'Y, v'_j Y) = 0 \qquad \text{for } j = 1, \dots, i.$$

Throughout the paper, we call the (u_i, v_i)'s canonical correlation directions. It was shown by Hotelling [20] that the $(\Sigma_x^{1/2} u_i, \Sigma_y^{1/2} v_i)$'s are the successive singular vector pairs of

$$(2) \qquad \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2},$$

where $\Sigma_x = \text{Cov}(X)$, $\Sigma_y = \text{Cov}(Y)$ and $\Sigma_{xy} = \text{Cov}(X, Y)$. When one is only given a random sample $\{(X_i, Y_i) : i = 1, \dots, n\}$ of size n, classical CCA estimates the canonical correlation directions by performing singular value decomposition (SVD) on the sample counterpart of (2) first and then premultiply the singular vectors by the inverse of square roots of the sample covariance matrices. For fixed dimensions p and m, the estimators are well behaved when the sample size is large [2].

However, in contemporary datasets, we typically face the situation where the ambient dimension in which we observe data is very high while the sample size is small. The dimensions p and m can be much larger than the sample size n. For example, in cancer genomic studies, X and Y can be gene expression and DNA methylation measurements, respectively, where the dimensions p and m can be as large as tens of thousands while the sample size n is typically no larger than several hundreds [12]. When applied to datasets of such nature, classical CCA faces at least three key challenges. First, the canonical correlation directions obtained through classical CCA procedures involve all the variables measured on each subject, and hence are difficult to interpret. Second, due to the amount of noise that increases dramatically as the ambient dimension grows, it is typically impossible to consistently estimate even the leading canonical correlation directions without any additional structural assumption [5, 22]. Third, successive canonical correlation directions should be orthogonal with respect to the population covariance matrices which are notoriously hard to estimate in high-dimensional settings. Indeed, it is not possible to obtain a substantially better estimator than the sample covariance matrix [27] which usually behaves poorly [21]. So, the estimation of such nuisance parameters further complicates the problem of high-dimensional CCA.

Motivated by genomics, neuroimaging and other applications, there have been growing interests in imposing sparsity assumptions on the leading canonical correlation directions. See, for example, [4, 19, 25, 29, 34, 36, 38, 39] for some recent methodological developments and applications. By seeking sparse canonical correlation directions, the estimated (u_i, v_i) vectors only involve a small number of variables, and hence are easier to interpret.

Despite these recent methodological advances, theoretical understanding about the sparse CCA problem is lacking. It is unclear whether the sparse CCA algorithms proposed in the literature have consistency or certain rates of convergence if the population canonical correlation directions are indeed sparse. To the best of our limited knowledge, the only theoretical work available in the literature is [13]. In this paper, the authors gave a characterization for the sparse CCA problem and considered an idealistic single canonical pair model where Σ_{xy}, the covariance between X and Y, was assumed to have a rank one structure. They reparametrized Σ_{xy} as follows:

$$
\Sigma_{xy} = \Sigma_x \lambda u v' \Sigma_y, \tag{3}
$$

where $\lambda \in (0, 1)$ and $u' \Sigma_x u = v' \Sigma_y v = 1$. It can be shown that (u, v) is the solution to (1), so that they are the leading canonical correlation directions. It is worth noting that without knowledge of Σ_x and Σ_y, one is not able to obtain (resp., estimate) (u, v) by simply applying singular value decomposition to Σ_{xy} (resp., sample covariance $\widehat{\Sigma}_{xy}$). Under this model, Chen et al. [13] studied the minimax lower bound for estimating the individual vectors u and v, and proposed an iterative thresholding approach for estimating u and v, partially motivated by [26]. However, their results depend on how well the nuisance parameters Σ_x and Σ_y can be estimated, which to our surprise, turns out to be unnecessary as shown in this paper.

1.1. *Main contributions.* The main objective of the current paper is to understand the fundamental limits of the sparse CCA problem from a decision-theoretic point of view. Such an investigation is not only interesting in its own right, but will also inform the development and evaluation of practical methodologies in the future. The model considered in this work is very general. As shown in [13], Σ_{xy} can be reparametrized as follows:

$$
\Sigma_{xy} = \Sigma_x (U \Lambda V') \Sigma_y \qquad \text{with } U' \Sigma_x U = V' \Sigma_y V = I_{\bar{r}}, \tag{4}
$$

where $\bar{r} = \min(p, m)$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{\bar{r}})$ and $1 > \lambda_1 \geq \cdots \geq \lambda_{\bar{r}} \geq 0$. Then the successive columns of U and V are the leading canonical correlation directions. Therefore, (4) is the most general model for covariance structure, and sparse CCA actually means the leading columns of U and V are sparse.

We can split $U \Lambda V'$ as

$$
U \Lambda V' = U_1 \Lambda_1 V_1' + U_2 \Lambda_2 V_2', \tag{5}
$$

where $\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_r)$, $\Lambda_2 = \text{diag}(\lambda_{r+1}, \dots, \lambda_{\bar{r}})$, $U_1 \in \mathbb{R}^{p \times r}$, $V_1 \in \mathbb{R}^{m \times r}$, $U_2 \in \mathbb{R}^{p \times r_2}$ and $V_2 \in \mathbb{R}^{m \times r_2}$ for $r_2 = \bar{r} - r$. In what follows, we call (U_1, V_1) the *leading* and (U_2, V_2) the *residual* canonical correlation directions. Since our primary interest lies in U_1 and V_1, both the covariance matrices Σ_x and Σ_y and the residual canonical correlation directions U_2 and V_2 are nuisance parameters in

our problem. This model is more general than (3) considered in [13]. It captures the situation in real practice where one is interested in recovering the first few sparse canonical correlation directions while there might be additional directions in the population structure.

To measure the performance of a procedure, we propose to estimate the matrix $U_1 V_1'$ under the following loss function:

$$(6) \qquad L(U_1 V_1', \widehat{U_1 V_1'}) = \|U_1 V_1' - \widehat{U_1 V_1'}\|_F^2.$$

We choose this loss function for several reasons. First, even when the λ_i's are all distinct, U_1 and V_1 are only determined up to a simultaneous sign change of their columns. In contrast, the matrix $U_1 V_1'$ is uniquely defined as long as $\lambda_r > \lambda_{r+1}$. Second, (6) is stronger than the squared projection error loss. For any matrix A, let P_A stand for the projection matrix onto its column space. If the spectra of Σ_x and Σ_y are both bounded away from zero and infinity, then, in view of Wedin's sin-theta theorem [37], any upper bound on the loss function (6) leads to an upper bound on the loss functions $\|P_{U_1} - \widehat{P}_{U_1}\|_F^2$ and $\|P_{V_1} - \widehat{P}_{V_1}\|_F^2$ for estimating the column subspaces of U_1 and V_1, which have been used in the related problem of sparse principal component analysis [11, 33]. Third, this loss function comes up naturally as the key component in the Kullback–Leibler divergence calculation for a special class of normal distributions where $\Sigma_x = I_p$, $\Sigma_y = I_m$ and $\lambda_{r+1} = \cdots = \lambda_{\bar{r}} = 0$ in (4).

We use weak-ℓ_q balls to quantify sparsity. Let $\|(U_1)_{j*}\|$ denote the ℓ_2 norm of the jth row of U_1, and let $\|(U_1)_{(1)*}\| \geq \cdots \geq \|(U_1)_{(p)*}\|$ be the ordered row norms. One way to characterize the sparsity in U_1 (and V_1) is to look at its weak-ℓ_q radius for some $q \in [0, 2)$,

$$(7) \qquad \|U_1\|_{q,w} = \max_{j \in [p]} j \|(U_1)_{(j)*}\|^q$$

under the tradition that $0^q = 0$. For instance, in the case of exact sparsity, that is, $q = 0$, $\|U_1\|_{0,w}$ counts the number of nonzero rows in U_1. When $q \in (0, 2)$, (7) quantifies the decay of the ordered row norms of U_1, which is a form of approximate sparsity. Then we define the parameter space $\mathcal{F}_q(s_u, s_v, p, m, r, \lambda; \kappa, M)$, as the collection of all covariance matrices

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

with the CCA structure (4) and (5), which satisfies:

1. $U_1 \in \mathbb{R}^{p \times r}$ and $V_1 \in \mathbb{R}^{m \times r}$ satisfying $\|U_1\|_{q,w} \leq s_u$ and $\|V_1\|_{q,w} \leq s_v$;
2. $\|\Sigma_x^l\|_{op} \vee \|\Sigma_y^l\|_{op} \leq M$ for $l = \pm 1$;
3. $1 > \kappa\lambda \geq \lambda_1 \geq \cdots \geq \lambda_r \geq \lambda > 0$.

Throughout the paper, we assume $\kappa\lambda \leq 1 - c_0$ for some absolute constant $c_0 \in (0, 1)$. The key parameters s_u, s_v, p, m, r and λ are allowed to depend on the sample size n, while $\kappa, M > 1$ are treated as absolute constants. Compared with the

single canonical pair model (3) in [13], where rank$(\Sigma_{xy}) = 1$, in this paper, the rank of Σ_{xy} can be as high as p or m and r is allowed to grow. In addition, we do not need any structural assumption on Σ_x and Σ_y except for condition 2 on the largest and smallest eigenvalues, which implies that Σ_x and Σ_y are invertible.

Suppose we observe i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \sim N_{p+m}(0, \Sigma)$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \asymp b_n$ if for some absolute constant $C > 1$, $1/C \le a_n/b_n \le C$ for all n. By the minimax lower and upper bound results in Section 2, under mild conditions, we obtain the following tight nonasymptotic minimax rates for estimating the leading canonical directions when $q = 0$:

(8)
$$
\inf_{\widehat{U_1 V_1'}} \sup_{\Sigma \in \mathcal{F}_0(s_u, s_v, p, m, r, \lambda)} \mathbb{E}_\Sigma \| U_1 V_1' - \widehat{U_1 V_1'} \|_{\mathrm{F}}^2
$$
$$
\asymp \frac{1}{n\lambda^2} \left(r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v} \right).
$$

In Section 2, we give a precise statement of this result and tight minimax rates for the case of approximate sparsity, that is, $q \in (0, 2)$.

The result (8) provides a precise characterization of the statistical fundamental limit of the sparse CCA problem. It is worth noting that the conditions required for (8) do not involve any additional assumptions on the nuisance parameters Σ_x, Σ_y, U_2 and V_2. Therefore, we are able to establish the remarkable fact that the fundamental limit of the sparse CCA problem is *not* affected by those nuisance parameters. This optimality result can serve as an important guideline to evaluate procedures proposed in the literature.

To obtain minimax upper bounds, we propose an estimator by optimizing canonical correlation under sparsity constraints. A key element in analyzing the risk behavior of the estimator is a generalized sin-theta theorem. See Theorem 5 in Section 5.1. The theorem is of interest in its own right and can be useful in other problems where matrix perturbation analysis is needed. It is worth noting that the proposed procedure does *not* require sample splitting, which was needed in [11]. We bypass sample splitting by establishing a new empirical process bound for the supreme of Gaussian quadratic forms with rank constraint. See Lemma 7 in Section 5.1. The estimator is shown to be minimax rate optimal by establishing matching minimax lower bounds based on a local metric entropy approach [8, 11, 24, 41].

1.2. *Connection to and difference from sparse PCA.* The current paper is related to the problem of sparse principal component analysis (PCA), which has received a lot of recent attention in the literature. Most literature on sparse PCA considers the spiked covariance model [21, 32] where one observes an $n \times p$ data matrix, each row of which is independently sampled from a normal distribution $N_p(0, \Sigma_0)$ with

(9)
$$
\Sigma_0 = V \Lambda V' + \sigma^2 I_p.
$$

Here, $V \in \mathbb{R}^{p \times r}$ has orthonormal column vectors which are assumed to be sparse and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$. Since the first r eigenvalues of Σ_0 are $\{\lambda_i + \sigma^2\}_{i=1}^r$ and the rest are all σ^2, the λ_i's are referred as "spikes," and hence the name of the model. Johnstone and Lu [23] proposed a diagonal thresholding estimator of the sparse principal eigenvector which is provably consistent for a range of sparsity regimes. For fixed r, Birnbaum et al. [9] derived minimax rate optimal estimators for individual sparse principal eigenvectors, and Ma [26] proposed to directly estimate sparse principal subspaces, that is, the span of V, and constructed an iterative thresholding algorithm for this purpose which is shown to achieve near optimal rate of convergence adaptively. Cai et al. [11] studied minimax rates and adaptive estimation for sparse principal subspaces with little constraint on r. See also [33] for the case of a more general model. In addition, variable selection, rank detection, computational complexity and posterior contraction rates of sparse PCA have been studied. See, for instance, [1, 6, 10, 17] and the references therein.

Compared with sparse PCA, the sparse CCA problem studied in the current paper is different and arguably more challenging in three important ways.

- In sparse PCA, the sparse vectors of interest, that is, the columns of V in (9) are normalized with respect to the identity matrix. In contrast, in sparse CCA, the sparse vectors of interest, that is, the columns of U and V are normalized with respect to Σ_x and Σ_y, respectively, which are not only unknown but also hard to estimate in high-dimensional settings. The necessity of normalization with respect to nuisance parameters adds on to the difficulty of the sparse CCA problem.

- In sparse PCA, especially in the spiked covariance model, there is a clean separation between "signal" and "noise": the signal is in the spiked part and the rest are noise. However, in the parameter space considered in this paper, we allow the presence of residual canonical correlations $U_2 \Lambda_2 V_2'$, which is motivated by the situation statisticians face in practice. It is highly nontrivial to show that the presence of the residual canonical correlations does not influence the minimax estimation rates.

- The covariance structures in sparse PCA and sparse CCA have both sparsity and low-rank structures. However, there is a subtle difference between the two. In sparse PCA, the sparsity and orthogonality of V in (9) are coherent. This means that the columns of V are sparse and orthogonal to each other simultaneously. Such convenience is absent in the sparse CCA problem. It is implied from (4) that $\Sigma_x^{1/2} U_1$ and $\Sigma_y^{1/2} V_1$ have orthogonal columns, while it is the columns of U_1 and V_1 that are sparse. The orthogonal columns and the sparse columns are different. The consequence is that in order to estimate the sparse matrices U_1 and V_1, we must appeal to the orthogonality in the nonsparse matrices $\Sigma_x^{1/2} U_1$ and $\Sigma_y^{1/2} V_1$, even when the matrices Σ_x and Σ_y are unknown. If we naively treat sparse CCA as sparse PCA, the procedure can be inconsistent (see the simulation results in [13]).

1.3. *Organization of the paper.*   The rest of the paper is organized as follows. Section 2 presents the main results of the paper, including upper bounds in Section 2.1 and lower bounds in Section 2.2. Section 3 discusses some related issues. The proofs of the minimax upper bounds are gathered in Section 4, with some auxiliary results and technical lemmas proved in Section 5. The proof of the lower bounds and some further technical lemmas are given in the supplementary material [15].

1.4. *Notation.*   For any matrix $A = (a_{ij})$, the ith row of A is denoted by A_{i*} and the jth column by A_{*j}. For a positive integer p, $[p]$ denotes the index set $\{1, 2, \ldots, p\}$. For any set I, $|I|$ denotes its cardinality and I^c its complement. For two subsets I and J of indices, we write A_{IJ} for the $|I| \times |J|$ submatrices formed by a_{ij} with $(i, j) \in I \times J$. When I or J is the whole set, we abbreviate it with an $*$, and so if $A \in \mathbb{R}^{p \times k}$, then $A_{I*} = A_{I[k]}$ and $A_{*J} = A_{[p]J}$. For any square matrix $A = (a_{ij})$, denote its trace by $\mathsf{Tr}(A) = \sum_i a_{ii}$. Moreover, let $O(p, k)$ denote the set of all $p \times k$ orthonormal matrices and $O(k) = O(k, k)$. For any matrix $A \in \mathbb{R}^{p \times k}$, $\sigma_i(A)$ stands for its ith largest singular value. The Frobenius norm and the operator norm of A are defined as $\|A\|_F = \sqrt{\mathsf{Tr}(A'A)}$ and $\|A\|_{op} = \sigma_1(A)$, respectively. The support of A is defined as $\mathrm{supp}(A) = \{i \in [n] : \|A_{i*}\| > 0\}$. The trace inner product of two matrices $A, B \in \mathbb{R}^{p \times k}$ is defined as $\langle A, B \rangle = \mathsf{Tr}(A'B)$. For any number a, we use $\lceil a \rceil$ to denote the smallest integer that is no smaller than a. For any two numbers a and b, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For any event E, we use $\mathbf{1}_{\{E\}}$ to denote its indicator function. We use $\mathbb{P}_\Sigma$ to denote the probability distribution of $N_{p+m}(0, \Sigma)$ and $\mathbb{E}_\Sigma$ for the associated expectation.

2. Main results.

In this section, we state the main results of the paper. In Section 2.1, we introduce a method to estimate the leading canonical correlation directions. Minimax upper bounds are obtained. Section 2.2 gives minimax lower bounds which match the upper bounds up to a constant factor. We abbreviate the parameter space $\mathcal{F}_q(s_u, s_v, p, m, r, \lambda; \kappa, M)$ as $\mathcal{F}_q$.

2.1. *Upper bounds.*   The main idea of the estimator proposed in this paper is to maximize the canonical correlations under sparsity constraints. Note that the SVD approach of the classical CCA [20] can be written in the following optimization form:

$$(10) \qquad \max_{(A,B)} \mathsf{Tr}(A'\widehat{\Sigma}_{xy}B) \qquad \text{s.t.} \quad A'\widehat{\Sigma}_x A = B'\widehat{\Sigma}_y B = I_r.$$

We generalize (10) to the high-dimensional setting by adding sparsity constraints.

Since the leading canonical correlation directions (U_1, V_1) are weak ℓ_q sparse, we introduce effective sparsity for $q \in [0, 2)$, which plays a key role in defining

the procedure. Define

$$
(11) \qquad x_q^u = \max\left\{0 \le x \le p : x \le s_u\left(\frac{n\lambda^2}{r + \log(ep/x)}\right)^{q/2}\right\},
$$

$$
(12) \qquad x_q^v = \max\left\{0 \le x \le m : x \le s_v\left(\frac{n\lambda^2}{r + \log(em/x)}\right)^{q/2}\right\}.
$$

The effective sparsity of U_1 and V_1 are defined as

$$
(13) \qquad k_q^u = \lceil x_q^u \rceil, \qquad k_q^v = \lceil x_q^v \rceil.
$$

For $j \ge k_q^u$, it can be shown that

$$
\|(U_1)_{(j)*}\| \le \left(\frac{r + \log(ep/k_q^u)}{n\lambda^2}\right)^{1/2},
$$

for which the signal is not strong enough to be recovered from the data. It holds similarly for V_1.

For n i.i.d. observations (X_i, Y_i), $i \in [n]$, we compute the sample covariance matrix

$$
\widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\Sigma}_y \end{bmatrix}.
$$

The estimator $(\widehat{U}_1, \widehat{V}_1)$ for (U_1, V_1), the leading r canonical correlation directions, is defined as a solution to the following optimization problem:

$$
(14) \qquad
\begin{aligned}
&\max_{(A,B)} \ \mathsf{Tr}(A'\widehat{\Sigma}_{xy}B) \\
&\text{s.t.} \quad A'\widehat{\Sigma}_x A = B'\widehat{\Sigma}_y B = I_r \text{ and } \|A\|_{0,w} = k_q^u, \|B\|_{0,w} = k_q^v.
\end{aligned}
$$

When $q = 0$, we have $k_q^u = s_u$ and $k_q^v = s_v$. Then the program (14) is just a slight generalization of the classical approach of [20] with additional ℓ_0 constraints $\|A\|_{0,w} = s_u$ and $\|B\|_{0,w} = s_v$. By the definition of the parameter space, it is also natural to impose the ℓ_q constraints $\|A\|_{q,w} \le s_u$ and $\|B\|_{q,w} \le s_v$. Such constraints were used by [33] to solve the sparse PCA problem. However, their upper bounds require more assumptions due to the difficulty in analyzing ℓ_q constraints. We use ℓ_0 constraints on the effective sparsity and obtain the optimal upper bound under minimal assumptions.

Set

$$
(15) \qquad \varepsilon_n^2 = \frac{1}{n\lambda^2}\left(r(k_q^u + k_q^v) + k_q^u \log\frac{ep}{k_q^u} + k_q^v \log\frac{em}{k_q^v}\right),
$$

which is the minimax rate to be shown later.

THEOREM 1.   *We assume that*

$$(16) \qquad\qquad\qquad \varepsilon_n^2 \le c,$$

$$(17) \qquad\qquad\qquad \lambda_{r+1} \le c\lambda,$$

for some sufficiently small constant $c \in (0, 1)$. *For any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M, q, κ *and* C', *such that for any* $\Sigma \in \mathcal{F}_q$,

$$\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}^2 \le C\varepsilon_n^2,$$

with $\mathbb{P}_\Sigma$-*probability at least* $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$.

REMARK 1.   It will be shown in Section 2.2 that assumption (16) is necessary for consistent estimation. Assumption (17) implies $\lambda_{r+1} \le c\lambda_r$ for $c \in (0, 1)$, such that the eigengap is lower bounded as $\lambda_r - \lambda_{r+1} \ge (1 - c)\lambda_r > 0$.

REMARK 2.   The upper bound ε_n^2 has two parts. The first part $\frac{1}{n\lambda^2}(r(k_q^u + k_q^v))$ is caused by low rank structure, and the second part $\frac{1}{n\lambda^2}(k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v))$ is caused by sparsity. If $r \le \log(ep/k_q^u) \wedge \log(em/k_q^v)$, the second part dominates, while the first part dominates if $r \ge \log(ep/k_q^u) \vee \log(em/k_q^v)$.

REMARK 3.   The upper bound does not require any structural assumption on the marginal covariance matrices Σ_x and Σ_y other than bounds on the largest and the smallest eigenvalues. Although in the high-dimensional setting, the sample covariance $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_y$ are not good estimators of the matrices Σ_x, Σ_y, the normalization constraints $A'\widehat{\Sigma}_x A = B'\widehat{\Sigma}_y B = I_r$, together with the sparsity of A, B, only involve submatrices of $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_y$. Under the assumption (16), it can be shown that a $k_q^u \times k_q^u$ submatrix of $\widehat{\Sigma}_x$ converges to the corresponding submatrix of Σ_x with the rate $\sqrt{\frac{k_q^y \log(ep/k_q^u)}{n}}$ under operator norm uniformly over all $k_q^u \times k_q^u$ submatrices. Similar results hold for $\widehat{\Sigma}_y$ and Σ_y. See Lemma 12 in Section 5.4. These rates are dominated by the minimax rate ε_n in (15).

REMARK 4.   One of the major difficulties of sparse CCA is the presence of the unknown Σ_x and Σ_y. Suppose Σ_x and Σ_y are known, one may work with the transformed data $\{(\Sigma_x^{-1} X_i, \Sigma_y^{-1} Y_i) : i = 1, \ldots, n\}$. The cross-covariance of the transformed data is $\Sigma_x^{-1} \Sigma_{xy} \Sigma_y^{-1} = U\Lambda V'$, which is a sparse matrix. When $\mathrm{rank}(\Sigma_{xy}) = 1$, algorithms such as [13, 40] can obtain the sparse singular vectors from $\Sigma_x^{-1} \widehat{\Sigma}_{xy} \Sigma_y^{-1}$, which estimate U_1 and V_1 with optimal rate. When Σ_x and Σ_y are unknown, structural assumptions are required on the covariance matrices in order that Σ_x^{-1} and Σ_y^{-1} can be well estimated. Then one can use the estimated Σ_x^{-1} and Σ_y^{-1} to transform the data and apply the same sparse singular vector

estimator (see [13]). However, unless $\Sigma_x = I_p$ and $\Sigma_y = I_m$, this method cannot be extended to the case where $\mathrm{rank}(\Sigma_{xy}) \geq 2$, since the orthogonality of U and V is with respect to general covariance matrices Σ_x and Σ_y, respectively. In the case where $\Sigma_x = I_p$ and $\Sigma_y = I_m$, the problem is similar to sparse PCA, and the proof of Theorem 1 can be greatly simplified.

To obtain the convergence rate in expectation, we propose a modified estimator. The modification is inspired by the fact that $U_1 V_1'$ are bounded in Frobenius norm, because

$$(18) \qquad \|U_1 V_1'\|_{\mathrm{F}} \leq \|\Sigma_x^{-1/2}\|_{\mathrm{op}} \|\Sigma_x^{1/2} U_1\|_{\mathrm{F}} \|\Sigma_y^{1/2} V_1\|_{\mathrm{op}} \|\Sigma_y^{-1/2}\|_{\mathrm{op}} \leq M\sqrt{r}.$$

Define $\widehat{U_1 V_1'}$ to be the truncated version of $\widehat{U}_1 \widehat{V}_1'$ as

$$\widehat{U_1 V_1'} = \widehat{U}_1 \widehat{V}_1' \mathbf{1}_{\{\|\widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}} \leq 2M\sqrt{r}\}}.$$

The modification can be viewed as an improvement, because whenever $\|\widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}} > 2M\sqrt{r}$, we have

$$\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}} \geq \|\widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}} - \|U_1 V_1'\|_{\mathrm{F}} \geq M\sqrt{r} \geq \|0 - U_1 V_1'\|_{\mathrm{F}}.$$

Then it is better to estimate $U_1 V_1'$ by 0.

THEOREM 2. *Suppose* (16) *and* (17) *hold. In addition, assume that*

$$(19) \qquad \exp(C_1(k_q^u + \log(ep/k_q^u))) > n\lambda^2,$$

$$(20) \qquad \exp(C_1(k_q^v + \log(em/k_q^v))) > n\lambda^2,$$

for some $C_1 > 0$, then there exists $C_2 > 0$ only depending on M, q, κ and C_1, such that

$$\sup_{\Sigma \in \mathcal{F}_q} \mathbb{E}_{\Sigma} \|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \leq C_2 \varepsilon_n^2.$$

REMARK 5. The assumptions (19) and (20) imply the tail probability in Theorem 1 is sufficiently small. Once there exists a small constant $\delta > 0$, such that

$$p \vee e^{k_q^u} \geq n^\delta \quad \text{and} \quad m \vee e^{k_q^v} \geq n^\delta$$

hold, then (19) and (20) also hold with some $C_1 > 0$. Notice that $p > n^\delta$ is commonly assumed in high-dimensional statistics to have convergence results in expectation. The assumption here is weaker than that.

2.2. *Lower bounds*. Theorems 1 and 2 show that the procedure proposed in (14) attains the rate ε_n^2. In this section, we show that this rate is optimal among all estimators. More specifically, we show that the following minimax lower bounds hold for $q \in [0, 2)$.

THEOREM 3. *Assume that* $1 \leq r \leq \frac{k_q^u \wedge k_q^v}{2}$, *and that*

$$(21) \qquad n\lambda^2 \geq C_0 \left(r + \log \frac{ep}{k_q^u} \vee \log \frac{em}{k_q^v} \right)$$

for some sufficiently large constant C_0. *Then there exists a constant* $c > 0$ *depending only on q and an absolute constant* c_0 *such that the minimax risk for estimating* $U_1 V_1'$ *satisfies*

$$\inf_{(\widehat{U}_1, \widehat{V}_1)} \sup_{\Sigma \in \mathcal{F}_q} \mathbb{E}_\Sigma \|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_F^2 \geq c\varepsilon_n^2 \wedge c_0.$$

The proof of Theorem 3 is given in the supplementary material [15].

REMARK 6. Assumption (21) is necessary for consistent estimation.

3. Discussion. We include below discussions on two related issues.

3.1. *Minimax rates for individual sparsity.* In this paper, we have derived tight minimax estimation rates for the leading sparse canonical correlation directions where the sparsity is depicted by the rapid decay of the ordered row norms in U_1 and V_1 (as characterized by the weak-ℓ_q notion).

Another interesting case of sparsity is when the individual column vectors of U_1 and V_1 are sparse. For instance, when

$$(22) \qquad \|u_i\|_{q,w} \leq t_u \quad \text{and} \quad \|v_i\|_{q,w} \leq t_v \qquad \forall i \in [r],$$

where the $\|\cdot\|_{q,w}$ is defined as in (7) by treating any p-vector as a $p \times 1$ matrix. Let $\mathcal{F}_q^c = \mathcal{F}_q^c(t_u, t_v, p, m, r, \lambda; \kappa, M)$ be defined as in Section 1 following (7) but with the sparsity notion changed to that in (22). Similar to (11)–(13), let

$$y_q^u = \max \left\{ 0 \leq y \leq p : y \leq t_u \left(\frac{n\lambda^2}{\log(ep/(ry))} \right)^{q/2} \right\}, \qquad j_q^u = \lceil y_q^u \rceil,$$

and y_q^v and j_q^v be analogously defined. Then we have:

THEOREM 4. *Assume that* $1 \leq r \leq \frac{j_q^u \wedge j_q^v}{2}$, $2rj_q^u \leq p$, $2rj_q^v \leq m$ *and* $n\lambda^2 \geq C_0(r + \log \frac{ep}{rj_q^u} \vee \log \frac{em}{rj_q^v})$ *for some sufficiently large constant* C_0. *Then there is a constant* $c > 0$ *depending only on q and an absolute constant* $c_0 > 0$ *such that*

$$(23) \qquad \inf_{\widehat{U_1 V_1'}} \sup_{\Sigma \in \mathcal{F}_q^c} \mathbb{E}_\Sigma \|U_1 V_1' - \widehat{U_1 V_1'}\|_F^2 \geq c_0 \wedge \frac{c}{n\lambda^2} r \left(j_q^u \log \frac{ep}{rj_q^u} + j_q^v \log \frac{em}{rj_q^v} \right).$$

If in addition $rj_q^u \leq p^{1-\alpha}, rj_q^v \leq m^{1-\alpha}$ for some small $\alpha \in (0, 1), r \leq C \log(p \wedge m)$ for some $C > 0$ and the conditions of Theorem 2 are satisfied with $k_q^u = rj_q^u$ and $k_q^v = rj_q^v$, then a matching upper bound is achieved by the estimator in Theorem 2 with $k_q^u = rj_q^u$ and $k_q^v = rj_q^v$.

The proof of Theorem 4 is given in the supplementary material [15]. The lower bound (23) for individual sparsity is larger than the minimax rate (15) for joint sparsity when $t_u = s_u$ and $t_v = s_v$.

3.2. *Adaptation, computation and some recent work.* The main purpose of proposing the estimator in (14) is to determine the minimax estimation rates in sparse CCA problem under weak assumptions. Admittedly, it requires the knowledge of parameter space and is computationally intensive.

Designing adaptive and computationally efficient procedures to achieve statistically optimal performance is an interesting and important research direction. Built upon the insights developed in the current paper, Gao et al. [16] have proposed an adaptive and efficient procedure for sparse CCA. The procedure first obtains a crude estimator via a convex relaxation of the problem (14) here which is then refined by a group sparse linear regression. The resulting estimator achieves optimal rates of convergence in estimating the leading sparse canonical directions under a prediction loss without imposing any structural assumption on Σ_x and Σ_y, when the residual directions are absent. Notably, the procedure in [16] requires a larger sample size than in the present paper, which has been shown to be essentially necessary for any computational efficient procedure under the Gaussian CCA model considered here under the assumption of planted clique hardness. The argument has also led to a computational lower bounds for the sparse PCA problem under the Gaussian spiked covariance model, bridging the gap between the sparse PCA literature and the computational lower bounds in [6] and [35].

It is of great interest to further investigate if there is some adaptive and efficient estimator that attains the statistical optimality established in the current paper under full generality.

4. Proof of main results. This section is devoted to the proof of Theorems 1–2. The proof of Theorems 3–4 is given in the supplementary material [15].

4.1. *Outline of proof and preliminaries.* To prove both Theorems 1 and 2, we go through the following three steps:

1. We decompose the value of the loss function into multiple terms which result from different sources;

2. We derive individual high probability bound for each term in the decomposition;

3. We assemble the individual bounds to obtain the desired upper bounds on the loss and the risk functions.

In the rest of this subsection, we carry out these three steps in order. To facilitate the presentation, we introduce below several important quantities to be used in the proof.

Recall the effective sparsity (k_q^u, k_q^v) defined in (13). Let S_u be the index set of the rows of U_1 with the k_q^u largest ℓ_2 norms. In case U_1 has no more than k_q^u nonzero rows, we include in S_u the smallest indices of the zero rows in U_1 such that $|S_u| = k_q^u$. We also define S_v analogously. In what follows, we refer to them as the *effective support sets*.

We define (U_1^*, V_1^*) as a solution to

$$\max_{(A,B)} \text{Tr}(A'\Sigma_{xy}B)$$

(24)

$$\text{s.t.} \quad A'\Sigma_x A = B'\Sigma_y B = I_r \text{ and } \text{supp}(A) \subset S_u, \text{supp}(B) \subset S_v.$$

In what follows, we refer to them as the *sparse approximations* to U_1 and V_1. By definition, when $q = 0$, $U_1^*(V_1^*)' = U_1 V_1'$, which can be derived rigorously from Theorem 5.

In addition, we define the *oracle estimator* $(\widehat{U}_1^*, \widehat{V}_1^*)$ as a solution to

$$\max_{(A,B)} \text{Tr}(A'\widehat{\Sigma}_{xy}B)$$

(25)

$$\text{s.t.} \quad A'\widehat{\Sigma}_x A = B'\widehat{\Sigma}_y B = I_r \text{ and } \text{supp}(A) = S_u, \text{supp}(B) = S_v.$$

In case the program (24) [or (25)] has multiple global optimizers, we define (U_1^*, V_1^*) [or $(\widehat{U}_1^*, \widehat{V}_1^*)$] by picking an arbitrary one.

REMARK 7. The introduction of (24) and (25) is to separate the error brought by not knowing the covariance Σ_x and Σ_y and by not knowing the effective supports S_u and S_v. The program (25) assumes known effective supports but unknown covariance and the program (24) assumes both known effective supports and known covariance.

We note that

$$(U_1^*)_{S_u^c *} = (\widehat{U}_1^*)_{S_u^c *} = 0, \qquad (V_1^*)_{S_v^c *} = (\widehat{V}_1^*)_{S_v^c *} = 0.$$

By definition, the matrices (U_1^*, V_1^*) are normalized with respect to Σ_x and Σ_y, and $(\widehat{U}_1^*, \widehat{V}_1^*)$ are normalized with respect to $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_y$. Note the notation A_{S*} stands for the submatrix of A with rows in S and all columns.

Last but not least, let

(26) $$\widehat{S}_u = \text{supp}(\widehat{U}_1), \qquad \widehat{S}_v = \text{supp}(\widehat{V}_1).$$

By the definition of $(\widehat{U}_1, \widehat{V}_1)$ in (14), we have $|\widehat{S}_u| = k_q^u$ and $|\widehat{S}_v| = k_q^v$ with probability one. Remember the minimax rate ε_n^2 defined in (15).

4.2. *Loss decomposition.* In the first step, we decompose the loss function into five terms as follows.

LEMMA 1. *Assume $\frac{1}{n}(k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v)) < c$ for sufficiently small $c > 0$. For any constant $C' > 0$, there exists a constant $C > 0$ only depending on M and C', such that*

$$\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_F^2$$

$$(27) \qquad \leq 3\|U_1^*(V_1^*)' - U_1 V_1'\|_F^2$$

$$(28) \qquad + 3\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1^*(V_1^*)'\|_F^2$$

$$(29) \qquad - \frac{6C}{\lambda_r}\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1 \widehat{V}_1'\rangle$$

$$(30) \qquad + \frac{6C}{\lambda_r}\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1 \widehat{V}_1'\rangle$$

$$(31) \qquad + \frac{6C}{\lambda_r}\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \widehat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1 \widehat{V}_1'\rangle,$$

with probability at least $1 - \exp(-C' k_q^u \log(ep/k_q^u)) - \exp(-C' k_q^v \log(em/k_q^v))$.

PROOF. See Section 5.2. □

In particular, Lemma 1 decomposes the total loss into the sum of the sparse approximation error in (27), the oracle loss in (28) which is present even if we have the oracle knowledge of the effective support sets S_u and S_v, the bias term in (29) caused by the presence of the residual term $U_2 \Lambda_2 V_2'$ in the CCA structure (4) and the two excess loss terms in (30) and (31) resulting from the uncertainty about the effective support sets. When $q = 0$, the sparse approximation error term (27) vanishes.

4.3. *Bounds for individual terms.* We now state the bounds for the individual terms obtained in Lemma 1 as five separate lemmas. The proofs of these lemmas are deferred to Sections 5.3–5.6.

LEMMA 2 (Sparse approximation). *Suppose (16) and (17) hold. There exists a constant $C > 0$ only depending on M, κ, q, such that*

$$(32) \qquad \|U_1^*(V_1^*)' - U_1 V_1'\|_F^2 \leq \frac{Cq}{2-q}\varepsilon_n^2,$$

$$(33) \qquad \|U_1^* \Lambda_1 (V_1^*)' - U_1 \Lambda_1 V_1'\|_F^2 \leq \frac{Cq}{2-q}\lambda^2 \varepsilon_n^2.$$

LEMMA 3 (Oracle loss).   *Suppose* $\frac{1}{n\lambda^2}(k_q^u + k_q^v + \log(ep/k_q^u) + \log(em/k_q^v)) < c$ *and that* (17) *holds for some sufficiently small* $c > 0$. *For any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M, q, κ *and* C', *such that*

$$(34) \quad \|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1^*(V_1^*)'\|_F^2 \leq \frac{Cr}{n\lambda^2}\left[k_q^u + k_q^v + \log\left(\frac{ep}{k_q^u}\right) + \log\left(\frac{em}{k_q^v}\right)\right],$$

with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$. *Moreover, if* (16) *also holds, then with the same probability*

$$(35) \qquad\qquad \|\widehat{U}_1^*\Lambda_1(\widehat{V}_1^*)' - U_1^*\Lambda_1(V_1^*)'\|_F^2 \leq C\lambda^2\varepsilon_n^2.$$

The proof of Lemma 3 is given in the supplementary material [15]. Since $r \leq k_q^u \wedge k_q^v$, (34) is bounded above by $C\varepsilon_n^2$. The error bounds in Lemma 3 are due to the estimation error of true covariance matrices by sample covariance matrices on the subset $S_u \times S_v$.

LEMMA 4 (Bias).   *Suppose* $\frac{1}{n}(k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v)) < C_1$ *for some constant* $C_1 > 0$. *For any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M, κ, C_1 *and* C', *such that*

$$\left|\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right|$$
$$\leq C\lambda_{r+1}(\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1 V_1'\|_F^2 + \|U_1 V_1' - \widehat{U}_1\widehat{V}_1'\|_F^2),$$

with probability at least $1 - \exp(-C'k_q^u \log(ep/k_q^u)) - \exp(-C'k_q^v \log(em/k_q^v))$.

The bias in Lemma 4 is 0 when $U_2\Lambda_2 V_2'$ is 0.

LEMMA 5 (Excess loss 1).   *Suppose* (16) *holds. For any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M *and* C', *such that*

$$\left|\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right| \leq C\lambda\varepsilon_n \|\widehat{U}_1\widehat{V}_1' - \widehat{U}_1^*(\widehat{V}_1^*)'\|_F,$$

with probability at least $1 - \exp(-C'(r(k_q^u + k_q^v) + k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v)))$.

LEMMA 6 (Excess loss 2).   *Suppose* (16) *and* (17) *hold. For any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M, κ, q *and* C', *such that*

$$\left|\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 (\widehat{V}_1^*)' \widehat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right|$$
$$\leq C\lambda\varepsilon_n \|\widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\|_F,$$

with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$.

4.4. *Proof of Theorem* 1. For notational convenience, let

$$R = \|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}, \qquad \theta = \|U_1^*(V_1^*)' - U_1V_1'\|_{\mathrm{F}},$$

$$\delta = \|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1^*(V_1^*)'\|_{\mathrm{F}}.$$

Consider the event such that the conclusions of Lemmas 1–6 hold, which occurs with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$ according to the union bound. On this event, Lemmas 2 and 3 imply that

$$\theta^2 \le C\varepsilon_n^2 \quad \text{and} \quad \delta^2 \le C\varepsilon_n^2.$$

Moreover, Lemma 4 implies

$$\left|\frac{1}{\lambda_r}\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right| \le \frac{C\lambda_{r+1}}{\lambda}(R^2 + \theta^2 + \delta^2).$$

Lemma 5 implies

$$\left|\frac{1}{\lambda_r}\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right| \le C\varepsilon_n(R + \theta + \delta),$$

and Lemma 6 implies

$$\left|\frac{1}{\lambda_r}\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 (\widehat{V}_1^*)' \widehat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, U_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1'\rangle\right| \le C\varepsilon_n(R + \theta + \delta).$$

Together with Lemma 1, the above bounds lead to

$$R^2 \le C(\theta^2 + \delta^2) + \frac{C\lambda_{r+1}}{\lambda}(R^2 + \theta^2 + \delta^2) + C\varepsilon_n(R + \theta + \delta)$$

$$\le \frac{C\lambda_{r+1}}{\lambda}R^2 + C\varepsilon_n R + C\varepsilon_n^2.$$

Under assumption (17), we have $\frac{1}{2}R^2 \le C\varepsilon_n R + C\varepsilon_n^2$, implying

$$R^2 \le C\varepsilon_n^2,$$

for some $C > 0$. We complete the proof by noting that the conditions of Lemmas 1–6 are satisfied under assumptions (16) and (17).

4.5. *Proof of Theorem* 2. Recall the definition of ε_n in (15), and let C_1 be the constant in (19) and (20). The result of Theorem 1 implies that we can choose an arbitrarily large constant C' such that $C' > C_1$. Given C', there exists a constant C, by which we can bound the risk as follows:

$$\mathbb{E}_\Sigma \|\widehat{U_1V_1'} - U_1V_1'\|_{\mathrm{F}}^2$$

$$\le \mathbb{E}_\Sigma\left[\|\widehat{U_1V_1'} - U_1V_1'\|_{\mathrm{F}}^2 \mathbf{1}_{\{\|\widehat{U_1V_1'} - U_1V_1'\|_{\mathrm{F}}^2 \le C\varepsilon_n^2\}}\right]$$

$$+ \mathbb{E}_\Sigma\big[\|\widehat{U_1 V_1'} - U_1 V_1'\|_F^2 \mathbf{1}_{\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_F^2 > C\varepsilon_n^2\}}\big]$$

$$(36) \qquad \leq C\varepsilon_n^2 + \mathbb{E}_\Sigma\big[(2\|\widehat{U_1 V_1'}\|_F^2 + 2\|U_1 V_1'\|_F^2)\mathbf{1}_{\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_F^2 > C\varepsilon_n^2\}}\big]$$

$$(37) \qquad \leq C\varepsilon_n^2 + 6M^2 r \mathbb{P}_\Sigma\big(\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_F^2 > C\varepsilon_n^2\big)$$

$$(38) \qquad \leq C_2 \varepsilon_n^2.$$

Here, inequality (36) is due to the triangle inequality and the fact that

$$\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_F^2 > C\varepsilon_n^2\} \subset \{\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_F^2 > C\varepsilon_n^2\}.$$

In fact, if $\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_F^2 \leq C\varepsilon_n^2$, then $\|\widehat{U}_1\widehat{V}_1'\|_F^2 \leq C\varepsilon_n^2 + M^2 r \leq 2M^2 r$. By our definition of the estimator, this means $\widehat{U_1 V_1'} = \widehat{U}_1\widehat{V}_1'$, which further implies $\|\widehat{U_1 V_1'} - U_1 V_1'\|_F^2 \leq C\varepsilon_n^2$. Inequality (37) follows from our definition of estimator $\widehat{U_1 V_1'}$ and (18). The last inequality follows from the conclusion of Theorem 1 and assumptions (19) and (20). This completes the proof.

5. Proof of auxiliary results. In this section, we prove Lemmas 1–2 and 4–6 used in the proof of Theorem 1 and 2. The proof of Lemma 3 is given in the supplementary material [15]. Throughout the section, without further notice, ε_n^2 is defined as in (15).

5.1. *A generalized sin-theta theorem and Gaussian quadratic form with rank constraint.* We first introduce two key results used in the proof of Lemmas 1–6 that might be of independent interest.

The first result is a generalized sin-theta theorem. For the definition of unitarily invariant norms, we refer the readers to [7, 30]. In particular, both Frobenius norm $\|\cdot\|_F$ and operator norm $\|\cdot\|_{op}$ are unitarily invariant.

THEOREM 5. *Consider matrices $X, Y \in \mathbb{R}^{p \times m}$. Let the SVD of X and Y be*

$$X = A_1 D_1 B_1' + A_2 D_2 B_2', \qquad Y = \widehat{A}_1 \widehat{D}_1 \widehat{B}_1' + \widehat{A}_2 \widehat{D}_2 \widehat{B}_2',$$

with $D_1 = \mathrm{diag}(d_1, \ldots, d_r)$ and $\widehat{D}_1 = \mathrm{diag}(\widehat{d}_1, \ldots, \widehat{d}_r)$. Suppose there is a positive constant $\delta \in (0, d_r]$ such that $\|\widehat{D}_2\|_{op} \leq d_r - \delta$. Let $\|\cdot\|$ be any unitarily invariant norm, and $\varepsilon = \|A_1'(X - Y)\| \vee \|(X - Y)B_1\|$. Then we have

$$(39) \qquad \|A_1 D_1 B_1' - \widehat{A}_1 \widehat{D}_1 \widehat{B}_1'\| \leq \left(\frac{\sqrt{2}(d_1 + \widehat{d}_1)}{\delta} + 1\right)\varepsilon.$$

If further there is an absolute constant $\bar{\kappa} \geq 1$ such that $d_1 \vee \widehat{d}_1 \leq \bar{\kappa} d_r$, then there is a constant $C > 0$ only depending on $\bar{\kappa}$, such that

$$(40) \qquad \|A_1 B_1' - \widehat{A}_1 \widehat{B}_1'\| \leq \frac{C\varepsilon}{\delta}.$$

REMARK 8. In addition, when X and Y are positive semi-definite, $A_l = B_l$, $\widehat{A}_l = \widehat{B}_l$ for $l = 1, 2$, we recover the classical Davis–Kahan sin-theta theorem [14] $\|A_1 A_1' - \widehat{A}_1 \widehat{A}_1'\| \leq C\varepsilon/\delta$ up to a constant multiplier.

The second result is an empirical process type bound for Gaussian quadratic forms with rank constraint.

LEMMA 7. Let $\{Z_i\}_{1 \leq i \leq n}$ be i.i.d. observations from $N(0, I_d)$. Then there exist some $C, C' > 0$, such that for any $t > 0$,

$$\mathbb{P}\left(\sup_{\{K: \|K\|_F \leq 1, \operatorname{rank}(K) \leq r\}} \left|\left\langle \frac{1}{n}\sum_{i=1}^n Z_i Z_i' - I_d, K \right\rangle\right| > t\right) \leq \exp(C'rd - Cn(t^2 \wedge t)).$$

The proofs of Theorem 5 and Lemma 7 are given in the supplementary material [15].

5.2. *Proof of Lemma* 1. Recall the definition of (S_u, S_v) and $(\widehat{S}_u, \widehat{S}_v)$ in Section 4.1. From here on, let

(41)
$$T_u = S_u \cup \widehat{S}_u \quad \text{and} \quad T_v = S_v \cup \widehat{S}_v.$$

The proof of Lemma 1 depends on the following two technical results. Their proofs are given in the supplementary material [15].

LEMMA 8. *For matrices* A, B, E, F *and a diagonal matrix* $D = (d_l)_{1 \leq l \leq r}$ *with* $d_1 \geq d_2 \geq \cdots \geq d_r > 0$ *and* $A'A = B'B = E'E = F'F = I_r$, *we have*

$$\frac{d_r}{2}\|AB' - EF'\|_F^2 \leq \langle ADB', AB' - EF' \rangle \leq \frac{d_1}{2}\|AB' - EF'\|_F^2.$$

LEMMA 9. *Under the assumption of Lemma* 1, *for any constant* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* M *and* C', *such that for any matrix* A *supported on the* $T_u \times T_v$, *we have*

$$C^{-1}\|A\|_F^2 \leq \|\widehat{\Sigma}_x^{1/2} A \widehat{\Sigma}_y^{1/2}\|_F^2 \leq C\|A\|_F^2,$$

with probability at least $1 - \exp(-C'k_q^u \log(ep/k_q^u)) - \exp(-C'k_q^v \log(em/k_q^v))$.

PROOF OF LEMMA 1. First of all, the triangle inequality and Jensen's inequality together lead to

(42)
$$\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_F^2$$
$$\leq 3(\|\widehat{U}_1^* \widehat{V}_1^{*'} - U_1^* V_1^{*'}\|_F^2 + \|\widehat{U}_1 \widehat{V}_1' - \widehat{U}_1^* \widehat{V}_1^{*'}\|_F^2 + \|U_1^* V_1^{*'} - U_1 V_1'\|_F^2).$$

Now, it remains to bound $\|\widehat{U}_1\widehat{V}_1' - \widehat{U}_1^*\widehat{V}_1^{*\prime}\|_F^2$. To this end, we have

$$\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\|_F^2$$

(43)
$$\leq C\|\widehat{\Sigma}_x^{1/2}(\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1')\widehat{\Sigma}_y^{1/2}\|_F^2$$

(44)
$$\leq \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_x^{1/2}\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2}(\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1')\widehat{\Sigma}_y^{1/2}\rangle$$

$$= \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

$$= \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y - \widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

$$+ \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

(45)
$$\leq \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y - \widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

(46)
$$= \frac{2C}{\lambda_r}\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y - \Sigma_x U_1\Lambda_1 V_1'\Sigma_y, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

$$- \frac{2C}{\lambda_r}\langle\Sigma_x U_2\Lambda_2 V_2'\Sigma_y, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle$$

$$+ \frac{2C}{\lambda_r}\langle\Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle.$$

Here, (43) is implied by Lemma 9 and (44) is implied by Lemma 8. To see (45), we note $(\widehat{U}_1, \widehat{V}_1)$ is the solution to (14), and so $\mathrm{Tr}(\widehat{U}_1'\widehat{\Sigma}_{xy}\widehat{V}_1) \geq \mathrm{Tr}((\widehat{U}_1^*)'\widehat{\Sigma}_{xy}\widehat{V}_1^*)$, or equivalently

$$\langle\widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1'\rangle \leq 0.$$

Equality (46) comes from the CCA structure (4) and (5). Combining (42)–(46) and rearranging the terms, we obtain the desired result. □

5.3. *Proof of Lemma* 2. The major difficulty in proving the lemma lies in the presence of the residual structure $U_2\Lambda_2 V_2'$ in (5) and the possible nondiagonality of covariance matrices Σ_x and Σ_y. To overcome the difficulty, we introduce intermediate matrices $(\widetilde{U}_1, \widetilde{V}_1)$ defined as follows. First, we write the SVD of $(\Sigma_{xS_uS_u})^{1/2}U_{1S_u*}\Lambda_1(V_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}$ as

(47)
$$(\Sigma_{xS_uS_u})^{1/2}U_{1S_u*}\Lambda_1(V_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2} = P\widetilde{\Lambda}_1 Q',$$

and let $\widetilde{U}_1^{S_u} = (\Sigma_{xS_uS_u})^{-1/2}P$ and $\widetilde{V}_1^{S_v} = (\Sigma_{yS_vS_v})^{-1/2}Q$. Finally, we define $\widetilde{U}_1 \in \mathbb{R}^{p\times r}$ and $\widetilde{V}_1 \in \mathbb{R}^{m\times r}$ by setting

(48) $(\widetilde{U}_1)_{S_u*} = \widetilde{U}_1^{S_u}, \qquad (\widetilde{U}_1)_{S_u^c*} = 0, \qquad (\widetilde{V}_1)_{S_v*} = \widetilde{V}_1^{S_v}, \qquad (\widetilde{V}_1)_{S_v^c*} = 0.$

By definition, we have $U_{1S_u*}\Lambda_1(V_{1S_u*})' = \widetilde{U}_{1S_u*}\widetilde{\Lambda}(\widetilde{V}_{1S_u*})'$. Last but not least, we define

$$
\begin{aligned}
\Xi = {} & P\widetilde{\Lambda}_1 Q' \\
& + (I - PP')(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xS_u*}U_2\Lambda_2 V_2'\Sigma_{y*S_v}(\Sigma_{yS_vS_v})^{-1/2}(I - QQ').
\end{aligned}
\tag{49}
$$

We now summarize the key properties of the $\widetilde{U}_1$, $\widetilde{V}_1$ and $\widetilde{\Lambda}_1$ matrices in the following two lemmas, the proofs of which are given in the supplementary material [15].

LEMMA 10. *Let P, Q and Ξ be defined in* (47) *and* (49). *Then we have*:

1. *The column vectors of P and Q are the r leading left and right singular vectors of Ξ.*
2. *The first and the rth singular values $\widetilde{\lambda}_1$ and $\widetilde{\lambda}_r$ of Ξ satisfy $1.1\kappa\lambda \geq \widetilde{\lambda}_1 \geq \widetilde{\lambda}_r \geq 0.9\lambda$, and the $(r+1)$th singular value $\widetilde{\lambda}_{r+1} \leq c\lambda$ for some sufficiently small constant $c > 0$.*
3. *The column vectors of $\Sigma_x^{1/2}\widetilde{U}_1$ and $\Sigma_y^{1/2}\widetilde{V}_1$ are the r leading left and right singular vectors of $\Sigma_x^{1/2}\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\Sigma_y^{1/2}$.*

LEMMA 11. *For some constant $C > 0$,*
$$
\|\widetilde{U}_1'\Sigma_x U_2\|_F^2 \leq C\|U_{1S_u^c*}\|_F^2 \quad and \quad \|\widetilde{V}_1'\Sigma_y V_2\|_F^2 \leq C\|V_{1S_v^c*}\|_F^2.
$$

In what follows, we prove claims (32) and (33) in order.

PROOF OF (32). By the triangle inequality,
$$
\|U_1^* V_1^{*\prime} - U_1 V_1'\|_F \leq \|U_1^* V_1^{*\prime} - \widetilde{U}_1\widetilde{V}_1'\|_F + \|\widetilde{U}_1\widetilde{V}_1' - U_1 V_1'\|_F.
\tag{50}
$$
It is sufficient to bound each of the two terms on the right-hand side.

$1°$ *Bound for $\|\widetilde{U}_1\widetilde{V}_1' - U_1 V_1'\|_F$*. Since the smallest eigenvalues of Σ_x and Σ_y are bounded from below by some absolute positive constant,
$$
\|\widetilde{U}_1\widetilde{V}_1' - U_1 V_1'\|_F \leq C\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{V}_1' - U_1 V_1')\Sigma_y^{1/2}\|_F.
$$
By Lemma 10, $\Sigma_x^{1/2}\widetilde{U}_1$ and $\Sigma_y^{1/2}\widetilde{V}_1$ collect the r leading left and right singular vectors of $\Sigma_x^{1/2}\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\Sigma_y^{1/2}$, and by (4), $\Sigma_x^{1/2}U_1$ and $\Sigma_y^{1/2}V_1$ collect the r leading left and right singular vectors of $\Sigma_x^{1/2}U_1\Lambda_1 V_1'\Sigma_y^{1/2}$. Thus, Theorem 5 implies
$$
\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{V}_1' - U_1 V_1')\Sigma_y^{1/2}\|_F \leq \frac{C}{\lambda}\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1 V_1')\Sigma_y^{1/2}\|_F.
$$
The right-hand side of the above inequality is bounded as
$$
\begin{aligned}
& \|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1 V_1'\|_F \\
& \quad \leq \|\widetilde{U}_{1S_u*}\widetilde{\Lambda}_1(\widetilde{V}_{1S_v*})' - U_{1S_u*}\Lambda_1(V_{1S_v*})'\|_F + \|U_{1S_u^c*}\Lambda_1(V_{1S_v*})'\|_F \\
& \quad\quad + \|U_{1S_u*}\Lambda_1(V_{1S_v^c*})'\|_F + \|U_{1S_u^c*}\Lambda_1(V_{1S_v^c*})'\|_F \\
& \quad \leq C\lambda\big(\|U_{1S_u^c*}\|_F + \|V_{1S_v^c*}\|_F\big).
\end{aligned}
\tag{51}
$$

Here, the last inequality is due to (47) and (48). For the last term, a similar argument to that used in Lemma 7 of [11] leads to

$$
\begin{aligned}
\|U_{1S_u^c*}\|_F^2 &\le \frac{Cq}{2-q} k_q^u (s_u/k_q^u)^{2/q} \le \frac{Cq}{2-q}\varepsilon_n^2, \\
\|V_{1S_v^c*}\|_F^2 &\le \frac{Cq}{2-q} k_q^v (s_v/k_q^v)^{2/q} \le \frac{Cq}{2-q}\varepsilon_n^2,
\end{aligned}
\tag{52}
$$

where the last inequalities in both displays are due to (11)–(13). Therefore, we obtain

$$
\|\widetilde{U}_1 \widetilde{V}_1' - U_1 V_1'\|_F^2 \le \frac{Cq}{2-q}\varepsilon_n^2.
\tag{53}
$$

2° *Bound for* $\|U_1^* V_1^{*\prime} - \widetilde{U}_1 \widetilde{V}_1'\|_F$. Let λ_{r+1}^* denote the $(r+1)$th singular value of $(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2}$. Then we have

$$
\begin{aligned}
&\|U_1^* V_1^{*\prime} - \widetilde{U}_1 \widetilde{V}_1'\|_F \\
&= \|U_{1S_u*}^*(V_{1S_v*}^*)' - \widetilde{U}_{1S_u*}(\widetilde{V}_{1S_v*})'\|_F \\
&\le C\|(\Sigma_{xS_uS_u})^{1/2}[U_{1S_u*}^*(V_{1S_v*}^*)' - \widetilde{U}_{1S_u*}(\widetilde{V}_{1S_v*})'](\Sigma_{yS_vS_v})^{1/2}\|_F \\
&\le \frac{C\|(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_F}{\widetilde{\lambda}_r - \lambda_{r+1}^*}.
\end{aligned}
\tag{54}
$$

Here, the first equality holds since both $U_1^* V_1^{*\prime}$ and $\widetilde{U}_1 \widetilde{V}_1'$ are supported on the $S_u \times S_v$ submatrix. Noting that by the discussion before (24), (48) and Lemma 10, $((\Sigma_{xS_uS_u})^{1/2}U_{1S_u*}^*, (\Sigma_{yS_vS_v})^{1/2}V_{1S_v*}^*)$ and $((\Sigma_{xS_uS_u})^{1/2}\widetilde{U}_{1S_u*}, (\Sigma_{yS_vS_v})^{1/2}\widetilde{V}_{1S_v*})$ collect the leading left and right singular vectors of $(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v} \times (\Sigma_{yS_vS_v})^{-1/2}$ and Ξ, respectively, we obtain the last inequality by applying (40) in Theorem 5. In what follows, we derive upper bound for the numerator and lower bound for the denominator in (54) in order.

Upper bound for $\|(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_F$. First, we decompose $\Sigma_{xyS_uS_v}$ as

$$
\begin{aligned}
\Sigma_{xyS_uS_v} &= \Sigma_{xS_u*}(U_1\Lambda_1 V_1' + U_2\Lambda_2 V_2')\Sigma_{y*S_v} \\
&= \Sigma_{xS_uS_u}U_{1S_u*}\Lambda_1 V_{1S_v*}'\Sigma_{yS_vS_v} + \Sigma_{xS_uS_u}U_{1S_u*}\Lambda_1 V_{1S_v^c*}'\Sigma_{yS_v^cS_v} \\
&\quad + \Sigma_{xS_uS_u^c}U_{1S_u^c*}\Lambda_1 V_1'\Sigma_{y*S_v} + \Sigma_{xS_u*}U_2\Lambda_2 V_2'\Sigma_{y*S_v}.
\end{aligned}
\tag{55}
$$

Then (55), (49) and (47) jointly imply that

$$
\begin{aligned}
&\|(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_F \\
&\le \|(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xS_uS_u^c}U_{1S_u^c*}\Lambda_1 V_1'\Sigma_{y*S_v}(\Sigma_{yS_vS_v})^{-1/2}\|_F \\
&\quad + \|(\Sigma_{xS_uS_u})^{1/2}U_{1S_u*}\Lambda_1 V_{1S_v^c*}'\Sigma_{yS_v^cS_v}(\Sigma_{yS_vS_v})^{-1/2}\|_F
\end{aligned}
$$

$$+ \left\| P P' (\Sigma_{x S_u S_u})^{-1/2} \Sigma_{x S_u *} U_2 \Lambda_2 V_2' \Sigma_{y * S_v} (\Sigma_{y S_v S_v})^{-1/2} (I - Q Q') \right\|_{\mathrm{F}}$$

$$+ \left\| (\Sigma_{x S_u S_u})^{-1/2} \Sigma_{x S_u *} U_2 \Lambda_2 V_2' \Sigma_{y * S_v} (\Sigma_{y S_v S_v})^{-1/2} Q Q' \right\|_{\mathrm{F}}$$

$$\leq C \lambda \big(\| U_{1 S_u^c *} \|_{\mathrm{F}} + \| V_{1 S_v^c *} \|_{\mathrm{F}} \big)$$

$$+ C \lambda_{r+1} \big(\| P' (\Sigma_{x S_u S_u})^{-1/2} \Sigma_{x S_u *} U_2 \|_{\mathrm{F}} + \| Q' (\Sigma_{y S_v S_v})^{-1/2} \Sigma_{y S_v *} V_2 \|_{\mathrm{F}} \big)$$

$$= C \lambda \big(\| U_{1 S_u^c *} \|_{\mathrm{F}} + \| V_{1 S_v^c *} \|_{\mathrm{F}} \big) + C \lambda_{r+1} \big(\| \widetilde{U}_1' \Sigma_x U_2 \|_{\mathrm{F}} + \| \widetilde{V}_1' \Sigma_y V_2 \|_{\mathrm{F}} \big).$$

Here, the last equality is due to the definition (48). The last display, together with (52) and Lemma 11, leads to

$$(56) \qquad \left\| (\Sigma_{x S_u S_u})^{-1/2} \Sigma_{xy S_u S_v} (\Sigma_{y S_v S_v})^{-1/2} - \Xi \right\|_{\mathrm{F}}^2 \leq \frac{C q}{2 - q} \lambda^2 \varepsilon_n^2.$$

Lower bound for $\widetilde{\lambda}_r - \lambda_{r+1}^*$. The bound (56), together with Weyl's inequality ([18], page 449 and Hoffman–Wielant inequality [31], page 63) implies

$$(57) \qquad \begin{aligned} & \left| \lambda_{r+1}^* - \widetilde{\lambda}_{r+1} \right| \vee \left\| \Lambda_1^* - \widetilde{\Lambda}_1 \right\|_{\mathrm{F}} \\ & \qquad \leq \left\| (\Sigma_{x S_u S_u})^{-1/2} \Sigma_{xy S_u S_v} (\Sigma_{y S_v S_v})^{-1/2} - \Xi \right\|_{\mathrm{F}} \leq C \sqrt{\frac{q}{2 - q}} \lambda \varepsilon_n \leq 0.1 \lambda. \end{aligned}$$

Together with Lemma 10, it further implies

$$(58) \qquad \widetilde{\lambda}_r - \lambda_{r+1}^* \geq \widetilde{\lambda}_r - \widetilde{\lambda}_{r+1} - \left| \widetilde{\lambda}_{r+1} - \lambda_{r+1}^* \right| \geq 0.7 \lambda.$$

Combining (54), (56) and (58), we obtain

$$(59) \qquad \left\| \widetilde{U}_1 \widetilde{V}_1' - U_1^* V_1^{*\prime} \right\|_{\mathrm{F}}^2 \leq \frac{C q}{2 - q} \varepsilon_n^2.$$

The proof of (32) is completed by combining (50), (53) and (59). $\quad\square$

PROOF OF (33).   Note that

$$\begin{aligned} & \left\| U_1^* \Lambda_1 V_1^{*\prime} - U_1 \Lambda_1 V_1' \right\|_{\mathrm{F}} \\ & \qquad \leq \left\| U_1^* \Lambda_1 V_1^{*\prime} - \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' \right\|_{\mathrm{F}} + \left\| \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' - U_1 \Lambda_1 V_1' \right\|_{\mathrm{F}} \\ & \qquad \leq \left\| U_1^* \Lambda_1^* V_1^{*\prime} - \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' \right\|_{\mathrm{F}} + \left\| \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' - U_1 \Lambda_1 V_1' \right\|_{\mathrm{F}} \\ & \qquad\quad + C \left\| \Lambda_1^* - \widetilde{\Lambda}_1 \right\|_{\mathrm{F}} + C \| \widetilde{\Lambda}_1 - \Lambda_1 \|_{\mathrm{F}} \\ & \qquad \leq \left\| U_1^* \Lambda_1^* V_1^{*\prime} - \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' \right\|_{\mathrm{F}} + C' \left\| \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' - U_1 \Lambda_1 V_1' \right\|_{\mathrm{F}} + C \left\| \Lambda_1^* - \widetilde{\Lambda}_1 \right\|_{\mathrm{F}}. \end{aligned}$$

Here, the last inequality is due to

$$(60) \qquad \| \widetilde{\Lambda}_1 - \Lambda_1 \|_{\mathrm{F}} \leq \left\| \Sigma_x^{1/2} \big(\widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' - U_1 \Lambda_1 V_1' \big) \Sigma_y^{1/2} \right\|_{\mathrm{F}},$$

a consequence of Lemma 10 and the Hoffman–Wielandt inequality [31], page 63.

We now control each of the three terms on the rightmost-hand side of the second last display. First, the bound we derived for (51), up to a constant multiplier, $\|\widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1' - U_1 \Lambda_1 V_1'\|_\mathrm{F}$ is upper bounded by the right-hand side of (33). Next, the bound for $\|\Lambda_1^* - \widetilde{\Lambda}_1\|_\mathrm{F}$ has been shown in (57). Last but not least, applying (39) in Theorem 5, we obtain

$$
\|U_1^* \Lambda_1^* V_1^{*'} - \widetilde{U}_1 \widetilde{\Lambda}_1 \widetilde{V}_1'\|_\mathrm{F}
$$

$$
\leq \frac{C(\widetilde{\lambda}_1 + \lambda_1^*)}{\widetilde{\lambda}_r - \lambda_{r+1}^*} \|(\Sigma_{x S_u S_u})^{-1/2} \Sigma_{xy S_u S_v} (\Sigma_{y S_v S_v})^{-1/2} - \Xi\|_\mathrm{F} \leq C \sqrt{\frac{q}{2-q}} \lambda \varepsilon_n,
$$

where the last inequality is due to (56), (57), (58) and Lemma 10. The proof is completed by assembling the bounds for the three terms. $\quad\square$

5.4. *Proof of Lemma 4.*   In this proof, we need the following technical result, which is a direct consequence of Lemma 3 in [15] by applying union bound. Remember the notation T_u and T_v defined in (41).

LEMMA 12.   *Assume* $\frac{1}{n}(k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v)) < C_1$ *for some constant* $C_1 > 0$. *For any constant* $C' > 0$, *there exists some constant* $C > 0$ *only depending on* M, C_1 *and* C', *such that*

$$
\|\widehat{\Sigma}_{x T_u T_u} - \Sigma_{x T_u T_u}\|_\mathrm{op}^2 \leq \frac{C}{n}(k_q^u \log(ep/k_q^u)),
$$

$$
\|\widehat{\Sigma}_{y T_v T_v} - \Sigma_{y T_v T_v}\|_\mathrm{op}^2 \leq \frac{C}{n}(k_q^v \log(em/k_q^v)),
$$

with probability at least $1 - \exp(-C' k_q^u \log(ep/k_q^u)) - \exp(-C' k_q^v \log(em/k_q^v))$.

In addition, we need the following result.

LEMMA 13 (Stewart and Sun [30], Theorem II.4.11).   *For any matrices* A, B *with* $A'A = B'B = I$, *we have*

$$
\inf_W \|A - BW\|_\mathrm{F} \leq \|AA' - BB'\|_\mathrm{F}.
$$

We first bound $\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1 \widehat{V}_1' \rangle$. By the definition of trace product, we have

$$
\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1 \widehat{V}_1' \rangle = \langle \Lambda_2 V_2' \Sigma_y \widehat{V}_1', U_2' \Sigma_x \widehat{U}_1 \rangle
$$

$$
\leq \|\Lambda_2 V_2' \Sigma_y \widehat{V}_1'\|_\mathrm{F} \|U_2' \Sigma_x \widehat{U}_1\|_\mathrm{F}
$$

$$
\leq \lambda_{r+1} \|V_2' \Sigma_y \widehat{V}_1'\|_\mathrm{F} \|U_2' \Sigma_x \widehat{U}_1\|_\mathrm{F}.
$$

Define the SVD of matrices U_1 and $\widehat{U}_1$ to be

$$
U_1 = \Theta R H', \qquad \widehat{U}_1 = \widehat{\Theta} \widehat{R} \widehat{H}'.
$$

For any matrix W, we have

$$\|\widehat{U}_1'\Sigma_x U_2\|_{\mathrm{F}} = \|(\widehat{U}_1 - U_1 H R^{-1} W \widehat{R}\widehat{H}')'\Sigma_x U_2\|_{\mathrm{F}}$$
$$\leq C\|\widehat{U}_1 - U_1 H R^{-1} W \widehat{R}\widehat{H}'\|_{\mathrm{F}}$$
$$\leq C\|\widehat{R}\|_{\mathrm{op}}\|\widehat{\Theta} - \Theta W\|_{\mathrm{F}},$$

where $\|\widehat{R}\|_{\mathrm{op}} \leq \|\widehat{U}_1\|_{\mathrm{op}} \leq \|(\widehat{\Sigma}_{x T_u T_u})^{-1/2}\|_{\mathrm{op}}\|(\widehat{\Sigma}_{x T_u T_u})^{1/2}\widehat{U}_{1 T_u *}\|_{\mathrm{op}} \leq C$ with probability at least $1 - \exp(-C'k_q^u \log(ep/k_q^u)) - \exp(-C'k_q^v \log(em/k_q^v))$ by Lemma 12. Hence, by Lemma 13, we have

$$(61) \qquad \|\widehat{U}_1'\Sigma_x U_2\|_{\mathrm{F}} \leq C \inf_W \|\widehat{\Theta} - \Theta W\|_{\mathrm{F}} \leq C\|\widehat{\Theta}\widehat{\Theta}' - \Theta\Theta'\|_{\mathrm{F}}.$$

We note that both $\widehat{\Theta}\widehat{\Theta}'$ and $\Theta\Theta'$ are the projection matrices of the left singular spaces of $\widehat{U}_1\widehat{V}_1'$ and $U_1 V_1'$, respectively, and the eigengap is at constant level since the rth singular value of $U_1 V_1'$ is bounded below by some constant and the $(r+1)$th singular value of $\widehat{U}_1\widehat{V}_1'$ is zero. Then a direct consequence of Wedin's sin-theta theorem [37] gives

$$(62) \qquad \|\widehat{\Theta}\widehat{\Theta}' - \Theta\Theta'\|_{\mathrm{F}} \leq C\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}.$$

Combining (61) and (62), we have $\|\widehat{U}_1'\Sigma_x U_2\|_{\mathrm{F}} \leq C_1\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}$. The same argument also implies $\|V_2'\Sigma_y \widehat{V}_1'\|_{\mathrm{F}} \leq C_1\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}$. Therefore,

$$|\langle \Sigma_x U_2 \Lambda_2 V_2'\Sigma_y, \widehat{U}_1\widehat{V}_1'\rangle| \leq C_2\lambda_{r+1}\|\widehat{U}_1\widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}^2.$$

Using a similar argument, we also obtain

$$|\langle \Sigma_x U_2 \Lambda_2 V_2'\Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)'\rangle| \leq C_2\lambda_{r+1}\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1 V_1'\|_{\mathrm{F}}^2.$$

By the triangle inequality, we complete the proof.

5.5. Proof of Lemma 5. Define

$$W = \begin{bmatrix} 0 & \widehat{U}_1^*\widehat{V}_1^{*'} - \widehat{U}_1\widehat{V}_1' \\ (\widehat{U}_1^*\widehat{V}_1^{*'} - \widehat{U}_1\widehat{V}_1')' & 0 \end{bmatrix}.$$

Then simple algebra leads to

$$(63) \qquad \langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*'} - \widehat{U}_1\widehat{V}_1'\rangle = \tfrac{1}{2}\langle \Sigma - \widehat{\Sigma}, W\rangle.$$

In the rest of the proof, we bound $\langle \Sigma - \widehat{\Sigma}, W\rangle$ by using Lemma 7.

Notice that the matrix $\widehat{U}_1^*\widehat{V}_1^{*'} - \widehat{U}_1\widehat{V}_1'$ has nonzero rows with indices in $T_u = S_u \cup \widehat{S}_u$ and nonzero columns with indices in $T_v = S_v \cup \widehat{S}_v$. Hence, the enlarged matrix W has nonzero rows and columns with indices in $T \times T$, where

$$T = T_u \cup (T_v + p)$$

with $T_v + p = \{j + p : j \in T_v\}$. The cardinality of T is $|T| = |T_u| + |T_v| \leq 2(k_q^u + k_q^v)$. Thus, we can rewrite (63) as

$$
\begin{aligned}
\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \rangle \\
&= \tfrac{1}{2} \langle \Sigma - \widehat{\Sigma}, W \rangle \\
&= \tfrac{1}{2} \langle \Sigma_{TT} - \widehat{\Sigma}_{TT}, W_{TT} \rangle \\
&= \tfrac{1}{2} \langle I_{|T|} - \Sigma_{TT}^{-1/2} \widehat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, \Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2} \rangle \\
&= \tfrac{1}{2} \| \Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2} \|_{\mathrm{F}} \langle I_{|T|} - \Sigma_{TT}^{-1/2} \widehat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle,
\end{aligned}
$$

where $K^T = \dfrac{\Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2}}{\| \Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2} \|_{\mathrm{F}}}$. Note that

$$
\tfrac{1}{2} \| \Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2} \|_{\mathrm{F}} \leq C \| W_{TT} \|_{\mathrm{F}} = C \| W \|_{\mathrm{F}} = \sqrt{2} C \| \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \|_{\mathrm{F}}.
$$

To obtain the desired bound, it suffices to show that

$$
(64) \qquad\qquad |\langle I_{|T|} - \Sigma_{TT}^{-1/2} \widehat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle|
$$

is upper bounded by $C \lambda \varepsilon_n$ with high probability.

To this end, we note that $T_u = S_u \cup \widehat{S}_u$ has at most $\binom{p}{k_q^u}$ different configurations since S_u is deterministic and $\widehat{S}_u$ is a random set of size k_q^u. For the same reason, T_v has at most $\binom{m}{k_q^v}$ different configurations. Therefore, the subset T has at most $N = \binom{p}{k_q^u} \binom{m}{k_q^v}$ different possible configurations, which can be listed as $T_1, T_2, \ldots, T_N$. Let

$$
K^{T_j} = \frac{\Sigma_{T_j T_j}^{1/2} W_{T_j T_j} \Sigma_{T_j T_j}^{1/2}}{\| \Sigma_{T_j T_j}^{1/2} W_{T_j T_j} \Sigma_{T_j T_j}^{1/2} \|_{\mathrm{F}}}
$$

for all $j \in [N]$. Since each $W_{T_j T_j}$ is of rank at most $2r$, so are the K^{T_j}'s. Therefore,

$$
\begin{aligned}
|(64)| &\leq \max_{1 \leq j \leq N} |\langle I_{|T_j|} - \Sigma_{T_j T_j}^{-1/2} \widehat{\Sigma}_{T_j T_j} \Sigma_{T_j T_j}^{-1/2}, K^{T_j} \rangle| \\
&\leq \max_{1 \leq j \leq N} \sup_{\| K \|_{\mathrm{F}} \leq 1, \mathrm{rank}(K) \leq 2r} |\langle I_{|T_j|} - \Sigma_{T_j T_j}^{-1/2} \widehat{\Sigma}_{T_j T_j} \Sigma_{T_j T_j}^{-1/2}, K \rangle|.
\end{aligned}
$$

Then the union bound leads to

$$
\mathbb{P}_\Sigma \big(|(64)| > t \big)
$$

$$
(65) \qquad \leq \sum_{j=1}^{N} \mathbb{P} \Big(\sup_{\| K \|_{\mathrm{F}} \leq 1, \mathrm{rank}(K) \leq 2r} |\langle I_{|T_j|} - \Sigma_{T_j T_j}^{-1/2} \widehat{\Sigma}_{T_j T_j} \Sigma_{T_j T_j}^{-1/2}, K \rangle| > t \Big)
$$

$$\leq \sum_{j=1}^{N} \exp(C'r|T_j| - Cn(t \wedge t^2))$$

$$\leq \binom{p}{k_q^u}\binom{m}{k_q^v} \exp(C_1 r(k_q^u + k_q^v) - Cn(t \wedge t^2))$$

$$\leq \exp\left(C_1 r(k_q^u + k_q^v) + k_q^u \log \frac{ep}{k_q^u} + k_q^v \log \frac{em}{k_q^v} - Cn(t \wedge t^2)\right),$$

where inequality (65) is due to Lemma 7. We complete the proof by choosing $t^2 = C_2\lambda^2\varepsilon_n^2$ in the last display for some sufficiently large constant $C_2 > 0$, which, by condition (16), is bounded.

5.6. *Proof of Lemma* 6. First, we apply a telescoping expansion to the quantity of interest as

$$\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \widehat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1' \rangle$$

$$(66) \qquad = \langle \Sigma_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \Sigma_y - \Sigma_x U_1^* \Lambda_1 V_1^{*\prime} \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1' \rangle$$

$$(67) \qquad + \langle \Sigma_x U_1^* \Lambda_1 V_1^{*\prime} \Sigma_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1' \rangle$$

$$(68) \qquad + \langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \widehat{\Sigma}_y - \Sigma_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1' \rangle.$$

In what follows, we bound each of the terms in (66)–(68) in order.

1° *Bound for* (66). Applying (35) in Lemma 3, we obtain that with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$,

$$|(66)| \leq C\|\widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} - U_1^* \Lambda_1 V_1^{*\prime}\|_{\mathrm{F}} \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}}$$

$$\leq C\sqrt{\frac{q}{2-q}} \lambda\varepsilon_n \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}}.$$

2° *Bound for* (67). Applying (33) in Lemma 2, we obtain

$$|(67)| \leq C\|U_1^* \Lambda_1 V_1^{*\prime} - U_1 \Lambda_1 V_1'\|_{\mathrm{F}} \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}}$$

$$\leq C\sqrt{\frac{q}{2-q}} \lambda\varepsilon_n \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}}.$$

3° *Bound for* (68). We turn to bound (68) based on a strategy similar to that used in proving Lemma 5. First, we write it in a form for which we could apply Lemma 7. Recall the random sets T_u and T_v defined in (41). Then for

$$H_x^{T_u} = (\Sigma_{xT_uT_u})^{1/2}(\widehat{U}_{1T_u*}^*(\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*}(\widehat{V}_{1T_v*})')$$

$$\times \widehat{\Sigma}_{yT_vT_v} \widehat{V}_{1T_v*}^* \Lambda_1 (\widehat{U}_{1T_u*}^*)' (\Sigma_{xT_uT_u})^{1/2},$$

$$H_y^{T_v} = (\Sigma_{yT_vT_v})^{1/2} \widehat{V}_{1T_v*}^* \Lambda_1 (\widehat{U}_{1T_u*}^*)'$$

$$\times \Sigma_{xT_uT_u}(\widehat{U}_{1T_u*}^*(\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*}(\widehat{V}_{1T_v*})')(\Sigma_{yT_vT_v})^{1/2},$$

and $\overline{H}_x^{T_u} = H_x^{T_u}/\|H_x^{T_u}\|_F$, $\overline{H}_y^{T_v} = H_y^{T_v}/\|H_y^{T_v}\|_F$, we have

$$
\begin{aligned}
&|(68)| \\
&= |\langle \widehat{\Sigma}_x - \Sigma_x, (\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1') \widehat{\Sigma}_y \widehat{V}_1^* \Lambda_1 \widehat{U}_1^{*\prime} \rangle \\
&\quad + \langle \widehat{\Sigma}_y - \Sigma_y, \widehat{V}_1^* \Lambda_1 \widehat{U}_1^{*\prime} \Sigma_x (\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1') \rangle| \\
&\leq |\langle \widehat{\Sigma}_{x T_u T_u} - \Sigma_{x T_u T_u}, (\widehat{U}_{1 T_u *}^* (\widehat{V}_{1 T_v *}^*)' \\
&\quad - \widehat{U}_{1 T_u *} (\widehat{V}_{1 T_v *})') \widehat{\Sigma}_{y T_v T_v} \widehat{V}_{1 T_v *}^* \Lambda_1 (\widehat{U}_{1 T_u *}^*)' \rangle| \\
&\quad + |\langle \widehat{\Sigma}_{y T_v T_v} \\
&\quad - \Sigma_{y T_v T_v}, \widehat{V}_{1 T_v *}^* \Lambda_1 (\widehat{U}_{1 T_u *}^*)' \Sigma_{x T_u T_u} (\widehat{U}_{1 T_u *}^* (\widehat{V}_{1 T_v *}^*)' - \widehat{U}_{1 T_u *} (\widehat{V}_{1 T_v *})') \rangle| \\
&= \|H_x^{T_u}\|_F |\langle (\Sigma_{x T_u T_u})^{-1/2} \widehat{\Sigma}_{x T_u T_u} (\Sigma_{x T_u T_u})^{-1/2} - I_{|T_u|}, \overline{H}_x^{T_u} \rangle| \\
&\quad + \|H_y^{T_v}\|_F |\langle (\Sigma_{y T_v T_v})^{-1/2} \widehat{\Sigma}_{y T_v T_v} (\Sigma_{y T_v T_v})^{-1/2} - I_{|T_v|}, \overline{H}_y^{T_v} \rangle|.
\end{aligned}
$$

We now bound each term on the rightmost side. Applying Lemma 7 with union bound and then following a similar analysis to that leading to (64) but with T replaced by T_u and T_v, we obtain that

$$
\begin{aligned}
&\left| \langle (\Sigma_{x T_u T_u})^{-1/2} \widehat{\Sigma}_{x T_u T_u} (\Sigma_{x T_u T_u})^{-1/2} - I_{|T_u|}, \overline{H}_x^{T_u} \rangle \right| \\
&\leq C \sqrt{\frac{k_q^u}{n} \left(r + \log \frac{ep}{k_q^u} \right)}, \\
(69) \quad\quad &\left| \langle (\Sigma_{y T_v T_v})^{-1/2} \widehat{\Sigma}_{y T_v T_v} (\Sigma_{y T_v T_v})^{-1/2} - I_{|T_v|}, \overline{H}_y^{T_v} \rangle \right| \\
&\leq C \sqrt{\frac{k_q^v}{n} \left(r + \log \frac{em}{k_q^v} \right)}
\end{aligned}
$$

with probability at least $1 - \exp(-C' k_q^u (r + \log(ep/k_q^u)))$ and $1 - \exp(-C' k_q^v (r + \log(em/k_q^v)))$, respectively.

To bound $\|H_x^{T_u}\|_F$ and $\|H_y^{T_v}\|_F$, we note that it follows from Lemma 12 that all eigenvalues of $\widehat{\Sigma}_{x T_u T_u}$ and $\widehat{\Sigma}_{y T_v T_v}$ are bounded from below and above by some universal positive constants with probability at least $1 - \exp(-C' k_q^u \log(ep/k_q^u)) - \exp(-C' k_q^v \log(em/k_q^v))$ under assumption (16). Thus, with the same probability we have

$$
\begin{aligned}
(70) \quad \|H_x^{T_u}\|_F &\leq C\lambda \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_F \|\widehat{\Sigma}_{y T_v T_v}^{1/2} \widehat{V}_{1 T_v *}^*\|_{op} \\
&\quad \times \|\widehat{\Sigma}_{y T_v T_v}^{1/2}\|_{op} \|\widehat{\Sigma}_{x T_u T_u}^{1/2} \widehat{U}_{1 T_u *}^*\|_{op} \|\widehat{\Sigma}_{x T_u T_u}^{-1/2}\|_{op} \\
&\leq C_1 \lambda \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1'\|_F
\end{aligned}
$$

and

$$\|H_y^{T_v}\|_{\mathrm{F}} \leq C\lambda \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1{}'\|_{\mathrm{F}} \|\widehat{\Sigma}_{yT_vT_v}^{1/2} \widehat{V}_{1T_v*}^*\|_{\mathrm{op}}$$

$$(71) \qquad\qquad \times \|\widehat{\Sigma}_{yT_vT_v}^{-1/2}\|_{\mathrm{op}} \|\widehat{\Sigma}_{xT_uT_u}^{1/2} \widehat{U}_{1T_u*}^*\|_{\mathrm{op}} \|\widehat{\Sigma}_{xT_uT_u}^{-1/2}\|_{\mathrm{op}}$$

$$\leq C_1\lambda \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1{}'\|_{\mathrm{F}}.$$

Combining (69), (70) and (71), we obtain

$$|(68)| \leq C\lambda^2 \varepsilon_n \|\widehat{U}_1^* \widehat{V}_1^{*\prime} - \widehat{U}_1 \widehat{V}_1{}'\|_{\mathrm{F}},$$

with probability at least $1 - \exp(-C'k_q^u \log(ep/k_q^u)) - \exp(-C'k_q^v \log(em/k_q^v))$. Noting that $\lambda < 1$, this completes the proof.

SUPPLEMENTARY MATERIAL

Supplement to "Minimax estimation in sparse canonical correlation analysis" (DOI: 10.1214/15-AOS1332SUPP; .pdf). The supplement [15] contains an Appendix to the current paper in which we prove Theorems 3–5 and Lemmas 3 and 7–11.

REFERENCES

[1] AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[2] ANDERSON, T. W. (1999). Asymptotic theory for canonical correlation analysis. *J. Multivariate Anal.* **70** 1–29. MR1701396

[3] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[4] AVANTS, B. B., COOK, P. A., UNGAR, L., GEE, J. C. and GROSSMAN, M. (2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage* **50** 1004–1016.

[5] BAO, Z., HU, G., PAN, G. and ZHOU, W. (2014). Canonical correlation coefficients of high-dimensional normal vectors: Finite rank case. Preprint. Available at arXiv:1407.7194.

[6] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.* **30** 1–21.

[7] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[8] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129

[9] BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

[10] CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

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

[12] CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

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

[14] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

[15] GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Supplement to "Minimax estimation in sparse canonical correlation analysis." DOI:10.1214/15-AOS1332SUPP.

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

[17] GAO, C. and ZHOU, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** 785–818. MR3325710

[18] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[19] HARDOON, D. R. and SHAWE-TAYLOR, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* **83** 331–353. MR3108214

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

[21] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[22] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010

[23] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

[24] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[25] LÊ CAO, K.-A., MARTIN, P. G. P., ROBERT-GRANIÉ, C. and BESSE, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics* **10** 1–34.

[26] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

[27] MA, Z. and WU, Y. (2013). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. Preprint. Available at arXiv:1306.3609.

[28] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

[29] PARKHOMENKO, E., TRITCHLER, D. and BEYENE, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* **8** Art. 1, 36. MR2471148

[30] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

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

[32] TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 611–622. MR1707864

[33] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452

[34] WAAIJENBORG, S. and ZWINDERMAN, A. H. (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics* **10** 315.

[35] WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. Preprint. Available at arXiv:1408.5369.

[36] WANG, Y. X. R., JIANG, K., FELDMAN, L. J., BICKEL, P. J. and HUANG, H. (2014). Inferring gene association networks using sparse canonical correlation analysis. Preprint. Available at arXiv:1401.6504.

[37] WEDIN, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968

[38] WIESEL, A., KLIGER, M. and HERO, A. O. III (2008). A greedy approach to sparse canonical correlation analysis. Preprint. Available at arXiv:0801.2748.

[39] WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

[40] YANG, D., MA, Z. and BUJA, A. (2011). A sparse SVD method for high-dimensional data. Preprint. Available at arXiv:1112.2433.

[41] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

C. GAO
H. H. ZHOU
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: chao.gao@yale.edu
        huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68

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

H. REN
DEPARTMENT OF STATISTICS
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15260
USA
E-MAIL: zren@pitt.edu
URL: http://www.pitt.edu/~zren