Submitted to the Annals of Statistics

MINIMAX ESTIMATION IN SPARSE CANONICAL CORRELATION ANALYSIS

BY CHAO GAO, * ZONGMING MA, † ZHAO REN, * AND HARRISON H. ZHOU *

Yale University, University of Pennsylvania, Yale University and Yale University

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 non-asymptotic 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) [17] is one of the most classical and important tools in multivariate statistics [3, 24]. 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

*The research of Chao Gao, Zhao Ren and Harrison H. Zhou is supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

†The research of Zongming Ma is supported in part by the Claude Marion Endowed Faculty Scholar Award of the Wharton School.

AMS 2000 subject classifications: Primary 62H12; secondary 62C20

Keywords and phrases: Covariance matrix, Minimax rates, Model selection, Nuisance parameter, Sin-theta theorem, Sparse CCA (SCCA).

specifically, (u_1, v_1) is the solution to the following optimization problem,

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

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

$$\mathrm{Cov}(u'X, u_j'X) = \mathrm{Cov}(v'Y, v_j'Y) = 0, \quad \text{for } \ j = 1, \ldots, i.$$

Throughout the paper, we call the (u_i, v_i)'s canonical correlation directions. It was shown by Hotelling [17] that the $(\Sigma_x^{1/2}u_i, \Sigma_y^{1/2}v_i)$'s are the successive singular vector pairs of

$$\Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2}, \tag{2}$$

where $\Sigma_x = \mathrm{Cov}(X), \Sigma_y = \mathrm{Cov}(Y)$ and $\Sigma_{xy} = \mathrm{Cov}(X, Y)$. When one is only given a random sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ 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 [25]. 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. 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 substantially better estimator than the sample covariance matrix [23] which usually behaves poorly [18]. 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, [36, 37, 26, 16, 21, 33, 4, 34] 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 [12]. 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 reparametrize 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. Under this model, Chen et al. [12] 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 [22]. 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 [12], Σ_{xy} can be reparametrized as follows,

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

where $\bar{r} = \min(p, m)$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{\bar{r}})$ 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 $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \Lambda_2 = \mathrm{diag}(\lambda_{r+1}, \ldots, \lambda_{\bar{r}})$, $U_1 \in \mathbb{R}^{p\times r}$, $V_1 \in \mathbb{R}^{m\times r}$, $U_2 \in \mathbb{R}^{p\times r_2}$ 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 [12]. 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_1V_1' under the following loss function

$$L(U_1V_1', \widehat{U_1V_1'}) = \|U_1V_1' - \widehat{U_1V_1'}\|_{\mathrm{F}}^2. \tag{6}$$

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_1V_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 [35], any upper bound on the loss function (6) leads to an upper bound on the loss functions $\|P_{U_1} - \widehat{P}_{U_1}\|_{\mathrm{F}}^2$ and $\|P_{V_1} - \widehat{P}_{V_1}\|_{\mathrm{F}}^2$ for estimating the column subspaces of U_1 and V_1, which have been used in the related problem of sparse principal component analysis [9, 32]. 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 j-th 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 $q \in [0, 2)$,

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

under the tradition that $0^q = 0$. For instance, in the case of exact sparsity, i.e., $q = 0$, $\|U_1\|_{0,w}$ 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 $\|U_1\|_{q,w} \le s_u$ and $\|V_1\|_{q,w} \le s_v$;
2. $\left\|\Sigma_x^l\right\|_{\text{op}} \vee \left\|\Sigma_y^l\right\|_{\text{op}} \le M$ for $l = \pm 1$;
3. $1 > \kappa\lambda \ge \lambda_1 \ge ... \ge \lambda_r \ge \lambda > 0$.

Throughout the paper, we assume $\kappa\lambda \le 1 - c_0$ 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 [12], where $\text{rank}(\Sigma_{xy}) = 1$, in this paper, the rank of Σ_{xy} can be as high as p or m, r is allowed to grow, and we do not need structural assumptions on Σ_x and Σ_y such as sparsity.

Suppose we observe i.i.d. pairs $(X_1, Y_1), \dots (X_n, Y_n) \sim N_{p+m}(0, \Sigma)$. 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 non-asymptotic minimax rates for estimating the leading canonical directions when $q = 0$:

$$\begin{aligned} &\inf_{\widehat{U_1V_1'}} \sup_{\Sigma\in\mathcal{F}_0(s_u,s_v,p,m,r,\lambda)} \mathbb{E}_\Sigma \|U_1V_1' - \widehat{U_1V_1'}\|_{\text{F}}^2 \\ \asymp \;& \frac{1}{n\lambda^2}\left(r(s_u+s_v) + s_u \log\frac{\text{e}p}{s_u} + s_v\log\frac{\text{e}m}{s_v}\right), \end{aligned} \tag{8}$$

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

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 4 in Section 4.1. 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 [9]. We bypass sample splitting by establishing a new empirical process bound for the supreme of Gaussian quadratic forms with rank constraint. See Lemma 8 in Section 4.1. The estimator is shown to be minimax rate optimal by establishing matching minimax lower bounds based on a local metric entropy approach [20, 7, 38, 9].

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 [30, 18] 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

$$\Sigma_0 = V \Lambda V' + \sigma^2 I_p. \tag{9}$$

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 [19] proposed a diagonal thresholding estimator of the sparse principal eigenvector which is provably consistent when $r = 1$ in (9). For fixed r, Birnbaum et al. [8] derived minimax rate optimal estimators for individual sparse principal eigenvectors, and Ma [22] proposed to directly estimate sparse principal subspaces, i.e., 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. [9] studied minimax rates and adaptive estimation for sparse principal subspaces with little constraint on r. See also [32] 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, 10, 5, 14] 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, i.e., the columns of V in (9) are normalized with respect to the identity matrix. In contrast, in sparse CCA, the sparse vectors of interest, i.e., 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 $U_2\Lambda_2V_2'$, which is motivated by the situation statisticians face in practice. It is highly non-trivial 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 non-sparse matrices $\Sigma_x^{1/2}U_1$ and $\Sigma_y^{1/2}V_1$.

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. All the proofs are gathered in Section 3, with some auxiliary results and technical lemmas proved in Section 4 and the appendix.

1.4. *Notation.* For any matrix $A = (a_{ij})$, the i-th row of A is denoted by A_{i*} and the j-th column by A_{*j}. For a positive integer p, $[p]$ denotes the index set $\{1, 2, ..., p\}$. 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 i-th largest singular value. The Frobenius norm and the operator norm of A are defined as $\|A\|_{\mathrm{F}} = \sqrt{\mathsf{Tr}(A'A)}$ and $\|A\|_{\mathrm{op}} = \sigma_1(A)$, respectively. The support of A is defined as $\text{supp}(A) = \{i \in [n] : \|A_{i*}\| > 0\}$. 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 [17] can be written in the following optimization form,

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

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

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

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

The effective sparsity of U_1 and V_1 are defined as

$$k_q^u = \lceil x_q^u \rceil, \quad k_q^v = \lceil x_q^v \rceil. \tag{13}$$

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,

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

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 [17] with additional l_0 constraints $\|A\|_{0,w} = s_u$ and $\|B\|_{0,w} = s_v$.

Set

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

which is the minimax rate to be shown later.

THEOREM 1. *Under the assumption that*

$$\epsilon_n^2 \leq c, \tag{16}$$

$$\lambda_{r+1} \leq c\lambda, \tag{17}$$

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 $\Sigma \in \mathcal{F}_q$,

$$\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}^2 \leq C\epsilon_n^2,$$

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 the assumption (16) is necessary for consistent estimation. The assumption (17) implies $\lambda_{r+1} \leq c\lambda_r$ for $c \in (0,1)$, such that the eigengap is lower bounded as $\lambda_r - \lambda_{r+1} \geq (1-c)\lambda_r > 0$.

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

To obtain the convergence rate in expectation, we propose a modified estimator. The modification is inspired by the fact that U_1V_1' are bounded in Frobenius norm, because

$$\|U_1V_1'\|_{\mathrm{F}} \leq \|\Sigma_x^{-1/2}\|_{\mathrm{op}}\|\Sigma_x^{1/2}U_1\|_{\mathrm{F}}\|\Sigma_y^{1/2}V_1\|_{\mathrm{op}}\|\Sigma_y^{-1/2}\|_{\mathrm{op}} \leq M\sqrt{r}. \tag{18}$$

Define $\widehat{U_1V_1'}$ to be the truncated version of $\widehat{U}_1\widehat{V}_1'$ as

$$\widehat{U_1V_1'} = \widehat{U}_1\widehat{V}_1'\,\mathbf{1}_{\left\{\|\widehat{U}_1\widehat{V}_1'\|_{\mathrm{F}} \leq 2M\sqrt{r}\right\}}.$$

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

$$\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}} \geq \|\widehat{U}_1\widehat{V}_1'\|_{\mathrm{F}} - \|U_1V_1'\|_{\mathrm{F}} \geq M\sqrt{r} \geq \left\|0 - U_1V_1'\right\|_{\mathrm{F}}.$$

Then it is better to estimate U_1V_1' by 0.

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

$$\exp(C_1(k_q^u + \log(ep/k_q^u))) \quad > \quad n\lambda^2, \tag{19}$$

$$\exp(C_1(k_q^v + \log(em/k_q^v))) \quad > \quad n\lambda^2, \tag{20}$$

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

$$\sup_{\Sigma \in \mathcal{F}_q} \mathbb{E}_\Sigma \|\widehat{U_1V_1'} - U_1V_1'\|_{\mathrm{F}}^2 \leq C_2\epsilon_n^2.$$

REMARK 3. 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.* Theorem 1 and Theorem 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* $1 \le r \le \frac{k_q^u \wedge k_q^v}{2}$, *and that*

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

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*

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

REMARK 4. The assumption (21) is necessary for consistent estimation.

3. Proof of Main Results.

3.1. *Proof of upper bounds.* In this part, we prove Theorems 1 and 2.

3.1.1. *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

$$\begin{aligned} &\max_{(A,B)} \mathsf{Tr}(A' \Sigma_{xy} B) \\ &\text{s.t. } A'\Sigma_x A = B' \Sigma_y B = I_r \text{ and } \mathrm{supp}(A) \subset S_u, \mathrm{supp}(B) \subset S_v. \end{aligned} \tag{22}$$

In what follows, we refer to them as the *sparse approximations* to U_1 and V_1. By definition, when $q = 0$, $U_1^*(V_1^*)' = U_1 V_1'$.

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

$$\begin{aligned}
&\max_{(A,B)} \mathsf{Tr}(A'\widehat{\Sigma}_{xy}B) \\
&\text{s.t. } A'\widehat{\Sigma}_x A = B'\widehat{\Sigma}_y B = I_r \text{ and } \mathrm{supp}(A) = S_u, \mathrm{supp}(B) = S_v.
\end{aligned} \tag{23}$$

In case the program (22) (or (23)) has multiple global optimizers, we define (U_1^*, V_1^*) (or $(\widehat{U}_1^*, \widehat{V}_1^*)$) by picking an arbitrary one.

We note that

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

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 $\widehat{\Sigma}_y$.

Last but not least, let

$$\widehat{S}_u = \mathrm{supp}(\widehat{U}_1), \qquad \widehat{S}_v = \mathrm{supp}(\widehat{V}_1). \tag{24}$$

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).

3.1.2. *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*

$$\begin{aligned}
&\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}^2 \\
\le\ & 3\|U_1^*(V_1^*)' - U_1V_1'\|_{\mathrm{F}}^2 && (25)\\
&+3\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1^*(V_1^*)'\|_{\mathrm{F}}^2 && (26)\\
&-\frac{6C}{\lambda_r}\left\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1' \right\rangle && (27)\\
&+\frac{6C}{\lambda_r}\left\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1' \right\rangle && (28)\\
&+\frac{6C}{\lambda_r}\left\langle \widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y - \Sigma_x U_1\Lambda_1 V_1'\Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1\widehat{V}_1' \right\rangle, && (29)
\end{aligned}$$

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 4.2. □

In particular, Lemma 1 decomposes the total loss into the sum of the sparse approximation error in (25), the oracle loss in (26) which is present even if we have the oracle knowledge of the effective support sets S_u and S_v, the bias term in (27) caused by the presence of the residual term $U_2\Lambda_2V_2'$ in the CCA structure (4), and the two excess loss terms in (28) and (29) resulting from the uncertainty about the effective support sets. When $q = 0$, the sparse approximation error term (25) vanishes.

3.1.3. *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 subsections 4.3 – 4.7.

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

$$\|U_1^*(V_1^*)' - U_1V_1'\|_{\mathrm{F}}^2 \leq \frac{Cq}{2-q}\,\epsilon_n^2, \tag{30}$$

$$\|U_1^*\Lambda_1(V_1^*)' - U_1\Lambda_1V_1'\|_{\mathrm{F}}^2 \leq \frac{Cq}{2-q}\lambda^2\epsilon_n^2. \tag{31}$$

LEMMA 3 (Oracle loss). *Suppose $\frac{1}{n\lambda^2}\Big(k_q^u + k_q^v + \log(ep/k_q^u) + \log(em/k_q^v)\Big) < c$ 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*

$$\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1^*(V_1^*)'\|_{\mathrm{F}}^2 \leq \frac{Cr}{n\lambda^2}\left[k_q^u + k_q^v + \log\left(\frac{ep}{k_q^u}\right) + \log\left(\frac{em}{k_q^v}\right)\right], \tag{32}$$

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

$$\|\widehat{U}_1^*\Lambda_1(\widehat{V}_1^*)' - U_1^*\Lambda_1(V_1^*)'\|_{\mathrm{F}}^2 \leq C\lambda^2\epsilon_n^2. \tag{33}$$

Since $r \leq k_q^u \wedge k_q^v$, (32) is bounded above by $C\epsilon_n^2$. 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

$$\begin{aligned}&\left|\left\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)' - \widehat{U}_1 \widehat{V}_1' \right\rangle\right| \\ &\quad \leq C\lambda_{r+1}\left(\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1V_1'\|_{\mathrm{F}}^2 + \|U_1V_1' - \widehat{U}_1\widehat{V}_1'\|_{\mathrm{F}}^2\right),\end{aligned}$$

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 is Lemma 4 is 0 when $U_2\Lambda_2V_2'$ 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*

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

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*

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

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

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

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

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, Lemma 2 and Lemma 3 imply that

$$\theta^2 \leq C\epsilon_n^2 \quad \text{and} \quad \delta^2 \leq C\epsilon_n^2.$$

Moreover, Lemma 4 implies

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

Lemma 5 implies

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

and Lemma 6 implies

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

Together with Lemma 1, the above bounds lead to

$$\begin{aligned} R^2 &\leq C(\theta^2 + \delta^2) + \frac{C\lambda_{r+1}}{\lambda}(R^2 + \theta^2 + \delta^2) + C\epsilon_n (R + \theta + \delta) \\ &\leq \frac{C\lambda_{r+1}}{\lambda} R^2 + C\epsilon_n R + C\epsilon_n^2. \end{aligned}$$

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

$$R^2 \leq C\epsilon_n^2,$$

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

3.1.5. *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

$$\begin{aligned} &\mathbb{E}_\Sigma \|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \\ &\leq \mathbb{E}_\Sigma \left[\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \mathbf{1}_{\left\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \leq C\epsilon_n^2\right\}} \right] \\ &\quad + \mathbb{E}_\Sigma \left[\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \mathbf{1}_{\left\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 > C\epsilon_n^2\right\}} \right] \end{aligned}$$

$$\leq C\epsilon_n^2 + \mathbb{E}_\Sigma \left[\left(2\|\widehat{U_1 V_1'}\|_{\mathrm{F}}^2 + 2\|U_1 V_1'\|_{\mathrm{F}}^2 \right) \mathbf{1}_{\left\{\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 > C\epsilon_n^2\right\}} \right] \tag{34}$$

$$\leq C\epsilon_n^2 + 6M^2 r\, \mathbb{P}_\Sigma \left(\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}^2 > C\epsilon_n^2 \right) \tag{35}$$

$$\leq C_2 \epsilon_n^2. \tag{36}$$

Here, the inequality (34) is due to the triangle inequality and the fact that

$$\left\{ \|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 > C\epsilon_n^2 \right\} \subset \left\{ \|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}^2 > C\epsilon_n^2 \right\}.$$

In fact, if $\|\widehat{U}_1 \widehat{V}_1' - U_1 V_1'\|_{\mathrm{F}}^2 \leq C\epsilon_n^2$, then $\|\widehat{U}_1 \widehat{V}_1'\|_{\mathrm{F}}^2 \leq C\epsilon_n^2 + M^2 r \leq 2M^2 r$. By our definition of the estimator, this means $\widehat{U_1 V_1'} = \widehat{U}_1 \widehat{V}_1'$, which further implies $\|\widehat{U_1 V_1'} - U_1 V_1'\|_{\mathrm{F}}^2 \leq C\epsilon_n^2$. The inequality (35) follows from our definition of estimator $\widehat{U_1 V_1'}$ and (18). The last inequality follows from the conclusion of Theorem 1 and the assumptions (19) and (20). This completes the proof.

3.2. *Proof of lower bounds.* In this part, we prove Theorem 3. We divide the proof into two parts. In the first part, we establish the desired lower bounds for the exact sparse case $q = 0$ in Section 3.2.1. In the second part, we extend the arguments to the approximate space case $q \in (0, 2)$ in Section 3.2.2. Without loss of generality, we assume $r \leq (p - k_q^u + 1) \wedge (m - k_q^v + 1)$.

Throughout the proof, we focus on the special case where $U_2 = 0$ and $V_2 = 0$ in (5). Thus, we omit the subscript 1 in U_1, Λ_1 and V_1 in the rest of the proof.

3.2.1. *The case of $q = 0$.* We first present a lemma on the Kullback-Leibler divergence between data distributions generated by a special kind of covariance matrices. The lemma also partially explains why (6) is a natural loss function to consider. Its proof is deferred to the appendix.

LEMMA 7. *For $i = 1, 2$, let $\Sigma_{(i)} = \begin{bmatrix} I_p & \lambda U_{(i)} V_{(i)}' \\ \lambda V_{(i)} U_{(i)}' & I_m \end{bmatrix}$ with $\lambda \in (0, 1)$, $U_{(i)} \in O(p, r)$ and $V_{(i)} \in O(m, r)$. Let $P_{(i)}$ denote the distribution of a random i.i.d. sample of size n from the $N_{p+m}(0, \Sigma_{(i)})$ distribution. Then*

$$D(P_{(1)}||P_{(2)}) = \frac{n\lambda^2}{2(1-\lambda^2)} \|U_{(1)}V_{(1)}' - U_{(2)}V_{(2)}'\|_{\mathrm{F}}^2.$$

The main tool for our proof is Fano's lemma, which is based on multiple hypothesis testing argument. The following version of Fano's lemma is from [39, Lemma 3].

PROPOSITION 1. *Let (Θ, ρ) be a metric space and $\{P_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote by $\mathcal{M}(T, \rho, \epsilon)$ the ϵ-packing number of T with respect to ρ, i.e., the maximal number of points in T whose pairwise minimum distance in ρ is at least ϵ. Define the* Kullback-Leibler diameter *of T by*

$$d_{\mathrm{KL}}(T) \triangleq \sup_{\theta, \theta' \in T} D(P_\theta \,||\, P_{\theta'}). \tag{37}$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\rho^2(\hat{\theta}(X), \theta)] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}\right). \tag{38}$$

PROOF OF THEOREM 3 (CASE $q = 0$). Note that in this case, $k_0^u = s_u$ and $k_0^v = s_v$.

1° We establish first the term involving $r(s_u + s_v)$. To this end, let $U_0 = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in O(p, r)$ and $V_0 = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in O(m, r)$. For some $\epsilon_0 \in (0, \sqrt{r \wedge (s_u - r)}]$ to be specified later, let

$$B(\epsilon_0) = \{U \in O(p, r) : \mathrm{supp}(U) \subset [s_u], \|U - U_0\|_{\mathrm{F}} \le \epsilon_0\}.$$

and

$$T_0 = \left\{ \Sigma = \begin{bmatrix} I_p & \lambda U V_0' \\ \lambda V_0 U' & I_m \end{bmatrix} : U \in B(\epsilon_0) \right\}.$$

It is straightforward to verify that $T_0 \subset \mathcal{F}_0$. By Lemma 7,

$$\begin{aligned} d_{\mathrm{KL}}(T_0) &= \sup_{U_{(i)} \in B(\epsilon_0)} \frac{n\lambda^2}{2(1-\lambda^2)} \|U_{(1)} V_0' - U_{(2)} V_0'\|_{\mathrm{F}}^2 \\ &= \sup_{U_{(i)} \in B(\epsilon_0)} \frac{n\lambda^2}{2(1-\lambda^2)} \|U_{(1)} - U_{(2)}\|_{\mathrm{F}}^2 = \frac{2n\lambda^2\epsilon_0^2}{1-\lambda^2}. \end{aligned} \tag{39}$$

Here, the second equality is due to the definition of V_0 and the third due to the definition of $B(\epsilon_0)$.

We now establish a lower bound for the packing number of T_0. For some $\alpha \in (0, 1)$ to be specified later, let $\{\widetilde{U}_{(1)}, \ldots, \widetilde{U}_{(N)}\} \subset O(p, r)$ be a maximal set such that $\mathrm{supp}(\widetilde{U}_i) \subset [s_u]$, and for any $i \neq j \in [N]$,

$$\|\widetilde{U}_{(i)} \widetilde{U}_{(i)}' - U_0 U_0'\|_{\mathrm{F}} \le \epsilon_0, \qquad \|\widetilde{U}_{(i)} \widetilde{U}_{(i)}' - \widetilde{U}_{(j)} \widetilde{U}_{(j)}'\|_{\mathrm{F}} \ge \sqrt{2}\alpha\epsilon_0. \tag{40}$$

Then by [9, Lemma 1], for some absolute constant $C > 1$,

$$N \ge \left(\frac{1}{C\alpha} \right)^{r(s_u - r)}.$$

For each $\widetilde{U}_{(i)}$, define

$$U_{(i)} = \widetilde{U}_{(i)} O_{(i)}, \quad \text{for } O_{(i)} = \operatorname*{argmin}_{O \in O(r)} \|\widetilde{U}_{(i)} O - U_0\|_{\mathrm{F}}. \tag{41}$$

Then for any $i \in [N]$, by definition, $U_{(i)} \in O(p, r)$, $\mathrm{supp}(U_{(i)}) \subset [s_u]$, and $U_{(i)} U_{(i)}' = \widetilde{U}_{(i)} \widetilde{U}_{(i)}'$. In addition, [28, Theorem II.4.11] implies

$$\|U_{(i)} - U_0\|_{\mathrm{F}} \le \|\widetilde{U}_{(i)} \widetilde{U}_{(i)}' - U_0 U_0'\|_{\mathrm{F}} \le \epsilon_0,$$

and so $U_{(i)} \in B(\epsilon_0)$. On the other hand, note that $\|\widetilde{U}_{(i)}\widetilde{U}_{(i)}' - \widetilde{U}_{(j)}\widetilde{U}_{(j)}'\|_{\mathrm{F}} \leq \sqrt{2}\|\widetilde{U}_{(i)} - \widetilde{U}_{(j)}\|_{\mathrm{F}}$, and hence for $i \neq j \in [N]$,

$$\|U_{(i)} - U_{(j)}\|_{\mathrm{F}} \geq \frac{1}{\sqrt{2}}\|U_{(i)}U_{(i)}' - U_{(j)}U_{(j)}'\|_{\mathrm{F}} = \frac{1}{\sqrt{2}}\|\widetilde{U}_{(i)}\widetilde{U}_{(i)}' - \widetilde{U}_{(j)}\widetilde{U}_{(j)}'\|_{\mathrm{F}} \geq \alpha\epsilon_0.$$

Let $\rho(\Sigma_{(1)}, \Sigma_{(2)}) = \|U_{(1)}V_0' - U_{(2)}V_0'\|_{\mathrm{F}} = \|U_{(1)} - U_{(2)}\|_{\mathrm{F}}$. Then the foregoing argument implies that for $\epsilon = \alpha\epsilon_0$,

$$\log \mathcal{M}(T_0, \rho, \epsilon) \geq r(s_u - r) \log \frac{1}{C\alpha}. \tag{42}$$

Setting $\epsilon = c_0[\sqrt{(r \wedge (s_u - r))} \wedge \sqrt{\frac{1-\lambda^2}{n\lambda^2} r(s_u - r)}]$ for a sufficiently small absolute constant c_0 and also setting $\alpha > 0$ to be a sufficiently small absolute constant, we obtain a lower bound of order

$$r \wedge (s_u - r) \wedge \frac{1 - \lambda^2}{n\lambda^2} r(s_u - r)$$

by applying Proposition 1 with (39) and (42). By symmetry, we also have the above lower bound with $r(s_u - r)$ replaced by $r(s_v - r)$. Noting that λ is bounded away from 1 and that $r \leq \frac{1}{2}(s_u \wedge s_v)$, we obtain the lower bound of order

$$\frac{r(s_u + s_v)}{n\lambda^2} \wedge r.$$

2° We turn to establishing the desired lower bound involving $s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v}$, which can be obtained from the rank-one argument spelled out in [12]. Without loss of generality, we may assume $s_u \leq \frac{p}{2}$ and $s_v \leq \frac{m}{2}$. To be rigorous, consider the following subset of the parameter space:

$$T_1 = \Bigg\{\Sigma = \begin{bmatrix} I_p & \lambda U V_0' \\ \lambda V_0 U' & I_m \end{bmatrix} : U = \begin{bmatrix} I_{r-1} & 0 \\ 0 & u_r \end{bmatrix}, \\ u_r \in \mathbb{S}^{p-r+1}, |\mathrm{supp}(u_r)| \leq s_u - r + 1\Bigg\}.$$

Restricting on the set T_1, the minimax risk for estimating UV' is the same as the minimax risk for estimating u_r under the squared error loss $\|u_r - \widehat{u}_r\|_{\mathrm{F}}^2$. Let $X = [X_1\ X_2]$ with $X_1 \in \mathbb{R}^{n\times(r-1)}$ and $X_2 \in \mathbb{R}^{n\times(p-r+1)}$, and $Y = [Y_1\ Y_2]$ with $Y_1 \in \mathbb{R}^{n\times(r-1)}$ and $Y_2 \in \mathbb{R}^{n\times(m-r+1)}$. Then it is further equivalent to estimating u_1 under squared error loss based on the observation (X_2, Y_2), because (X_2, Y_2) is a sufficient statistic for u_r. Applying the argument in [12, Appendix G], we obtain the lower bound of order

$$\frac{1}{n\lambda^2}\left(s_u \log\frac{ep}{s_u}\right) \wedge 1.$$

By symmetry, the same lower bound holds if we replace p and s_u by m and s_v. This completes the proof. □

3.2.2. *The case of $q \in (0, 2)$.*

Proof of Theorem 3 (Case $q \in (0, 2)$). 1° As in the case of $q = 0$, we first establish a lower bound of order

$$\frac{r(k_q^u + k_q^v)}{n\lambda^2} \wedge r.$$

Following the lines in the proof of [9, Theorem 2], we can find a collection of $\{\widetilde{U}_{(1)}, \ldots, \widetilde{U}_{(N)}\} \subset O(p, r)$ such that (40) holds for

$$\epsilon_0 = \sqrt{r \wedge (k_q^u - r) \wedge \frac{1 - \lambda^2}{2n\lambda^2} r k_q^u},$$

that $\|\widetilde{U}_{(i)}\|_{q,w} \le s_u$, and that for some absolute constant C, $N \ge \left(\frac{1}{C\alpha}\right)^{rk_q^u/2}$. For each $i \in [N]$, set $U_{(i)} = \widetilde{U}_{(i)} O_{(i)}$, which is defined in (41). Then that $O_{(i)} \in O(r)$ and [9, Eq.(110)] implies that $\|U_{(i)}\|_{q,w} = \|\widetilde{U}_{(i)}\|_{q,w} \le s_u$. The rest of the argument then follows that in Section 3.2.1.

2° Next, we establish a lower bound of order

$$\frac{1}{n\lambda^2}\left(k_q^u \log\frac{ep}{k_q^u} + k_q^v \log\frac{em}{k_q^v}\right) \wedge 1.$$

To this end, we apply the same reduction argument as in Section 3.2.1, and the argument in [12, Appendix G] leads to the desired claim. □

4. Proof of Auxiliary Results. In this section, we prove Lemmas 1 – 6 used in the proof of Theorem 1 and 2. Throughout the section, without further notice, ϵ_n^2 is defined as in (15).

4.1. *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 [6, 28]. In particular, both Frobenius norm $\|\cdot\|_{\mathrm{F}}$ and operator norm $\|\cdot\|_{\mathrm{op}}$ are unitarily invariant.

THEOREM 4. *Consider matrices $X, Y \in \mathbb{R}^{p\times m}$. Let the SVD of X and Y be*

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

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

$$\|A_1 D_1 B_1' - \widehat{A}_1 \widehat{D}_1 \widehat{B}_1'\| \le \left(\frac{\sqrt{2}(d_1 + \widehat{d}_1)}{\delta} + 1\right)\epsilon. \tag{43}$$

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

$$\|A_1 B_1' - \widehat{A}_1 \widehat{B}_1'\| \le \frac{C\epsilon}{\delta}. \tag{44}$$

REMARK 5. 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 [13] $\|A_1 A_1' - \widehat{A}_1 \widehat{A}_1'\| \le C\epsilon/\delta$ up to a constant multiplier.

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

LEMMA 8. *Let $\{Z_i\}_{1\le i\le n}$ be i.i.d. observations from $N(0, I_d)$. Then, there exist some $C, C' > 0$, such that for any $t > 0$,*

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

4.2. *Proof of Lemma 1.* Recall the definition of (S_u, S_v) and $(\widehat{S}_u, \widehat{S}_v)$ in Section 3.1.1. From here on, let

$$T_u = S_u \cup \widehat{S}_u \quad \text{and} \quad T_v = S_v \cup \widehat{S}_v. \tag{45}$$

The proof of Lemma 1 depends on the following two technical results. For their proofs, see the appendix.

LEMMA 9. *For matrices A, B, E, F and a diagonal matrix $D = (d_l)_{1\le l\le r}$ with $d_1 \ge d_2 \ge ... \ge d_r > 0$ and $A'A = B'B = E'E = F'F = I_r$, we have*

$$\frac{d_r}{2}\|AB' - EF'\|_{\mathrm{F}}^2 \le \langle ADB', AB' - EF'\rangle \le \frac{d_1}{2}\|AB' - EF'\|_{\mathrm{F}}^2.$$

LEMMA 10. *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*

$$C^{-1}\|A\|_{\mathrm{F}}^2\le\|\widehat{\Sigma}_x^{1/2}A\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2\le C\|A\|_{\mathrm{F}}^2,$$

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

$$\begin{aligned}&\|\widehat{U}_1\widehat{V}_1'-U_1V_1'\|_{\mathrm{F}}^2\\ &\le 3\left(\|\widehat{U}_1^*\widehat{V}_1^{*\prime}-U_1^*V_1^{*\prime}\|_{\mathrm{F}}^2+\|\widehat{U}_1\widehat{V}_1'-\widehat{U}_1^*\widehat{V}_1^{*\prime}\|_{\mathrm{F}}^2+\|U_1^*V_1^{*\prime}-U_1V_1'\|_{\mathrm{F}}^2\right).\end{aligned}\tag{46}$$

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

$$\begin{aligned}&\|\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\|_{\mathrm{F}}^2\\
&\le C\|\widehat{\Sigma}_x^{1/2}(\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1')\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 &(47)\\
&\le\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_x^{1/2}\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y^{1/2},\widehat{\Sigma}_x^{1/2}(\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1')\widehat{\Sigma}_y^{1/2}\right\rangle &(48)\\
&=\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y,\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle\\
&=\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y-\widehat{\Sigma}_{xy},\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle\\
&\quad+\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_{xy},\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle\\
&\le\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y-\widehat{\Sigma}_{xy},\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle &(49)\\
&=\frac{2C}{\lambda_r}\left\langle\widehat{\Sigma}_x\widehat{U}_1^*\Lambda_1\widehat{V}_1^{*\prime}\widehat{\Sigma}_y-\Sigma_xU_1\Lambda_1V_1'\Sigma_y,\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle &(50)\\
&\quad-\frac{2C}{\lambda_r}\left\langle\Sigma_xU_2\Lambda_2V_2'\Sigma_y,\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle\\
&\quad+\frac{2C}{\lambda_r}\left\langle\Sigma_{xy}-\widehat{\Sigma}_{xy},\widehat{U}_1^*\widehat{V}_1^{*\prime}-\widehat{U}_1\widehat{V}_1'\right\rangle.\end{aligned}$$

Here, (47) is implied by Lemma 10 and (48) is implied by Lemma 9. To see (49), we note $(\widehat{U}_1,\widehat{V}_1)$ is the solution to (14), and so $\mathsf{Tr}(\widehat{U}_1'\widehat{\Sigma}_{xy}\widehat{V}_1)\ge\mathsf{Tr}((\widehat{U}_1^*)'\widehat{\Sigma}_{xy}\widehat{V}_1^*)$, or equivalently

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

The equality (50) comes from the CCA structure (4) and (5). Combining (46)-(50) and rearranging the terms, we obtain the desired result. ☐

4.3. *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

$$(\Sigma_{xS_uS_u})^{1/2}U_{1S_u*}\Lambda_1(V_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2} = P\widetilde{\Lambda}_1 Q', \tag{51}$$

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

$$(\widetilde{U}_1)_{S_u*} = \widetilde{U}_1^{S_u},\ (\widetilde{U}_1)_{S_u^c*} = 0, \quad (\widetilde{V}_1)_{S_v*} = \widetilde{V}_1^{S_v},\ (\widetilde{V}_1)_{S_v^c*} = 0. \tag{52}$$

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

$$\Xi = P\widetilde{\Lambda}_1 Q' + (I-PP')(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xS_u*}U_2\Lambda_2V_2'\Sigma_{y*S_v}(\Sigma_{yS_vS_v})^{-1/2}(I-QQ'). \tag{53}$$

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 deferred to the appendix.

LEMMA 11. *Let P, Q and Ξ be defined in (51) and (53). 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 r-th 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 12. *For some constant $C > 0$,*

$$\|\widetilde{U}_1'\Sigma_x U_2\|_{\mathrm{F}}^2 \leq C\|U_{1S_u^c*}\|_{\mathrm{F}}^2 \quad \textit{and} \quad \|\widetilde{V}_1'\Sigma_y V_2\|_{\mathrm{F}}^2 \leq C\|V_{1S_v^c*}\|_{\mathrm{F}}^2.$$

In what follows, we prove claims (30) and (31) in order.

PROOF OF (30). By triangle inequality,

$$\|U_1^*V_1^{*\prime} - U_1V_1'\|_{\mathrm{F}} \leq \|U_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{V}_1'\|_{\mathrm{F}} + \|\widetilde{U}_1\widetilde{V}_1' - U_1V_1'\|_{\mathrm{F}}. \tag{54}$$

It is sufficient to bound each of the two terms on the right side.

1° *Bound for* $\|\widetilde{U}_1\widetilde{V}_1' - U_1V_1'\|_{\mathrm{F}}$. Since the smallest eigenvalues of Σ_x and Σ_y are bounded from below by some absolute positive constant,

$$\|\widetilde{U}_1\widetilde{V}_1' - U_1V_1'\|_{\mathrm{F}} \le C\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{V}_1' - U_1V_1')\Sigma_y^{1/2}\|_{\mathrm{F}}.$$

By Lemma 11, $\Sigma_x^{1/2}\widetilde{U}_1$ 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 $\Sigma_x^{1/2}U_1\Lambda_1V_1'\Sigma_y^{1/2}$. Thus, Theorem 4 implies

$$\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{V}_1' - U_1V_1')\Sigma_y^{1/2}\|_{\mathrm{F}} \le \frac{C}{\lambda}\|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1')\Sigma_y^{1/2}\|_{\mathrm{F}}.$$

The right side of the above inequality is bounded as

$$\begin{aligned} &\|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1'\|_{\mathrm{F}} \\ \le\ & \|\widetilde{U}_{1S_u*}\widetilde{\Lambda}_1(\widetilde{V}_{1S_v*})' - U_{1S_u*}\Lambda_1(V_{1S_v*})'\|_{\mathrm{F}} + \|U_{1S_u^c*}\Lambda_1(V_{1S_v*})'\|_{\mathrm{F}} \\ & + \|U_{1S_u*}\Lambda_1(V_{1S_v^c*})'\|_{\mathrm{F}} + \|U_{1S_u^c*}\Lambda_1(V_{1S_v^c*})'\|_{\mathrm{F}} \\ \le\ & C\lambda(\|U_{1S_u^c*}\|_{\mathrm{F}} + \|V_{1S_v^c*}\|_{\mathrm{F}}). \end{aligned} \tag{55}$$

Here, the last inequality is due to (51) and (52). For the last term, a similar argument to that used in Lemma 7 of [9] leads to

$$\begin{aligned} \|U_{1S_u^c*}\|_{\mathrm{F}}^2 &\le \frac{Cq}{2-q}k_q^u(s_u/k_q^u)^{2/q} \le \frac{Cq}{2-q}\epsilon_n^2, \\ \|V_{1S_v^c*}\|_{\mathrm{F}}^2 &\le \frac{Cq}{2-q}k_q^v(s_v/k_q^v)^{2/q} \le \frac{Cq}{2-q}\epsilon_n^2, \end{aligned} \tag{56}$$

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

$$\|\widetilde{U}_1\widetilde{V}_1' - U_1V_1'\|_{\mathrm{F}}^2 \le \frac{Cq}{2-q}\epsilon_n^2. \tag{57}$$

2° *Bound for* $\|U_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{V}_1'\|_{\mathrm{F}}$. 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

$$\begin{aligned} &\|U_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{V}_1'\|_{\mathrm{F}} \\ &= \|U_{1S_u*}^*(V_{1S_v*}^*)' - \widetilde{U}_{1S_u*}(\widetilde{V}_{1S_v*})'\|_{\mathrm{F}} \\ &\le C\|(\Sigma_{xS_uS_u})^{1/2}[U_{1S_u*}^*(V_{1S_v*}^*)' - \widetilde{U}_{1S_u*}(\widetilde{V}_{1S_v*})'](\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}} \\ &\le \frac{C\|(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_{\mathrm{F}}}{\widetilde{\lambda}_r - \lambda_{r+1}^*}. \end{aligned} \tag{58}$$

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 (22), (52) and Lemma 11, $((\Sigma_{xS_uS_u})^{1/2} U^*_{1S_u*}, (\Sigma_{yS_vS_v})^{1/2} V^*_{1S_v*})$ 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 $(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xyS_uS_v} (\Sigma_{yS_vS_v})^{-1/2}$ and Ξ respectively, we obtain the last inequality by applying (44) in Theorem 4. In what follows, we derive upper bound for the numerator and lower bound for the denominator in (58) in order.

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

$$
\begin{aligned}
\Sigma_{xyS_uS_v} &= \Sigma_{xS_u*}(U_1 \Lambda_1 V_1' + U_2 \Lambda_2 V_2') \Sigma_{y*S_v} \\
&= \Sigma_{xS_uS_u} U_{1S_u*} \Lambda_1 V'_{1S_v*} \Sigma_{yS_vS_v} + \Sigma_{xS_uS_u} U_{1S_u*} \Lambda_1 V'_{1S_v^c*} \Sigma_{yS_v^cS_v} \\
&\quad + \Sigma_{xS_uS_u^c} U_{1S_u^c*} \Lambda_1 V_1' \Sigma_{y*S_v} + \Sigma_{xS_u*} U_2 \Lambda_2 V_2' \Sigma_{y*S_v}.
\end{aligned} \tag{59}
$$

Then (59), (53) and (51) jointly imply that

$$
\begin{aligned}
&\|(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xyS_uS_v} (\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_{\mathrm{F}} \\
&\leq \|(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xS_uS_u^c} U_{1S_u^c*} \Lambda_1 V_1' \Sigma_{y*S_v} (\Sigma_{yS_vS_v})^{-1/2}\|_{\mathrm{F}} \\
&\quad + \|(\Sigma_{xS_uS_u})^{1/2} U_{1S_u*} \Lambda_1 V'_{1S_v^c*} \Sigma_{yS_v^cS_v} (\Sigma_{yS_vS_v})^{-1/2}\|_{\mathrm{F}} \\
&\quad + \|PP'(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xS_u*} U_2 \Lambda_2 V_2' \Sigma_{y*S_v} (\Sigma_{yS_vS_v})^{-1/2} (I - QQ')\|_{\mathrm{F}} \\
&\quad + \|(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xS_u*} U_2 \Lambda_2 V_2' \Sigma_{y*S_v} (\Sigma_{yS_vS_v})^{-1/2} QQ'\|_{\mathrm{F}} \\
&\leq C\lambda(\|U_{1S_u^c*}\|_{\mathrm{F}} + \|V_{1S_v^c*}\|_{\mathrm{F}}) \\
&\quad + C\lambda_{r+1}(\|P'(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xS_u*} U_2\|_{\mathrm{F}} + \|Q'(\Sigma_{yS_vS_v})^{-1/2} \Sigma_{yS_v*} V_2\|_{\mathrm{F}}) \\
&= C\lambda(\|U_{1S_u^c*}\|_{\mathrm{F}} + \|V_{1S_v^c*}\|_{\mathrm{F}}) + C\lambda_{r+1}(\|\widetilde{U}_1' \Sigma_x U_2\|_{\mathrm{F}} + \|\widetilde{V}_1' \Sigma_y V_2\|_{\mathrm{F}}).
\end{aligned}
$$

Here, the last equality is due to the definition (52). The last display, together with (56) and Lemma 12, leads to

$$
\|(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xyS_uS_v} (\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_{\mathrm{F}}^2 \leq \frac{Cq}{2-q} \lambda^2 \epsilon_n^2. \tag{60}
$$

Lower bound for $\widetilde{\lambda}_r - \lambda^*_{r+1}$. The bound (60), together with Weyl's inequality [15, p.449] and Hoffman-Wielant inequality [29, p.63] implies

$$
\begin{aligned}
&|\lambda^*_{r+1} - \widetilde{\lambda}_{r+1}| \vee \|\Lambda_1^* - \widetilde{\Lambda}_1\|_{\mathrm{F}} \\
&\leq \|(\Sigma_{xS_uS_u})^{-1/2} \Sigma_{xyS_uS_v} (\Sigma_{yS_vS_v})^{-1/2} - \Xi\|_{\mathrm{F}} \leq C\sqrt{\frac{q}{2-q}} \lambda \epsilon_n \leq 0.1\lambda.
\end{aligned} \tag{61}
$$

Together with Lemma 11, it further implies

$$\widetilde{\lambda}_r - \lambda^*_{r+1} \geq \widetilde{\lambda}_r - \widetilde{\lambda}_{r+1} - |\widetilde{\lambda}_{r+1} - \lambda^*_{r+1}| \geq 0.7\lambda. \tag{62}$$

Combining (58), (60) and (62), we obtain

$$\|\widetilde{U}_1\widetilde{V}_1' - U_1^*V_1^{*\prime}\|_{\mathrm{F}}^2 \leq \frac{Cq}{2-q}\epsilon_n^2. \tag{63}$$

The proof of (30) is completed by combining (54), (57) and (63). □

PROOF OF (31). Note that

$$\begin{aligned}
&\|U_1^*\Lambda_1V_1^{*\prime} - U_1\Lambda_1V_1'\|_{\mathrm{F}} \\
&\leq \|U_1^*\Lambda_1V_1^{*\prime} - \widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\|_{\mathrm{F}} + \|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1'\|_{\mathrm{F}} \\
&\leq \|U_1^*\Lambda_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\|_{\mathrm{F}} + \|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1'\|_{\mathrm{F}} \\
&\quad + C\|\Lambda_1^* - \widetilde{\Lambda}_1\|_{\mathrm{F}} + C\|\widetilde{\Lambda}_1 - \Lambda_1\|_{\mathrm{F}} \\
&\leq \|U_1^*\Lambda_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\|_{\mathrm{F}} + C'\|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1'\|_{\mathrm{F}} + C\|\Lambda_1^* - \widetilde{\Lambda}_1\|_{\mathrm{F}}.
\end{aligned}$$

Here the last inequality is due to

$$\|\widetilde{\Lambda}_1 - \Lambda_1\|_{\mathrm{F}} \leq \|\Sigma_x^{1/2}(\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1')\Sigma_y^{1/2}\|_{\mathrm{F}}, \tag{64}$$

a consequence of Lemma 11 and Hoffman-Wielandt inequality [29, p.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 (55), up to a constant multiplier, $\|\widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1' - U_1\Lambda_1V_1'\|_{\mathrm{F}}$ is upper bounded by the righthand side of (31). Next, the bound for $\|\Lambda_1^* - \widetilde{\Lambda}_1\|_{\mathrm{F}}$ has been shown in (61). Last but not least, applying (43) in Theorem 4, we obtain

$$\begin{aligned}
&\|U_1^*\Lambda_1^*V_1^{*\prime} - \widetilde{U}_1\widetilde{\Lambda}_1\widetilde{V}_1'\|_{\mathrm{F}} \\
&\leq \frac{C(\widetilde{\lambda}_1 + \lambda_1^*)}{\widetilde{\lambda}_r - \lambda^*_{r+1}}\|\left(\Sigma_{xS_uS_u}\right)^{-1/2}\Sigma_{xyS_uS_v}\left(\Sigma_{yS_vS_v}\right)^{-1/2} - \Xi\|_{\mathrm{F}} \leq C\sqrt{\frac{q}{2-q}}\lambda\epsilon_n\,,
\end{aligned}$$

where the last inequality is due to (60), (61), (62) and Lemma 11. The proof is completed by assembling the bounds for the three terms. □

4.4. *Proof of Lemma 3.* The proof relies on the following lemma, which is a simple consequence of Propositions $D.1$ and $D.2$ in [22].

LEMMA 13. *Assume* $\frac{1}{n}(k_q^u + k_q^v + \log(ep/k_q^u) + \log(em/k_q^v)) < C_1$ *for some constant* $C_1 > 0$. *Consider deterministic sets* $B_u \subset [p]$ *and* $B_v \subset [m]$ *with* $|B_u| = C_u k_q^u$ *and* $|B_v| = C_v k_q^v$ *for any constants* $C_u, C_v > 0$. *For any constant* $C' > 0$, *there exists* $C > 0$ *only depending on* M, C_1, C_u, C_v *and* C', *such that,*

$$\begin{aligned}
\|\widehat{\Sigma}_{xB_uB_u} - \Sigma_{xB_uB_u}\|_{\mathrm{op}}^2 &\leq \frac{C}{n}(k_q^u + \log(ep/k_q^u)),\\
\|\widehat{\Sigma}_{yB_vB_v} - \Sigma_{yB_vB_v}\|_{\mathrm{op}}^2 &\leq \frac{C}{n}(k_q^v + \log(em/k_q^v)),\\
\|\widehat{\Sigma}_{xyB_uB_v} - \Sigma_{xyB_uB_v}\|_{\mathrm{op}}^2 &\leq \frac{C}{n}(k_q^v + k_q^u + \log(ep/k_q^u) + \log(em/k_q^v)),
\end{aligned}$$

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 13 and Lemma 2.2 in [27] (see also Lemma 16 in appendix), we obtain the following concentration inequalities for square-roots of covariance matrices:

$$\|(\widehat{\Sigma}_{xS_uS_u})^{1/2} - (\Sigma_{xS_uS_u})^{1/2}\|_{\mathrm{op}}^2 \leq \frac{C}{n}(k_q^u + \log(ep/k_q^u)), \tag{65}$$

$$\|(\widehat{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{op}}^2 \leq \frac{C}{n}(k_q^v + \log(em/k_q^v)), \tag{66}$$

with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$. Furthermore, under the condition of Lemma 3, there exists some constant $C_1 > 0$ such that with the same probability,

$$\|\widehat{\Sigma}_{xS_uS_u}^i\|_{\mathrm{op}} \vee \|\widehat{\Sigma}_{yS_vS_v}^i\|_{\mathrm{op}} \leq C_1, \qquad i = \pm 1. \tag{67}$$

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

PROOF OF (32). Since both $\widehat{U}_1^*\widehat{V}_1^{*\prime}$ and $U_1^*V_1^{*\prime}$ are supported on the $S_u \times S_v$ submatrix, we have

$$\begin{aligned}
&\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - U_1^*V_1^{*\prime}\|_{\mathrm{F}} = \|\widehat{U}_{1S_u*}^*(\widehat{V}_{1S_v*}^*)' - U_{1S_u*}^*(V_{1S_v*}^*)'\|_{\mathrm{F}}\\
&\quad \leq C\|(\Sigma_{xS_uS_u})^{1/2}[\widehat{U}_{1S_u*}^*(\widehat{V}_{1S_v*}^*)' - U_{1S_u*}^*(V_{1S_v*}^*)'](\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}.
\end{aligned}$$

By the triangle inequality, up to a constant multiplier, the rightmost hand side of the last display is further upper bounded by the sum of the following two terms:

$$\begin{aligned}
&\|(\Sigma_{xS_uS_u})^{1/2}\widehat{U}_{1S_u*}^*(\widehat{V}_{1S_v*}^*)'(\Sigma_{yS_vS_v})^{1/2}\\
&\qquad - (\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}_{1S_u*}^*(\widehat{V}_{1S_v*}^*)'(\widehat{\Sigma}_{yS_vS_v})^{1/2}\|_{\mathrm{F}},
\end{aligned} \tag{68}$$

and

$$\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'(\widehat{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{xS_uS_u})^{1/2}U^*_{1S_u*}(V^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}. \tag{69}$$

In what follows, we show that up to constant multipliers, both (68) and (69) are upper bounded by the rate in (32).

1° *Bound for (68).* By the triangle inequality,

$$\begin{aligned}(68) &\le \|[(\Sigma_{xS_uS_u})^{1/2} - (\widehat{\Sigma}_{xS_uS_u})^{1/2}]\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}} \\ &\quad + \|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'[(\Sigma_{yS_vS_v})^{1/2} - (\widehat{\Sigma}_{yS_vS_v})^{1/2}]\|_{\mathrm{F}} \\ &\le \|\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}\|(\Sigma_{xS_uS_u})^{1/2} - (\widehat{\Sigma}_{xS_uS_u})^{1/2}\|_{\mathrm{op}} \\ &\quad + \|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'\|_{\mathrm{F}}\|(\Sigma_{yS_vS_v})^{1/2} - (\widehat{\Sigma}_{yS_vS_v})^{1/2}\|_{\mathrm{op}}.\end{aligned}$$

To further bound the rightmost side of the last display, we note that with high probability

$$\begin{aligned}\|\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}} &\le \|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}\|_{\mathrm{F}}\|(\widehat{\Sigma}_{xS_uS_u})^{-1/2}\|_{\mathrm{op}} \\ &\quad \times \|(\widehat{\Sigma}_{yS_vS_v})^{-1/2}\|_{\mathrm{op}}\|(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{op}} \\ &\le C\sqrt{r}.\end{aligned}$$

Together with (65), this implies that with high probability,

$$\begin{aligned}&\|\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}\|(\Sigma_{xS_uS_u})^{1/2} - (\widehat{\Sigma}_{xS_uS_u})^{1/2}\|_{\mathrm{op}} \\ &\qquad \le C\sqrt{\frac{r}{n}\left(k_q^u + k_q^v + \log\frac{ep}{k_q^u} + \log\frac{em}{k_q^v}\right)}.\end{aligned}$$

By a similar argument, $\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}(\widehat{V}^*_{1S_v*})'\|_{\mathrm{F}}\|(\Sigma_{yS_vS_v})^{1/2} - (\widehat{\Sigma}_{yS_vS_v})^{1/2}\|_{\mathrm{op}}$ satisfies the same upper bound. Thus, (68) is upper bounded by the rate in (32) with the desired probability by noting that $\lambda \le 1$.

2° *Bound for (69).* By definition, $(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}$ and $(\widehat{\Sigma}_{yS_vS_v})^{1/2}\widehat{V}^*_{1S_v*}$ collect the r leading left and right singular vectors of $(\widehat{\Sigma}_{xS_uS_u})^{-1/2}\widehat{\Sigma}_{xyS_uS_v}(\widehat{\Sigma}_{yS_vS_v})^{-1/2}$, and $(\Sigma_{xS_uS_u})^{1/2}U^*_{1S_u*}$ and $(\Sigma_{yS_vS_v})^{1/2}V^*_{1S_v*}$ collect the r leading left and right singular vectors of $(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2}$. To apply Theorem 4, let λ_r^* denote the r-th singular value of $(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2}$ and $\widehat{\lambda}^*_{r+1}$ the $(r+1)$-th singular value of $(\widehat{\Sigma}_{xS_uS_u})^{-1/2}\widehat{\Sigma}_{xyS_uS_v}(\widehat{\Sigma}_{yS_vS_v})^{-1/2}$. For

$$\begin{aligned}\Delta = &(\widehat{\Sigma}_{xS_uS_u})^{-1/2}\widehat{\Sigma}_{xyS_uS_v}(\widehat{\Sigma}_{yS_vS_v})^{-1/2} \\ &\quad - (\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2},\end{aligned}$$

Theorem 4 implies

$$(69) \le \frac{C\|\Delta\|_{\mathrm{F}}}{\lambda_r - \widehat{\lambda}^*_{r+1}} \le \frac{C\sqrt{r}\|\Delta\|_{\mathrm{op}}}{\lambda_r - \widehat{\lambda}^*_{r+1}}. \tag{70}$$

In what follows, we upper bound $\|\Delta\|_{\mathrm{op}}$ and lower bound $(\lambda_r - \widehat{\lambda}^*_{r+1})$ respectively.

To control $\|\Delta\|_{\mathrm{op}}$, observe that $\Delta = \sum_{i=1}^3 \Delta_i$, where

$$\begin{aligned}
\Delta_1 &= (\widehat{\Sigma}_{xS_uS_u})^{-1/2}(\Sigma_{xyS_uS_v} - \widehat{\Sigma}_{xyS_uS_v})(\Sigma_{yS_vS_v})^{-1/2},\\
\Delta_2 &= (\widehat{\Sigma}_{xS_uS_u})^{-1/2}((\widehat{\Sigma}_{xS_uS_u})^{1/2} - (\Sigma_{xS_uS_u})^{1/2})(\Sigma_{xS_uS_u})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_vS_v})^{-1/2},\\
\Delta_3 &= (\widehat{\Sigma}_{xS_uS_u})^{-1/2}\widehat{\Sigma}_{xyS_uS_v}(\widehat{\Sigma}_{yS_vS_v})^{-1/2}((\widehat{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{yS_vS_v})^{1/2})(\Sigma_{yS_vS_v})^{-1/2}.
\end{aligned}$$

By Lemma 13 and the bounds (65) – (67), we obtain

$$\|\Delta\|_{\mathrm{op}} \le \sum_{i=1}^3 \|\Delta_i\|_{\mathrm{op}} \le \sqrt{\frac{C}{n}(k_q^v + k_q^u + \log(ep/k_q^u) + \log(em/k_q^v))} \tag{71}$$

with the desired probability.

Turning to $\lambda^*_r - \widehat{\lambda}^*_{r+1}$, on the event such that (71) holds, the condition of the lemma further implies that the rightmost side is upper bounded by $\lambda/4$. Thus, Weyl's inequality [15, p.449] leads to

$$|\widehat{\lambda}^*_{r+1} - \lambda^*_{r+1}| \vee |\widehat{\lambda}^*_r - \lambda^*_r| \le \|\Delta\|_{\mathrm{op}} \le \frac{\lambda}{4}.$$

Together with the results on the λ^*_l's in (61) and Lemma 11, the last display leads to

$$\lambda^*_r - \widehat{\lambda}^*_{r+1} \ge 0.8\lambda - 0.1\lambda - |\lambda^*_{r+1} - \widehat{\lambda}^*_{r+1}| \ge 0.4\lambda. \tag{72}$$

We obtain the desired bound hence complete the proof of (32) by assembling (70) – (72). $\square$

PROOF OF (33). Now we provide a bound for $\|\widehat{U}^*_1\Lambda_1\widehat{V}^{*\prime}_1 - U^*_1\Lambda_1 V^{*\prime}_1\|_{\mathrm{F}}$. Following the lines of the proof of (32), up to a constant multiplier, this quantity can be upper bounded by the sum of the following two terms:

$$\begin{aligned}
&\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}\Lambda_1(\widehat{V}^*_{1S_v*})'(\widehat{\Sigma}_{yS_vS_v})^{1/2}\\
&\quad - (\Sigma_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}\Lambda_1(\widehat{V}^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}},
\end{aligned} \tag{73}$$

and

$$\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}\Lambda_1(\widehat{V}^*_{1S_v*})'(\widehat{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{xS_uS_u})^{1/2}U^*_{1S_u*}\Lambda_1(V^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}. \tag{74}$$

Using arguments similar to those for bounding (68), the term (73) can be upper bounded by using (65) and (66). Namely, we have

$$(73) \le C\lambda\sqrt{\frac{r}{n}(k_q^u + \log(ep/k_q^u) + k_q^v + \log(em/k_q^v))},$$

with probability at least $1 - \exp(-C'(k_q^u + \log(ep/k_q^u))) - \exp(-C'(k_q^v + \log(em/k_q^v)))$. The only difference from the bound of (68) is the extra factor λ due to the presence of Λ_1.

We now turn to (74). Using the triangle inequality, it can be bounded by the sum of the following three terms:

$$\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}\widehat{\Lambda}^*_1(\widehat{V}^*_{1S_v*})'(\widehat{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{xS_uS_u})^{1/2}U^*_{1S_u*}\Lambda^*_1(V^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}, \tag{75}$$

$$\|(\widehat{\Sigma}_{xS_uS_u})^{1/2}\widehat{U}^*_{1S_u*}(\widehat{\Lambda}^*_1 - \Lambda_1)(\widehat{V}^*_{1S_v*})'(\widehat{\Sigma}_{yS_vS_v})^{1/2}\|_{\mathrm{F}}, \tag{76}$$

$$\|(\Sigma_{xS_uS_u})^{1/2}U^*_{1S_u*}(\Lambda^*_1 - \Lambda_1)(V^*_{1S_v*})'(\Sigma_{yS_vS_v})^{1/2}\|_{\mathrm{F}}. \tag{77}$$

In the rest of the proof, we derive upper bounds for these three terms in order.

By (43) of Theorem 4, we can bound (75) by

$$\frac{C\sqrt{r}(\lambda^*_1 + \widehat{\lambda}^*_1)}{\lambda^*_r - \widehat{\lambda}^*_{r+1}}\|\Delta\|_{\mathrm{op}} \le \sqrt{\frac{Cr}{n}(k_q^v + k_q^u + \log(ep/k_q^u) + \log(em/k_q^v))}.$$

By Weyl's inequality [15, p.449], we can bound (76) by

$$C\sqrt{r}\|\widehat{\Lambda}^*_1 - \Lambda^*_1\|_{\mathrm{op}} \le C\sqrt{r}\|\Delta\|_{\mathrm{op}} \le \sqrt{\frac{Cr}{n}(k_q^v + k_q^u + \log(ep/k_q^u) + \log(em/k_q^v))}.$$

Finally, by (61) and the bound for (64) (which is from (55) and (56)), we can bound (77) by

$$C'\|\Lambda^*_1 - \Lambda_1\|_{\mathrm{F}} \le C'\|\Lambda^*_1 - \widetilde{\Lambda}_1\|_{\mathrm{F}} + C'\|\widetilde{\Lambda}_1 - \Lambda_1\|_{\mathrm{F}} \le C\sqrt{\frac{q}{2-q}}\lambda\epsilon_n.$$

The last three displays joint give the bound for (74). Together with the bound for (73), it leads to the desired upper bound for (33). □

4.5. *Proof of Lemma 4.* In this proof, we need the following technical result, which is a direct consequence of Lemma 13 by applying union bound. Remember the notations T_u and T_v defined in (45).

LEMMA 14. *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*

$$\begin{aligned} \|\widehat{\Sigma}_{xT_uT_u} - \Sigma_{xT_uT_u}\|_{\mathrm{op}}^2 &\leq \frac{C}{n}(k_q^u \log(ep/k_q^u)), \\ \|\widehat{\Sigma}_{yT_vT_v} - \Sigma_{yT_vT_v}\|_{\mathrm{op}}^2 &\leq \frac{C}{n}(k_q^v \log(em/k_q^v)), \end{aligned}$$

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 15 (Stewart and Sun [28], 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 $\left\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1 \widehat{V}_1' \right\rangle$. By the definition of trace product, we have

$$\begin{aligned} \left\langle \Sigma_x U_2 \Lambda_2 V_2' \Sigma_y, \widehat{U}_1 \widehat{V}_1' \right\rangle &= \left\langle \Lambda_2 V_2' \Sigma_y \widehat{V}_1', U_2' \Sigma_x \widehat{U}_1 \right\rangle \\ &\leq \|\Lambda_2 V_2' \Sigma_y \widehat{V}_1'\|_{\mathrm{F}} \|U_2' \Sigma_x \widehat{U}_1\|_{\mathrm{F}} \\ &\leq \lambda_{r+1} \|V_2' \Sigma_y \widehat{V}_1'\|_{\mathrm{F}} \|U_2' \Sigma_x \widehat{U}_1\|_{\mathrm{F}}. \end{aligned}$$

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

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

For any matrix W, we have

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

where $\|\widehat{R}\|_{\mathrm{op}} \leq \|\widehat{U}_1\|_{\mathrm{op}} \leq \|(\widehat{\Sigma}_{xT_uT_u})^{-1/2}\|_{\mathrm{op}} \|(\widehat{\Sigma}_{xT_uT_u})^{1/2} \widehat{U}_{1T_u*}\|_{\mathrm{op}} \leq C$ 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 14. Hence, by Lemma 15, we have

$$\|\widehat{U}_1' \Sigma_x U_2\|_{\mathrm{F}} \leq C \inf_W \|\widehat{\Theta} - \Theta W\|_{\mathrm{F}} \leq C \|\widehat{\Theta}\widehat{\Theta}' - \Theta\Theta'\|_{\mathrm{F}}. \tag{78}$$

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_1V_1' respectively and the eigen-gap is at constant level since the r-th singular value of U_1V_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 [35] gives

$$\|\widehat{\Theta}\widehat{\Theta}' - \Theta\Theta'\|_{\mathrm{F}} \le C\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}. \tag{79}$$

Combining (78) and (79), we have $\|\widehat{U}_1'\Sigma_x U_2\|_{\mathrm{F}} \le C_1\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}$. The same argument also implies $\|V_2'\Sigma_y\widehat{V}_1'\|_{\mathrm{F}} \le C_1\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}$. Therefore,

$$\left|\left\langle \Sigma_x U_2\Lambda_2 V_2'\Sigma_y, \widehat{U}_1\widehat{V}_1'\right\rangle\right| \le C_2\lambda_{r+1}\|\widehat{U}_1\widehat{V}_1' - U_1V_1'\|_{\mathrm{F}}^2.$$

Using the similar argument, we also obtain

$$\left|\left\langle \Sigma_x U_2\Lambda_2 V_2'\Sigma_y, \widehat{U}_1^*(\widehat{V}_1^*)'\right\rangle\right| \le C_2\lambda_{r+1}\|\widehat{U}_1^*(\widehat{V}_1^*)' - U_1V_1'\|_{\mathrm{F}}^2.$$

By triangle inequality, we complete the proof.

4.6. *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

$$\left\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widehat{U}_1^*\widehat{V}_1^{*'} - \widehat{U}_1\widehat{V}_1'\right\rangle = \frac{1}{2}\left\langle \Sigma - \widehat{\Sigma}, W\right\rangle. \tag{80}$$

In the rest of the proof, we bound $\langle \Sigma - \widehat{\Sigma}, W\rangle$ by using Lemma 8.

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 $|T| = |T_u| + |T_v| \le 2(k_q^u + k_q^v)$. Thus, we can rewrite (80) as

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

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

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

To obtain the desired bound, it suffices to show that

$$\left|\left\langle I_{|T|} - \Sigma_{TT}^{-1/2} \widehat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, K^T \right\rangle\right| \tag{81}$$

is upper bounded by $C\lambda\epsilon_n$ 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 possible 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 possible 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 $T_1, T_2, ..., T_N$. Let

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

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

$$\begin{aligned}
|(81)| &\le \max_{1\le j\le N} \left|\left\langle I_{|T_j|} - \Sigma_{T_jT_j}^{-1/2} \widehat{\Sigma}_{T_jT_j} \Sigma_{T_jT_j}^{-1/2}, K^{T_j} \right\rangle\right| \\
&\le \max_{1\le j\le N} \sup_{\|K\|_{\mathrm{F}}\le 1, \mathrm{rank}(K)\le 2r} \left|\left\langle I_{|T_j|} - \Sigma_{T_jT_j}^{-1/2} \widehat{\Sigma}_{T_jT_j} \Sigma_{T_jT_j}^{-1/2}, K \right\rangle\right|.
\end{aligned}$$

Then the union bound leads to

$$\begin{aligned}
&\mathbb{P}_{\Sigma}(|(81)| > t) \\
\le\ & \sum_{j=1}^{N} \mathbb{P}\left(\sup_{\|K\|_{\mathrm{F}}\le 1, \mathrm{rank}(K)\le 2r} \left|\left\langle I_{|T_j|} - \Sigma_{T_jT_j}^{-1/2} \widehat{\Sigma}_{T_jT_j} \Sigma_{T_jT_j}^{-1/2}, K \right\rangle\right| > t \right) \\
\le\ & \sum_{j=1}^{N} \exp(C' r|T_j| - Cn(t \wedge t^2)) \qquad (82) \\
\le\ & \binom{p}{k_q^u}\binom{m}{k_q^v} \exp(C_1 r(k_q^u + k_q^v) - Cn(t\wedge t^2)) \\
\le\ & \exp\left(C_1 r(k_q^u + k_q^v) + k_q^u \log\frac{ep}{k_q^u} + k_q^v \log\frac{em}{k_q^v} - Cn(t\wedge t^2)\right)
\end{aligned}$$

where the inequality (82) is due to Lemma 8. We complete the proof by choosing $t^2 = C_2\lambda^2\epsilon_n^2$ in the last display for some sufficiently large constant $C_2 > 0$, which, by condition (16), is bounded.

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

$$
\begin{aligned}
& \left\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \widehat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \right\rangle \\
= & \left\langle \Sigma_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \Sigma_y - \Sigma_x U_1^* \Lambda_1 V_1^{*\prime} \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \right\rangle \qquad (83) \\
& + \left\langle \Sigma_x U_1^* \Lambda_1 V_1^{*\prime} \Sigma_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \right\rangle \qquad (84) \\
& + \left\langle \widehat{\Sigma}_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \widehat{\Sigma}_y - \Sigma_x \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} \Sigma_y, \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \right\rangle . \qquad (85)
\end{aligned}
$$

In what follows, we bound each of the terms in (83) – (85) in order.

1° *Bound for* (83). Applying (33) in Lemma 3, we obtain that with probability at least $1-\exp(-C'(k_q^u+\log(ep/k_q^u)))-\exp(-C'(k_q^v+\log(em/k_q^v)))$,

$$
\begin{aligned}
|(83)| & \leq C \| \widehat{U}_1^* \Lambda_1 \widehat{V}_1^{*\prime} - U_1^* \Lambda_1 V_1^{*\prime} \|_{\mathrm{F}} \| \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \|_{\mathrm{F}} \\
& \leq C \sqrt{\frac{q}{2-q}} \lambda \epsilon_n \| \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \|_{\mathrm{F}}.
\end{aligned}
$$

2° *Bound for* (84). Applying (31) in Lemma 2, we obtain

$$
\begin{aligned}
|(84)| & \leq C \| U_1^* \Lambda_1 V_1^{*\prime} - U_1 \Lambda_1 V_1' \|_{\mathrm{F}} \| \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \|_{\mathrm{F}} \\
& \leq C \sqrt{\frac{q}{2-q}} \lambda \epsilon_n \| \widehat{U}_1^* \widehat{V}_1^{*'} - \widehat{U}_1 \widehat{V}_1' \|_{\mathrm{F}}.
\end{aligned}
$$

3° *Bound for* (85). We turn to bound (85) 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 8. Recall the random sets T_u and T_v defined in (45). Then for

$$
\begin{aligned}
H_x^{T_u} & = (\Sigma_{xT_uT_u})^{1/2} (\widehat{U}_{1T_u*}^* (\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*} (\widehat{V}_{1T_v*})') \\
& \quad \times \widehat{\Sigma}_{yT_vT_v} \widehat{V}_{1T_v*}^* \Lambda_1 (\widehat{U}_{1T_u*}^*)' (\Sigma_{xT_uT_u})^{1/2}, \\
H_y^{T_v} & = (\Sigma_{yT_vT_v})^{1/2} \widehat{V}_{1T_v*}^* \Lambda_1 (\widehat{U}_{1T_u*}^*)' \\
& \quad \times \Sigma_{xT_uT_u} (\widehat{U}_{1T_u*}^* (\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*} (\widehat{V}_{1T_v*})') (\Sigma_{yT_vT_v})^{1/2},
\end{aligned}
$$

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

$$\begin{aligned}
&|(85)| \\
= \ & \left|\left\langle \widehat{\Sigma}_x - \Sigma_x, (\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}')\widehat{\Sigma}_y\widehat{V}_1^*\Lambda_1\widehat{U}_1^{*\prime}\right\rangle\right. \\
& \left. + \left\langle \widehat{\Sigma}_y - \Sigma_y, \widehat{V}_1^*\Lambda_1\widehat{U}_1^{*\prime}\Sigma_x(\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}')\right\rangle\right| \\
\leq \ & \left|\left\langle \widehat{\Sigma}_{xT_uT_u} - \Sigma_{xT_uT_u}, (\widehat{U}_{1T_u*}^*(\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*}(\widehat{V}_{1T_v*})')\widehat{\Sigma}_{yT_vT_v}\widehat{V}_{1T_v*}^*\Lambda_1(\widehat{U}_{1T_u*}^*)'\right\rangle\right| \\
& + \left|\left\langle \widehat{\Sigma}_{yT_vT_v} - \Sigma_{yT_vT_v}, \widehat{V}_{1T_v*}^*\Lambda_1(\widehat{U}_{1T_u*}^*)'\Sigma_{xT_uT_u}(\widehat{U}_{1T_u*}^*(\widehat{V}_{1T_v*}^*)' - \widehat{U}_{1T_u*}(\widehat{V}_{1T_v*})')\right\rangle\right| \\
= \ & \|H_x^{T_u}\|_{\mathrm{F}}\left|\left\langle (\Sigma_{xT_uT_u})^{-1/2}\widehat{\Sigma}_{xT_uT_u}(\Sigma_{xT_uT_u})^{-1/2} - I_{|T_u|}, \overline{H}_x^{T_u}\right\rangle\right| \\
& + \|H_y^{T_v}\|_{\mathrm{F}}\left|\left\langle (\Sigma_{yT_vT_v})^{-1/2}\widehat{\Sigma}_{yT_vT_v}(\Sigma_{yT_vT_v})^{-1/2} - I_{|T_v|}, \overline{H}_y^{T_v}\right\rangle\right|.
\end{aligned}$$

We now bound each term on the rightmost hand side. Applying Lemma 8 with union bound and then following a similar analysis to that leading to (81) but with T replaced by T_u and T_v, we obtain that

(86)

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

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 $\|H_x^{T_u}\|_{\mathrm{F}}$ and $\|H_y^{T_v}\|_{\mathrm{F}}$, we note that it follows from Lemma 14 that all eigenvalues of $\widehat{\Sigma}_{xT_uT_u}$ and $\widehat{\Sigma}_{yT_vT_v}$ 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

$$\begin{aligned}
\|H_x^{T_u}\|_{\mathrm{F}} &\leq C\lambda\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}'\|_{\mathrm{F}}\|\widehat{\Sigma}_{yT_vT_v}^{1/2}\widehat{V}_{1T_v*}^*\|_{\mathrm{op}} \\
&\quad \|\widehat{\Sigma}_{yT_vT_v}^{1/2}\|_{\mathrm{op}}\|\widehat{\Sigma}_{xT_uT_u}^{1/2}\widehat{U}_{1T_u*}^*\|_{\mathrm{op}}\|\widehat{\Sigma}_{xT_uT_u}^{-1/2}\|_{\mathrm{op}} \\
&\leq C_1\lambda\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}'\|_{\mathrm{F}}, \qquad (87)
\end{aligned}$$

and

$$\begin{aligned}
\|H_y^{T_v}\|_{\mathrm{F}} &\leq C\lambda\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}'\|_{\mathrm{F}}\|\widehat{\Sigma}_{yT_vT_v}^{1/2}\widehat{V}_{1T_v*}^*\|_{\mathrm{op}} \\
&\quad \|\widehat{\Sigma}_{yT_vT_v}^{-1/2}\|_{\mathrm{op}}\|\widehat{\Sigma}_{xT_uT_u}^{1/2}\widehat{U}_{1T_u*}^*\|_{\mathrm{op}}\|\widehat{\Sigma}_{xT_uT_u}^{-1/2}\|_{\mathrm{op}} \\
&\leq C_1\lambda\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}'\|_{\mathrm{F}}. \qquad (88)
\end{aligned}$$

Combining (86), (87) and (88), we obtain

$$|(85)| \leq C\lambda^2\epsilon_n\|\widehat{U}_1^*\widehat{V}_1^{*\prime} - \widehat{U}_1\widehat{V}_1{}'\|_{\mathrm{F}},$$

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.

References.

[1] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. 37(5B):2877–2921, 2009.

[2] T. W. Anderson. Asymptotic theory for canonical correlation analysis. *Journal of Multivariate Analysis*, 70(1):1–29, 1999.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, 3rd edition, 2003.

[4] B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage*, 50(3): 1004–1016, 2010.

[5] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[6] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[7] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Probability Theory and Related Fields*, 65(2):181–237, 1983.

[8] A. Birnbaum, I.M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *arXiv preprint arXiv:1203.0967*, 2012.

[9] T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[10] T. T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, To appear., 2014.

[11] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[12] M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse CCA via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

[13] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

[14] C. Gao and H. H. Zhou. Rate-optimal posterior contraction for sparse PCA. *arXiv preprint arXiv:1312.0142*, 2013.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[16] D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[17] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[18] I. M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics*, 29:295–327, 2001.

[19] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104 (486):682–693, 2009.

[20] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[21] K.-A. Lê Cao, P. G. P. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34, 2009.

[22] Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[23] Z. Ma and Y. Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *arXiv preprint arXiv:1306.3609*, 2013.

[24] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[25] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[26] E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[27] B. A. Schmitt. Perturbation bounds for matrix square roots and pythagorean sums. *Linear Algebra and Its Applications*, 174:215–227, 1992.

[28] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990.

[29] T. Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

[30] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Stat. Soc. B*, 61(3):611–622, 1999.

[31] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[32] V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[33] S. Waaijenborg and A. H. Zwinderman. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics*, 10(1):315, 2009.

[34] Y. X. R. Wang, K. Jiang, L. J. Feldman, P. J. Bickel, and H. Huang. Inferring gene association networks using sparse canonical correlation analysis. *arXiv preprint arXiv:1401.6504*, 2014.

[35] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.

[36] A. Wiesel, M. Kliger, and A. O. Hero III. A greedy approach to sparse canonical correlation analysis. *arXiv preprint arXiv:0801.2748*, 2008.

[37] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009.

[38] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[39] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.