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

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

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

\[
\text{Cov}(u'X, u'_jX) = \text{Cov}(v'Y, v'_jY) = 0, \quad \text{for} \quad j = 1, \ldots, i.
\]

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

\[
\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2},
\]

where \(\Sigma_x = \text{Cov}(X), \Sigma_y = \text{Cov}(Y)\) and \(\Sigma_{xy} = \text{Cov}(X, Y)\). When one is only given a random sample \(\{(X_i, Y_i) : i = 1, \ldots, n\}\) 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 avail-
able 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 \(\Sigma_{xy}\), the covariance between \(X\) and \(Y\), was assumed to have a
rank one structure. They reparametrize \(\Sigma_{xy}\) as follows,

\[
\Sigma_{xy} = \Sigma_x \lambda u'v'\Sigma_y,
\]

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 direc-
tions. 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 \(\Sigma_x\) and
\(\Sigma_y\) 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 interest-
ing 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], \(\Sigma_{xy}\) can be reparametrized as follows,

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

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 $UAV'$ as

$$UAV' = U_1\Lambda_1 V_1' + U_2\Lambda_2 V_2',$$

where $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_r)$, $\Lambda_2 = \text{diag}(\lambda_{r+1}, \ldots, \lambda_r)$, $U_1 \in \mathbb{R}^{p \times r}$, $V_1 \in \mathbb{R}^{m \times r}$, $U_2 \in \mathbb{R}^{p \times r_2}$ and $V_2 \in \mathbb{R}^{m \times r_2}$ for $r_2 = r - r$. 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 $\Sigma_x$ and $\Sigma_y$ 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

$$L(U_1V_1', \hat{U}_1\hat{V}_1') = ||U_1V_1' - \hat{U}_1\hat{V}_1'||_F^2.$$  

We choose this loss function for several reasons. First, even when the $\lambda_i$’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 $\Sigma_x$ and $\Sigma_y$ 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 $\|P_{U_1} - \hat{P}_{U_1}\|_F^2$ and $\|P_{V_1} - \hat{P}_{V_1}\|_F^2$ for estimating the column subspaces of $U_1$ and $V_1$, which have been used in the related problem of sparse principal component analysis [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 $\lambda_r = \cdots = \lambda_p = 0$ in (4).

We use weak-$l_q$ balls to quantify sparsity. Let $\|(U_1)_{j*}\|$ denote the $\ell_2$ 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-$\ell_q$ radius for some $q \in [0, 2)$,

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

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

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

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

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

Throughout the paper, we assume $\kappa \lambda \leq 1 - c_0$ for some absolute constant $c_0 \in (0, 1)$. The key parameters $s_u, s_v, p, m, r$ and $\lambda$ 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 rank($\Sigma_{xy}) = 1$, in this paper, the rank of $\Sigma_{xy}$ can be as high as $p$ or $m$, $r$ is allowed to grow, and we do not need structural assumptions on $\Sigma_x$ and $\Sigma_y$ such as sparsity.

Suppose we observe i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \sim N_{p+m}(0, \Sigma)$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \asymp b_n$ if for some absolute constant $C > 1, 1/C \leq a_n/b_n \leq C$ 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 $q = 0$:

$$
\begin{align*}
\inf_{U_1, V_1} & \sup_{\Sigma \in \mathcal{F}_q(s_u, s_v, p, m, r, \lambda)} \mathbb{E}_\Sigma \|U_1 V_1^T - \hat{U}_1 \hat{V}_1^T\|^2_F \\
\asymp & \frac{1}{n \lambda^2} \left( r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v} \right),
\end{align*}
$$

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 $\Sigma_x, \Sigma_y, U_2$ 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

\[
\Sigma_0 = V \Lambda V' + \sigma^2 I_p.
\]

Here \( V \in \mathbb{R}^{p \times r} \) has orthonormal column vectors which are assumed to be sparse and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r) \) with \( \lambda_1 \geq \cdots \geq \lambda_r > 0 \). Since the first \( r \) eigenvalues of \( \Sigma_0 \) are \( \{\lambda_i + \sigma^2\}_{i=1}^r \) and the rest are all \( \sigma^2 \), the \( \lambda_i \)'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 $\Sigma_x$ and $\Sigma_y$ 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 $\{1, 2, \ldots, p\}$. For any set $I$, $|I|$ denotes its cardinality and $I^c$ its complement. For two subsets $I$ and $J$ of indices, we write $A_{IJ}$ for the $|I| \times |J|$ submatrices formed by $a_{ij}$ with $(i, j) \in I \times J$. When $I$ or $J$ is the whole set, we abbreviate it with an $*$, and so if $A \in \mathbb{R}^{p \times k}$, then $A_{I*} = A_{I[k]}$ and $A_{*J} = A_{[p]J}$. For any square matrix $A = (a_{ij})$, denote its trace by $\text{Tr}(A) = \sum_i a_{ii}$. 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 $\|A\|_F = \sqrt{\text{Tr}(AA^T)}$ and $\|A\|_{op} = \sigma_1(A)$, respectively.
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 $\langle A, B \rangle = \text{Tr}(A'B)$. For any number $a$, we use $\lfloor a \rfloor$ 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 $1_{\{E\}}$ 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,

$$\max_{(A, B)} \text{Tr}(A'\Sigma_{xy}B) \quad \text{s.t.} \quad A'\Sigma_xA = B'\Sigma_yB = I_r.$$  \hfill (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

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

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

The effective sparsity of $U_1$ and $V_1$ are defined as

$$k_q^u = \lfloor x_q^u \rfloor, \quad k_q^v = \lceil x_q^v \rceil. \hfill (13)$$

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

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

$$||(V_1)_{(j)*}|| \leq \left( \frac{r + \log(em/k_q^v)}{n\lambda^2} \right)^{1/2},$$

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

$$||(V_1)_{(j)*}|| \leq \left( \frac{r + \log(em/k_q^v)}{n\lambda^2} \right)^{1/2}. \hfill (14)$$
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,

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

subject to $A^T \widehat{\Sigma}_x A = B^T \widehat{\Sigma}_y B = I_r$ and \( \|A\|_{0,w} = k_q^u, \|B\|_{0,w} = k_q^v \).

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

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

which is the minimax rate to be shown later.

**Theorem 1.** Under the assumption that

$$\begin{align*}
\epsilon_n^2 &\leq c, \\
\lambda_{r+1} &\leq c \lambda,
\end{align*}$$

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

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

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 $\lambda_r - \lambda_{r+1} \geq (1 - c) \lambda_r > 0$. 

---

imsart-aos ver. 2011/05/20 file: SCCA_paper.tex date: May 6, 2014
Remark 2. The upper bound $c_n^2$ has two parts. The first part $\frac{1}{M^2} (r(k_q^u + k_q^v))$ is caused by low rank structure, and the second part $\frac{1}{M^2} (k_q^u \log(ep/k_q^u) + k_q^v \log(em/k_q^v))$ is caused by sparsity. If $r \leq \log(ep/k_q^u) \vee \log(em/k_q^v)$, 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 $U_1 V_1'$ are bounded in Frobenius norm, because

$$
\|U_1 V_1'\|_F \leq \|\Sigma_x^{-1/2}\|_{op} \|\Sigma_x^{1/2} U_1\|_F \|\Sigma_y^{1/2} V_1\|_{op} \|\Sigma_y^{-1/2}\|_{op} \leq M \sqrt{r}.
$$

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

$$
\hat{U}_1 V_1' = \hat{U}_1 V_1' \mathbf{1}_{\{\|\hat{U}_1 V_1'\|_F \leq 2M \sqrt{r}\}}.
$$

The modification can be viewed as an improvement, because whenever $\|\hat{U}_1 V_1'\|_F > 2M \sqrt{r}$, we have

$$
\|\hat{U}_1 V_1' - U_1 V_1\|_F \geq \|\hat{U}_1 V_1'\|_F - \|U_1 V_1'\|_F \geq M \sqrt{r} \geq \|0 - U_1 V_1'\|_F.
$$

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

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

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

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

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

$$
\sup_{\Sigma \in \mathcal{F}_q} \mathbb{E}_\Sigma \|\hat{U}_1 V_1' - U_1 V_1'\|_F^2 \leq C_2 c_n^2.
$$

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 $c_n^2$. In this section, we show that this rate is optimal among all estimators. More specifically, we show that the following minimax lower bounds hold for $q \in [0, 2)$.

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

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

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

$$\inf_{(U_1, V_1) \in \mathcal{F}_q} \sup_{\Sigma \in \mathcal{F}_q} \mathbb{E} \| \tilde{U}_1 \tilde{V}_1^t - U_1 V_1^t \|_F^2 \geq cn^2 \land c_0.$$

**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

$$\begin{align*}
\max_{(A, B)} &\quad \text{Tr}(A' \Sigma_{xy} B) \\
\text{s.t.} &\quad A' \Sigma_x A = B' \Sigma_y B = I_r \quad \text{and} \quad \text{supp}(A) \subset S_u, \text{supp}(B) \subset S_v.
\end{align*}$$

(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 $(\hat{U}_1^*, \hat{V}_1^*)$ as a solution to

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

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

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

We note that

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

By definition, the matrices $(U_1^*, V_1^*)$ are normalized with respect to $\Sigma_x$ and $\Sigma_y$, and $(\hat{U}_1^*, \hat{V}_1^*)$ are normalized with respect to $\hat{\Sigma}_x$ and $\hat{\Sigma}_y$.

Last but not least, let

$$\hat{S}_u = \text{supp}(\hat{U}_1), \quad \hat{S}_v = \text{supp}(\hat{V}_1).$$

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

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

$$\leq \frac{6C}{\lambda_r} \left\langle \Sigma_x U_2 A_2 V_2' \Sigma_y, \hat{U}_1^* (\hat{V}_1^*)' - \hat{U}_1 \hat{V}_1' \right\rangle$$

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

$$+ \frac{6C}{\lambda_r} \left\langle \hat{\Sigma}_{xy}, \hat{U}_1^* \Lambda_1 \hat{V}_1' - \hat{\Sigma}_y - \Sigma_x U_1 A_1 V_1' \Sigma_y, \hat{U}_1^* (\hat{V}_1^*)' - \hat{U}_1 \hat{V}_1' \right\rangle,$$

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


Proof. See Section 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 $U_2 \Lambda_2 V_2'$ 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 $M, \kappa, q$, such that

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

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

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

$$
\|\hat{U}_1^* (\hat{V}_1^*)' - U_1^* (V_1^*)'\|_F^2 \leq \frac{C r}{n\lambda^2} \left[ k_u^u + k_q^u + \log \left( \frac{ep}{k_q^u} \right) + \log \left( \frac{em}{k_q^u} \right) \right], \tag{32}
$$

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

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

Since $r \leq k_u^u \wedge k_q^u$, (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 $\frac{1}{n} (k_u^u \log(ep/k_q^u) + k_q^u \log(em/k_q^u)) < C_1$ for some constant $C_1 > 0$. For any constant $C' > 0$, there exists a constant
$C > 0$ only depending on $M, \kappa, C_1$ and $C'$, such that
\[
\left| \langle \Sigma_x U_2 \Lambda_2 V'_2 \Sigma_y, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \\
\leq C \lambda_{r+1} \left( \| \hat{U}'_1 (\hat{V}'_1) - U_1 V'_1 \|_F^2 + \| U_1 V'_1 - \hat{U}_1 \hat{V}'_1 \|_F^2 \right),
\]
with probability at least $1 - \exp(\frac{-C' k_q^u \log(ep/k_q^u)}{C}) - \exp(\frac{-C'' k_q^v \log(em/k_q^v)})$.

The bias is Lemma 4 is 0 when $U_2 \Lambda_2 V'_2$ is 0.

**Lemma 5 (Excess loss 1).** Suppose (16) holds. For any constant $C' > 0$, there exists a constant $C > 0$ only depending on $M$ and $C'$, such that
\[
\left| \langle \Sigma_{xy} - \hat{\Sigma}_{xy}, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \leq C \lambda \epsilon_n \| \hat{U}_1 \hat{V}'_1 - \hat{U}'_1 (\hat{V}'_1) \|_F,
\]
with probability at least $1 - \exp(\frac{-C' (k_q^u + k_q^v) \log(ep/k_q^u)}{C}) - \exp(\frac{-C' (k_q^v \log(em/k_q^v)))}{C}.

**Lemma 6 (Excess loss 2).** Suppose (16) and (17) hold. For any constant $C' > 0$, there exists a constant $C > 0$ only depending on $M, \kappa, q$ and $C'$, such that
\[
\left| \langle \hat{\Sigma}_x \hat{U}'_1 \Lambda_1 (\hat{V}'_1) \hat{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V'_1 \Sigma_y, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \leq C \lambda \epsilon_n \| \hat{U}_1 \hat{V}'_1 - \hat{U}'_1 (\hat{V}'_1) \|_F,
\]
with probability at least $1 - \exp(\frac{-C' (k_q^u + \log(ep/k_q^u))}{C}) - \exp(\frac{-C' (k_q^v \log(em/k_q^v)))}{C}.

3.1.4. **Proof of Theorem 1.** For notational convenience, let
\[ R = \| \hat{U}_1 \hat{V}'_1 - U_1 V'_1 \|_F, \quad \theta = \| U_1 (V'_1) - U_1 V'_1 \|_F, \quad \delta = \| (V'_1 - U_1 (V'_1)) \|_F. \]
Consider the event that the conclusions of Lemmas 1 – 6 hold, which occurs with probability at least $1 - \exp(\frac{-C' (k_q^u + \log(ep/k_q^u))}{C}) - \exp(\frac{-C' (k_q^v \log(em/k_q^v)))}{C}$ according to the union bound. On this event, Lemma 2 and Lemma 3 imply that
\[ \theta^2 \leq C \epsilon_n^2 \quad \text{and} \quad \delta^2 \leq C \epsilon_n. \]
Moreover, Lemma 4 implies
\[ \left| \frac{1}{\lambda_r} \langle \Sigma_x U_2 \Lambda_2 V'_2 \Sigma_y, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \leq \frac{C \lambda_{r+1}}{\lambda} \left( R^2 + \theta^2 + \delta^2 \right), \]
Lemma 5 implies
\[ \left| \frac{1}{\lambda_r} \langle \Sigma_{xy} - \hat{\Sigma}_{xy}, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \leq C \epsilon_n (R + \theta + \delta), \]
\[ \text{Lemma 5 implies} \]
\[ \left| \frac{1}{\lambda_r} \langle \Sigma_{xy} - \hat{\Sigma}_{xy}, \hat{U}'_1 (\hat{V}'_1) - \hat{U}_1 \hat{V}'_1 \rangle \right| \leq C \epsilon_n (R + \theta + \delta), \]
and Lemma 6 implies
\[
\left\| \frac{1}{\lambda_r} \left( \bar{\Sigma}_x \widehat{U}_1^* \Lambda_1 (\widehat{V}_1^*)' \bar{\Sigma}_y - \Sigma_x U_1 \Lambda_1 V_1' \Sigma_y, U_1 (\widehat{V}_1^*)' - \widehat{U}_1 \widehat{V}_1' \right) \right\| \leq C \epsilon_n (R + \theta + \delta).
\]
Together with Lemma 1, the above bounds lead to
\[
R^2 \leq C (\theta^2 + \delta^2) + \frac{C \lambda_r + 1}{\lambda} (R^2 + \theta^2 + \delta^2) + C \epsilon_n (R + \theta + \delta)
\]
\[
\leq \frac{C \lambda_r + 1}{\lambda} R^2 + C \epsilon_n R + C \epsilon_n^2.
\]
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 \(\epsilon_n\) in (15), and let \(C_1\) be the constant in (19) and (20). The result of Theorem 1 implies that we can choose an arbitrarily large constant \(C'\) such that \(C' > C_1\). Given \(C'\), there exists a constant \(C\), by which we can bound the risk as follows
\[
\mathbb{E}_\Sigma \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F
\]
\[
\leq \mathbb{E}_\Sigma \left[ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F 1 \left\{ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F \leq C \epsilon_n^2 \right\} \right]
\]
\[
+ \mathbb{E}_\Sigma \left[ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F 1 \left\{ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F > C \epsilon_n^2 \right\} \right]
\]
\[
(34) \leq C \epsilon_n^2 + \mathbb{E}_\Sigma \left[ \left( 2 \| \widehat{U}_1 \widehat{V}_1' \|^2_F + 2 \| U_1 V_1' \|^2_F \right) 1 \left\{ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F > C \epsilon_n^2 \right\} \right]
\]
\[
(35) \leq C \epsilon_n^2 + 6 M^2 r \mathbb{P}_\Sigma \left( \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F > C \epsilon_n^2 \right)
\]
\[
(36) \leq C \epsilon_n^2.
\]
Here, the inequality (34) is due to the triangle inequality and the fact that
\[
\left\{ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F > C \epsilon_n^2 \right\} \subset \left\{ \| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F > C \epsilon_n^2 \right\}.
\]
In fact, if \(\| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F \leq C \epsilon_n^2\), then \(\| \widehat{U}_1 \widehat{V}_1' \|^2_F \leq C \epsilon_n^2 + M^2 r \leq 2 M^2 r\). By our definition of the estimator, this means \(\widehat{U}_1 \widehat{V}_1' = \widehat{U}_1 \widehat{V}_1'\), which further implies \(\| \widehat{U}_1 \widehat{V}_1' - U_1 V_1' \|^2_F \leq C \epsilon_n^2\). The inequality (35) follows from our definition of estimator \(\widehat{U}_1 \widehat{V}_1'\) 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 \( r \leq (p - k^u_q + 1) \land (m - k^v_q + 1) \).

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 \( U_1, \Lambda_1 \) 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 \( I \) 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

\[
D(P(1) \| P(2)) = \frac{n\lambda^2}{2(1 - \lambda^2)} \|U(1)V'(1) - U(2)V'(2)\|^2_F.
\]

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

\[
d_{KL}(T) \triangleq \sup_{\theta, \theta' \in T} D(P_\theta \| P_{\theta'}).
\]

Then

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

**Proof of Theorem 3** (Case \( q = 0 \)). Note that in this case, \( k^u_0 = s_u \) and \( k^v_0 = s_v \).
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

\[
B(\epsilon_0) = \{ U \in O(p, r) : \text{supp}(U) \subset [s_u], \| U - U_0 \|_F \leq \epsilon_0 \}.
\]

and

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

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

\[
d_{KL}(T_0) = \sup_{U(i) \in B(\epsilon_0)} \frac{n \lambda^2}{2(1 - \lambda^2)} \left\| U(1)V_0' - U(2)V_0' \right\|_F^2
\]

\[
= \sup_{U(i) \in B(\epsilon_0)} \frac{n \lambda^2}{2(1 - \lambda^2)} \left\| U(1) - U(2) \right\|_F^2 = \frac{2n \lambda^2 \epsilon_0^2}{1 - \lambda^2}.
\]

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 \( \{ \tilde{U}(1), \ldots, \tilde{U}(N) \} \subset O(p, r) \) be a maximal set such that \( \text{supp}(\tilde{U}_i) \subset [s_u] \), and for any \( i \neq j \in [N] \),

\[
\| \tilde{U}(i)\tilde{U}'(i) - U_0U_0' \|_F \leq \epsilon_0, \quad \| \tilde{U}(i)\tilde{U}'(i) - \tilde{U}(j)\tilde{U}'(j) \|_F \geq \sqrt{2} \alpha \epsilon_0.
\]

Then by [9, Lemma 1], for some absolute constant \( C > 1 \),

\[
N \geq \left( \frac{1}{C \alpha} \right)^{r(s_u - r)}.
\]

For each \( \tilde{U}(i) \), define

\[
U(i) = \tilde{U}(i)O(i), \quad \text{for } O(i) = \arg\min_{O \in O(r)} \| \tilde{U}(i)O - U_0 \|_F.
\]

Then for any \( i \in [N] \), by definition, \( U(i) \in O(p, r) \), supp\( (U(i)) \subset [s_u] \), and

\[
U(i)U'(i) = \tilde{U}(i)\tilde{U}'(i).
\]

In addition, [28, Theorem II.4.11] implies

\[
\| U(i) - U_0 \|_F \leq \| \tilde{U}(i)\tilde{U}'(i) - U_0U_0' \|_F \leq \epsilon_0.
\]
and so $U(i) \in B(\epsilon)$. On the other hand, note that $\|\tilde{U}(i) - \tilde{U}(j)\|_F \geq \sqrt{2}\|\tilde{U}(i) - \tilde{U}(j)\|_F$, and hence for $i \neq j \in [N],$

$$\|U(i) - U(j)\|_F \geq \frac{1}{\sqrt{2}}\|U(i)U'_(i) - U(j)U'_j\|_F = \frac{1}{\sqrt{2}}\|\tilde{U}(i)\tilde{U}'(i) - \tilde{U}(j)\tilde{U}'(j)\|_F \geq \alpha \epsilon 0.$$ Let $\rho(\Sigma(1), \Sigma(2)) = \|U(1) V'_0 - U(2) V'_0\|_F = \|U(1) - U(2)\|_F$. Then the foregoing argument implies that for $\epsilon = \alpha \epsilon 0$,

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

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 $\lambda$ 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 $s_u \log \frac{m}{s_v}$, 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:

$$T_1 = \left\{ \Sigma = \begin{bmatrix} I_p & \lambda UV' \alpha \end{bmatrix}, \begin{bmatrix} I_{r-1} & 0 \\ \lambda V_0 U' I_m & 0 \end{bmatrix} : U = \begin{bmatrix} I_{r-1} \\ \lambda V_0 U' I_m \end{bmatrix}, u_r \in \mathcal{G}^{r+1}, \text{supp}(u_r) \leq s_u - r + 1 \right\}.$$ 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 $\|u_r - \hat{u}_r\|_F^2$. 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. \(\square\)

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 \(\Vert \widetilde{U}(i) \Vert_{q, w} \leq s_u\), and that for some absolute constant \(C, N \geq (\frac{1}{C^2})^{r k_q^u/2}\).

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 \(\Vert U(i) \Vert_{q, w} = \Vert \widetilde{U}(i) \Vert_{q, w} \leq s_u\). 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. \(\square\)

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, \(\epsilon_n^2\) 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 \(\| \cdot \|_F\) and operator norm \(\| \cdot \|_{op}\) are unitarily invariant.
THEOREM 4. Consider matrices $X, Y \in \mathbb{R}^{p \times m}$. Let the SVD of $X$ and $Y$ be

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

with $D_1 = \text{diag}(d_1, \ldots, d_r)$ and $\hat{D}_1 = \text{diag}(\hat{d}_1, \ldots, \hat{d}_r)$. Suppose there is a positive constant $\delta \in (0, d_r]$ such that $\|D_2\|_{\text{op}} \leq d_r - \delta$. Let $\|\cdot\|$ be any unitarily invariant norm, and $\epsilon = \|A_1'(X - Y)\| \vee \|(X - Y) B_1\|$. Then, we have

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

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

$$\|A_1 B_1' - \hat{A}_1 \hat{B}_1'\| \leq \frac{C \epsilon}{\delta}. \quad (44)$$

Remark 5. In addition, when $X$ and $Y$ are positive semi-definite, $A_l = B_l$, $\hat{A}_l = \hat{B}_l$ for $l = 1, 2$, we recover the classical Davis–Kahan sin-theta theorem [13] $\|A_1 A_1' - \hat{A}_1 \hat{A}_1'\| \leq C \epsilon / \delta$ up to a constant multiplier.

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

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

$$\mathbb{P} \left( \sup_{\{K: \|K\|_2 \leq 1, \text{rank}(K) \leq \ell\}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i' - I_d, K \right| > t \right) \leq \exp(C' rd - C n(t^2 \wedge t)).$$

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

$$T_u = S_u \cup \hat{S}_u \quad \text{and} \quad T_v = S_v \cup \hat{S}_v. \quad (45)$$

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 $D = (d_l)_{1 \leq l \leq r}$ with $d_1 \geq d_2 \geq \ldots \geq d_r > 0$ and $A'A = B'B = E'E = F'F = I_r$, we have

$$\frac{d_r}{2} \|AB' - EF\|_F^2 \leq \langle ADB', AB' - EF' \rangle \leq \frac{d_1}{2} \|AB' - EF'\|_F^2.$$
Lemma 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 $T_x \times T_y$, we have

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

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

Proof of Lemma 1. First of all, the triangle inequality and Jensen’s inequality together lead to

$$\|\hat{U}_1 \hat{V}_1' - U_1 V_1'\|_F^2 \leq 3 \left(\|\hat{U}_1 \hat{V}_1' - U_1 V_1'\|_F^2 + \|\hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}'_1\|_F^2 + \|U_1 V_1' - \hat{U}_1 \hat{V}'_1\|_F^2\right).$$

(46)

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

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

(47)

$$\leq \frac{2C}{\lambda_r} \left\langle \hat{\Sigma}_x^{1/2} \hat{U}_1 \hat{V}_1' \hat{\Sigma}_y^{1/2}, \hat{\Sigma}_x^{1/2}(\hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}'_1)\hat{\Sigma}_y^{1/2} \right\rangle$$

$$= \frac{2C}{\lambda_r} \left\langle \hat{\Sigma}_x \hat{U}_1 \hat{V}_1' \hat{\Sigma}_y, \hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}'_1 \right\rangle$$

$$\leq \frac{2C}{\lambda_r} \left\langle \hat{\Sigma}_x \hat{U}_1 \hat{V}_1' \hat{\Sigma}_y - \hat{\Sigma}_x \hat{U}_1 \hat{V}_1' \hat{\Sigma}_y, \hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}'_1 \right\rangle$$

$$\leq \frac{2C}{\lambda_r} \left\langle \hat{\Sigma}_x \hat{U}_1 \hat{V}_1' \hat{\Sigma}_y, \hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}'_1 \right\rangle$$

(48)

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

$$\left\langle \hat{\Sigma}_{xy}, \hat{U}_1^\top \hat{V}_1' - \hat{U}_1^\top \hat{V}'_1 \right\rangle \leq 0.$$

The equality (50) comes from the CCA structure (4) and (5). Combining (46)-(50) and rearranging the terms, we obtain the desired result. $\square$
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 $\Sigma_x$ and $\Sigma_y$. To overcome the difficulty, we introduce intermediate matrices $(\tilde{U}_1, \tilde{V}_1)$ defined as follows. First, we write the SVD of $(\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2}U_{1_{S_{u_{S_{u}}}}}\Lambda_1(V_{1_{S_{v_{S_{v}}}}})'(\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2})^{1/2}$ as

$$
(\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2}U_{1_{S_{u_{S_{u}}}}}\Lambda_1(V_{1_{S_{v_{S_{v}}}}})'(\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2})^{1/2} = P\tilde{\Lambda}_1 Q',
$$
and let $\tilde{U}_1 S_{u} = (\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2})^{-1/2}P$ and $\tilde{V}_1 S_{v} = (\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2})^{-1/2}Q$. Finally, we define $\tilde{U}_1 \in R^{p \times r}$ and $\tilde{V}_1 \in R^{m \times r}$ by setting

$$
(\tilde{U}_1)_{S_{u}} = \tilde{U}_1 S_{u}, \quad (\tilde{V}_1)_{S_{u}} = 0, \quad (\tilde{V}_1)_{S_{v}} = \tilde{V}_1 S_{v}, \quad (\tilde{V}_1)_{S_{v}} = 0.
$$

By definition, we have $U_{1_{S_{u_{S_{u}}}}}\Lambda_1(V_{1_{S_{v_{S_{v}}}}})' = \tilde{U}_1 S_{u}\tilde{\Lambda}(\tilde{V}_1 S_{v})'$. Last but not least, we define

$$
\Xi = P\tilde{\Lambda}_1 Q' + (I - P'P)(\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2})^{-1/2}U_{2} \Lambda_2 V_{2}'\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2}(\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2})^{-1/2}(I - QQ').
$$

We now summarize the key properties of the $\tilde{U}_1$, $\tilde{V}_1$ and $\tilde{\Lambda}$ matrices in the following two lemmas, the proofs of which are deferred to the appendix.

**Lemma 11.** Let $P, Q$ and $\Xi$ 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 $\Xi$;
2. The first and the $r$-th singular values $\tilde{\lambda}_1$ and $\tilde{\lambda}_r$ of $\Xi$ satisfy $1.1\kappa \lambda \geq \tilde{\lambda}_1 \geq \tilde{\lambda}_r \geq 0.9\lambda$, and the $(r + 1)$-th singular value $\tilde{\lambda}_{r+1} \leq \epsilon\lambda$ for some sufficiently small constant $\epsilon > 0$;
3. The column vectors of $\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2}\tilde{U}_1$ and $\Sigma_{y_{S_{u_{S_{u}}}}}^{1/2}\tilde{V}_1$ are the $r$ leading left and right singular vectors of $\Sigma_{x_{S_{u_{S_{u}}}}}^{1/2}\tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}_1$.

**Lemma 12.** For some constant $C > 0$,

$$
\|\tilde{U}'_{1}\Sigma_{x_{S_{u_{S_{u}}}}}U_{2}\|_{F} \leq C\|U_{1_{S_{u_{S_{u}}}}}\|_{F}^{2} \quad \text{and} \quad \|\tilde{V}'_{1}\Sigma_{y_{S_{u_{S_{u}}}}}V_{2}\|_{F} \leq C\|V_{1_{S_{u_{S_{u}}}}}\|_{F}^{2}.
$$

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

**Proof of (30).** By triangle inequality,

$$
\|U'_{1}V_{1} - U_{1}V'_{1}\|_{F} \leq \|U'_{1}V_{1} - U_{1}'V_{1}\|_{F} + \|U_{1}'V_{1} - U_{1}V'_{1}\|_{F}.
$$

It is sufficient to bound each of the two terms on the right side.
1° Bound for $\|\tilde{U}_1 \tilde{V}_1' - U_1 V_1'\|_F$. Since the smallest eigenvalues of $\Sigma_x$ and $\Sigma_y$ are bounded from below by some absolute positive constant,

$$\|\tilde{U}_1 \tilde{V}_1' - U_1 V_1'\|_F \leq C\|\Sigma_x^{1/2} (\tilde{U}_1 \tilde{V}_1' - U_1 V_1') \Sigma_y^{1/2}\|_F.$$ 

By Lemma 11, $\Sigma_x^{1/2} \tilde{U}_1$ and $\Sigma_y^{1/2} \tilde{V}_1$ collect the $r$ leading left and right singular vectors of $\Sigma_x^{1/2} \tilde{U}_1 \tilde{V}_1' \Sigma_y^{1/2}$, 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 $\Sigma_x^{1/2} U_1 V_1' \Sigma_y^{1/2}$, Thus, Theorem 4 implies

$$\|\Sigma_x^{1/2} (\tilde{U}_1 \tilde{V}_1' - U_1 V_1') \Sigma_y^{1/2}\|_F \leq C\|\Sigma_x^{1/2} (\tilde{U}_1 \tilde{V}_1' - U_1 V_1') \Sigma_y^{1/2}\|_F.$$ 

The right side of the above inequality is bounded as

$$\|\tilde{U}_1 \tilde{V}_1' - U_1 V_1'\|_F \leq \|\tilde{U}_1 S_u \tilde{V}_1 - U_1 \lambda_1 V_1\|_F + \|U_1 S_u \Lambda_1 (V_1 S_u)\|_F.$$ 

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

$$\|U_1 S_u\|_F^2 \leq \frac{Cq}{2 - q} \frac{k_u^u (s_u/k_q)}{2} \leq \frac{Cq}{2 - q} e_n,$$

$$\|V_1 S_u\|_F^2 \leq \frac{Cq}{2 - q} \frac{k_v^v (s_v/k_q)}{2} \leq \frac{Cq}{2 - q} e_n.$$ 

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

$$\|\tilde{U}_1 \tilde{V}_1' - U_1 V_1'\|_F \leq \frac{Cq}{2 - q} e_n.$$ 

2° Bound for $\|U_1^{*} V_1^{*} - \tilde{U}_1 \tilde{V}_1'\|_F$. Let $\lambda_{r+1}$ denote the $(r + 1)$-th singular value of $(\Sigma_x S_u S_u)^{-1/2} \Sigma_{xy} S_u S_u (\Sigma_y S_u S_u)^{-1/2}$. Then we have

$$\|U_1^{*} V_1^{*} - \tilde{U}_1 \tilde{V}_1'\|_F = \|U_1^{*} (V_1 S_u') - \tilde{U}_1 S_u \Lambda_1 (V_1 S_u)\|_F$$

$$\leq C\|\Sigma_x S_u (U_1 S_u')^{1/2} \|_F \|\Sigma_y S_u (V_1 S_u')^{1/2} \|_F$$

$$\leq C\|\Sigma_x S_u \Sigma_y S_u (\Sigma_y S_u S_u)^{-1/2} - \Xi\|_F$$

$$\leq \lambda_r - \lambda_{r+1}.$$
Here, the first equality holds since both $U_1^*V_1'$ and $\tilde{U}_1\tilde{V}_1'$ are supported on the $S_u \times S_v$ submatrix. Noting that by the discussion before (22), (52) and Lemma 11, $(\Sigma_{xS_uS_v})^{1/2}U_{1S_u*}, (\Sigma_{yS_uS_v})^{1/2}V_{1S_v*}$ and $(\Sigma_{xS_uS_v})^{-1/2}\tilde{U}_{1S_u*}, (\Sigma_{yS_uS_v})^{-1/2}\tilde{V}_{1S_v*}$ collect the leading left and right singular vectors of $(\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_uS_v})^{-1/2}$ and $\Xi$ 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 $\| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_uS_v})^{-1/2} - \Xi \|^2_F$**. First, we decompose $\Sigma_{xyS_uS_v}$ as

$$
\Sigma_{xyS_uS_v} = \Sigma_{xS_u*}(U_1\Lambda_1V_1' + U_2\Lambda_2V_2')\Sigma_{yS_v}
$$

(59)

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

$$
\| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_uS_v})^{-1/2} - \Xi \|^2_F \\
\leq \| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}U_{1S_u*}\Lambda_1V_1'(\Sigma_{yS_uS_v})^{-1/2} \|_F \\
+ \| (\Sigma_{xS_uS_v})^{1/2}U_{1S_u*}\Lambda_1V_1'\Sigma_{yS_uS_v}(\Sigma_{yS_uS_v})^{-1/2} \|_F \\
+ \| PP'(\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}U_2\Lambda_2V_2'(\Sigma_{yS_uS_v})^{-1/2}(I - QQ') \|_F \\
+ \| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}U_2\Lambda_2V_2'(\Sigma_{yS_uS_v})^{-1/2}QQ' \|_F \\
\leq C\lambda \|U_{1S_u*}\|_F + \|V_{1S_v*}\|_F \\
+ C\lambda_{r+1} \|P'(\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}U_2\|_F + \|Q'(\Sigma_{yS_uS_v})^{-1/2}\Sigma_{yS_uS_v}V_2\|_F \\
= C\lambda \|U_{1S_u*}\|_F + \|V_{1S_v*}\|_F + C\lambda_{r+1}(\|\tilde{U}_1\Sigma_2U_2\|_F + \|\tilde{V}_1\Sigma_2V_2\|_F).
$$

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

$$
(60) \quad \| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_uS_v})^{-1/2} - \Xi \|^2_F \leq \frac{Cq}{2-q} \lambda^2 \epsilon^2_n.
$$

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

$$
|\lambda_{r+1}^* - \bar{\lambda}_{r+1}| \leq \|\Lambda_{r+1}^* - \bar{\Lambda}_1\|_F \\
\leq \| (\Sigma_{xS_uS_v})^{-1/2}\Sigma_{xyS_uS_v}(\Sigma_{yS_uS_v})^{-1/2} - \Xi \|^2_F \leq C\left(\frac{q}{2-q}\right)^{1/2} \lambda \epsilon_n \leq 0.1\lambda.
$$

$$
(61)
$$
Together with Lemma 11, it further implies

\[ \tilde{\lambda}_r - \lambda^*_r \geq \tilde{\lambda}_r - \tilde{\lambda}_{r+1} - |\tilde{\lambda}_{r+1} - \lambda^*_r| \geq 0.7 \lambda. \]

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

\[ \| \tilde{U}_1 \tilde{V}'_1 - U_1^* V'_1 \|_F^2 \leq \frac{C q}{2 - q} \epsilon_n^2. \]

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

\[ \square \]

**Proof of (31).** Note that

\[
\| U_1^* \Lambda_1 V''_1 - U_1 \Lambda_1 V'_1 \|_F \\
\leq \| U_1^* \Lambda_1 V''_1 - \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 \|_F + \| \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 - U_1 \Lambda_1 V'_1 \|_F \\
\leq \| U_1^* \Lambda_1 V''_1 - \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 \|_F + \| \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 - U_1 \Lambda_1 V'_1 \|_F \\
+ C \| \Lambda_1^* - \tilde{\Lambda}_1 \|_F + C \| \tilde{\Lambda}_1 - \Lambda_1 \|_F \\
\leq \| U_1^* \Lambda_1 V''_1 - \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 \|_F + C' \| \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 - U_1 \Lambda_1 V'_1 \|_F + C \| \Lambda_1^* - \tilde{\Lambda}_1 \|_F.
\]

Here the last inequality is due to

\[ \| \tilde{\Lambda}_1 - \Lambda_1 \|_F \leq \| \Sigma_x^{1/2} (\tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 - U_1 \Lambda_1 V'_1) \Sigma_y^{1/2} \|_F, \]

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, \( \| \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 - U_1 \Lambda_1 V'_1 \|_F \) is upper bounded by the righthand side of (31). Next, the bound for \( \| \Lambda_1^* - \tilde{\Lambda}_1 \|_F \) has been shown in (61). Last but not least, applying (43) in Theorem 4, we obtain

\[
\| U_1^* \Lambda_1 V''_1 - \tilde{U}_1 \tilde{\Lambda}_1 \tilde{V}'_1 \|_F \\
\leq \frac{C (\tilde{\lambda}_1 + \lambda^*_1)}{\lambda_r - \lambda^*_{r+1}} \| (\Sigma_x S_u S_u) \Sigma_x \Sigma_y \|_F \\
\leq \frac{C \sqrt{q}}{2 - q} \lambda \epsilon_n,
\]

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.

\[ \square \]

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 $\frac{1}{n}(k_u^v + k^v_q + \log(ep/k^u_q) + \log(em/k^v_q)) < C_1$ for some constant $C_1 > 0$. Consider deterministic sets $B_u \subset [p]$ and $B_v \subset [m]$ with $|B_u| = C_u k^u_q$ and $|B_v| = C_v k^v_q$ 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,

$$
\|\hat{\Sigma}_x B_u - \Sigma x B_u\|_{op}^2 \leq \frac{C}{n}(k^u_q + \log(ep/k^u_q)),
$$

$$
\|\hat{\Sigma}_y B_u - \Sigma y B_u\|_{op}^2 \leq \frac{C}{n}(k^v_q + \log(em/k^v_q)),
$$

$$
\|\hat{\Sigma}_{xy} B_u - \Sigma_{xy} B_u\|_{op}^2 \leq \frac{C}{n}(k^v_q + k^u_q + \log(ep/k^u_q) + \log(em/k^v_q)),
$$

with probability at least 1 - $\exp(-C'(k^u_q + \log(ep/k^u_q))) - \exp(-C'(k^v_q + \log(em/k^v_q)))$.

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:

$$
\|\hat{\Sigma}_{x S_u S_u}^{1/2} - (\Sigma x S_u S_u)^{1/2}\|_{op}^2 \leq \frac{C}{n}(k^u_q + \log(ep/k^u_q)),
$$

$$
\|\hat{\Sigma}_{y S_v S_v}^{1/2} - (\Sigma y S_v S_v)^{1/2}\|_{op}^2 \leq \frac{C}{n}(k^v_q + \log(em/k^v_q)),
$$

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

$$
\|\hat{\Sigma}_{x S_u S_u}^i\|_{op} \vee \|\hat{\Sigma}_{y S_v S_v}^i\|_{op} \leq C_1, \quad i = \pm 1.
$$

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

Proof of (32). Since both $\hat{U}_1^s \hat{V}_1^{s'}$ and $U_1^s V_1^{s'}$ are supported on the $S_u \times S_v$ submatrix, we have

$$
\|\hat{U}_1^s \hat{V}_1^{s'} - U_1^s V_1^{s'}\|_F = \|\hat{U}_1^s (\hat{V}_1^{s'})^\top - U_1^s (V_1^{s'})^\top\|_F
\leq C \|(\Sigma x S_u S_u)^{1/2} (\hat{U}_1^s (\hat{V}_1^{s'})^\top - U_1^s (V_1^{s'})^\top)(\Sigma y S_v S_v)^{1/2}\|_F. 
$$

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:

$$
\|(\Sigma x S_u S_u)^{1/2} (\hat{U}_1^s (\hat{V}_1^{s'})^\top (\Sigma y S_v S_v)^{1/2} 
- (\hat{\Sigma}_x S_u S_u)^{1/2} (\hat{V}_1^{s'})^\top (\hat{\Sigma}_y S_v S_v)^{1/2}\|_F,
$$
and
\begin{equation}
\| \left( \widetilde{\Sigma}_{x,S_u} \right)^{1/2} \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \check{\Sigma}_{y,S_v} \right)^{1/2} \right. \\
- \left. \left( \Sigma_{x,S_u} S_u \right)^{1/2} U_{1S_u}^* \left( V_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F.
\end{equation}

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{align}
(68) & \leq \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F \\
& + \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F \\
& \leq \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} - \left( \Sigma_{x,S_u} S_u \right)^{1/2} \|_{op} \\
& + \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F \| \left( \Sigma_{y,S_v} \right)^{1/2} - \left( \Sigma_{y,S_v} \right)^{1/2} \|_{op}.
\end{align}

To further bound the rightmost side of the last display, we note that with high probability
\begin{align}
\| \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F & \leq \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \|_F \| \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \|_{op} \\
& \times \| \left( \Sigma_{y,S_v} \right)^{-1/2} \|_{op} \| \left( \Sigma_{y,S_v} \right)^{1/2} \|_{op} \\
& \leq C \sqrt{r}.
\end{align}

Together with (65), this implies that with high probability,
\begin{align}
\| \hat{U}_{1S_u}^* \left( \hat{V}_{1S_u}^* \right)' \left( \Sigma_{y,S_v} \right)^{1/2} \|_F \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} - \left( \Sigma_{x,S_u} S_u \right)^{1/2} \|_{op} \\
& \leq 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{align}

By a similar argument, \( \| \left( \Sigma_{x,S_u} S_u \right)^{1/2} \check{U}_{1S_u}^* \left( \check{V}_{1S_u}^* \right)' \|_F \| \left( \Sigma_{y,S_v} \right)^{-1/2} - \left( \Sigma_{y,S_v} \right)^{-1/2} \|_{op} \) satisfies the same upper bound. Thus, (68) is upper bounded by the rate in (32) with the desired probability by noting that \( \lambda \leq 1 \).

2° Bound for (69). By definition, \( \left( \Sigma_{x,S_u} S_u \right)^{1/2} \hat{U}_{1S_u}^* \) and \( \left( \Sigma_{y,S_v} \right)^{1/2} \hat{V}_{1S_u}^* \) collect the \( r \) leading left and right singular vectors of \( \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \check{\Sigma}_{y,S_v} \right)^{-1/2} \), and \( \left( \Sigma_{x,S_u} S_u \right)^{1/2} \check{U}_{1S_u}^* \) and \( \left( \Sigma_{y,S_v} \right)^{1/2} \check{V}_{1S_u}^* \) collect the \( r \) leading left and right singular vectors of \( \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \Sigma_{y,S_v} \right)^{-1/2} \). To apply Theorem 4, let \( \lambda_r^* \) denote the \( r \)-th singular value of \( \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \Sigma_{y,S_v} \right)^{-1/2} \), and \( \hat{\lambda}_{r+1}^* \) the \((r+1)\)-th singular value of \( \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \Sigma_{y,S_v} \right)^{-1/2} \). For
\begin{align}
\Delta &= \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \Sigma_{y,S_v} \right)^{-1/2} \\
& - \left( \Sigma_{x,S_u} S_u \right)^{-1/2} \Sigma_{xy,S_v} \left( \Sigma_{y,S_v} \right)^{-1/2},
\end{align}
Theorem 4 implies

\begin{equation}
(69) \quad \frac{C\|\Delta\|_F}{\lambda_r - \tilde{\lambda}_{r+1}^{\star}} \leq \frac{C\sqrt{r}\|\Delta\|_{op}}{\lambda_r - \tilde{\lambda}_{r+1}^{\star}}.
\end{equation}

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

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

\begin{align*}
\Delta_1 &= (\tilde{\Sigma}_{xS_uS_u})^{-1/2}(\Sigma_{xyS_uS_u} - \tilde{\Sigma}_{xyS_uS_u})(\Sigma_{yS_vS_v})^{-1/2}, \\
\Delta_2 &= (\tilde{\Sigma}_{xS_uS_u})^{-1/2}((\tilde{\Sigma}_{xS_uS_u})^{1/2} - (\Sigma_{xS_uS_u})^{1/2})(\Sigma_{xyS_uS_v} - \Sigma_{xyS_uS_v})(\Sigma_{yS_vS_v})^{-1/2}, \\
\Delta_3 &= (\tilde{\Sigma}_{xS_uS_u})^{-1/2}(\tilde{\Sigma}_{xyS_uS_v} - \tilde{\Sigma}_{xyS_uS_v})(\Sigma_{yS_vS_v})^{-1/2}((\tilde{\Sigma}_{yS_vS_v})^{1/2} - (\Sigma_{yS_vS_v})^{1/2})(\Sigma_{yS_vS_v})^{-1/2}.
\end{align*}

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

\begin{equation}
(71) \quad \|\Delta\|_{op} \leq \sum_{i=1}^{3} \|\Delta_i\|_{op} \leq \sqrt{\frac{C}{n}}(k_i^n + k^{n\prime}_n + \log(ep/k^n_q) + \log(em/k^{n\prime}_q))
\end{equation}

with the desired probability.

Turning to \(\lambda_r^{\star} - \tilde{\lambda}_{r+1}^{\star}\), 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

\[|\lambda_r^{\star} - \tilde{\lambda}_{r+1}^{\star}| + |\lambda_r^{\star} - \bar{\lambda}_r^{\star}| \leq \|\Delta\|_{op} \leq \frac{\lambda}{4}.
\]

Together with the results on the \(\lambda_i^{\star}\)’s in (61) and Lemma 11, the last display leads to

\begin{equation}
(72) \quad \lambda_r^{\star} - \tilde{\lambda}_{r+1}^{\star} \geq 0.8\lambda - 0.1\lambda - |\lambda_r^{\star} - \tilde{\lambda}_{r+1}^{\star}| \geq 0.4\lambda.
\end{equation}

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

**Proof of (33).** Now we provide a bound for \(\|\hat{U}_1^{\star} \Lambda_1 \hat{V}_1^{\star} - U_1^{\star} \Lambda_1 V_1^{\star}\|_F\). 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{align}
&(\tilde{\Sigma}_{xS_uS_u})^{1/2}(\tilde{\Sigma}_{xS_uS_u})^{1/2} - (\Sigma_{xS_uS_u})^{1/2}(\Sigma_{xS_uS_u})^{1/2}||_{F},
\end{align}

(73)
and

\[
\| (\hat{\Sigma}_{xS_uS_v})^{1/2} \hat{U}^*_{1S_u*} \Lambda_1 (\hat{V}^*_{1S_v*})' (\hat{\Sigma}_{yS_uS_v})^{1/2} \\
- (\Sigma_{xS_uS_v})^{1/2} U^*_{1S_u*} \Lambda_1 (V^*_{1S_v*})' (\Sigma_{yS_uS_v})^{1/2} \|_F.
\]

(74)

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

\[
(73) \leq C \lambda \sqrt{\frac{r}{n}} (k^u_q + \log(ep/k^u_q) + k^v_q + \log(em/k^v_q)),
\]

with probability at least \(1 - \exp(-C'(k^u_q + \log(ep/k^u_q)) - \exp(-C'(k^v_q + \log(em/k^v_q)))\). The only difference from the bound of (68) is the extra factor \(\lambda\) due to the presence of \(\Lambda_1\).

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

\[
\| (\hat{\Sigma}_{xS_uS_v})^{1/2} \hat{U}^*_{1S_u*} \tilde{\Lambda}_1^* (\hat{V}^*_{1S_v*})' (\hat{\Sigma}_{yS_uS_v})^{1/2} \\
- (\Sigma_{xS_uS_v})^{1/2} U^*_{1S_u*} \tilde{\Lambda}_1^* (V^*_{1S_v*})' (\Sigma_{yS_uS_v})^{1/2} \|_F,
\]

(75)

\[
\| (\hat{\Sigma}_{xS_uS_v})^{1/2} \hat{U}^*_{1S_u*} (\tilde{\Lambda}_1^* - \Lambda_1) (\hat{V}^*_{1S_v*})' (\hat{\Sigma}_{yS_uS_v})^{1/2} \|_F.
\]

(76)

\[
\| (\Sigma_{xS_uS_v})^{1/2} U^*_{1S_u*} (\tilde{\Lambda}_1^* - \Lambda_1) (V^*_{1S_v*})' (\Sigma_{yS_uS_v})^{1/2} \|_F.
\]

(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

\[
\frac{C \sqrt{\tau} (\lambda_1^u + \hat{\lambda}_1^u)}{\lambda_1^u - \hat{\lambda}_1^u + 1} \| \Delta \|_{op} \leq \sqrt{\frac{C r}{n}} (k^u_q + k^u_q + \log(ep/k^u_q) + \log(em/k^u_q)).
\]

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

\[
C \sqrt{\tau} \| \tilde{\Lambda}_1^* - \Lambda_1 \|_{op} \leq C \sqrt{\tau} \| \Delta \|_{op} \leq \sqrt{\frac{C r}{n}} (k^u_q + k^u_q + \log(ep/k^u_q) + \log(em/k^u_q)).
\]

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

\[
C' \| \Lambda_1^* - \Lambda_1 \|_F \leq C' \| \tilde{\Lambda}_1^* - \Lambda_1 \|_F + C' \| \tilde{\Lambda}_1 - \Lambda_1 \|_F \leq C \sqrt{\frac{q}{2 - q}} \lambda_{en}.
\]

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). 

\(\square\)
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 $\frac{1}{n}(k_u^u \log(ep/k_q^u) + k_v^v \log(em/k_q^v)) < C_1$ 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

$$\|\tilde{\Sigma}_{xT_u} - \Sigma_{xT_u}\|_{op}^2 \leq \frac{C}{n}(k_u^u \log(ep/k_q^u)),$$

$$\|\tilde{\Sigma}_{yT_v} - \Sigma_{yT_v}\|_{op}^2 \leq \frac{C}{n}(k_v^v \log(em/k_q^v)),$$

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

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

$$\inf_W \|A - BW\|_F \leq \|AA' - BB'\|_F.$$

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

$$\langle \Sigma_x U_2 \Lambda_2 V'_y \Sigma_y, \tilde{U}_1 \tilde{V}'_1 \rangle = \langle \Lambda_2 V'_y \Sigma_y \tilde{V}'_1, U_2 \Sigma_x \tilde{U}_1 \rangle \leq \|\Lambda_2 V'_y \Sigma_y \tilde{V}'_1\|_F \|U_2 \Sigma_x \tilde{U}_1\|_F \leq \lambda_{r+1} \|V'_y \Sigma_y \tilde{V}'_1\|_F \|U_2 \Sigma_x \tilde{U}_1\|_F.$$

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

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

For any matrix $W$, we have

$$\|\tilde{U}'_1 \Sigma_x U_2\|_F = \|(\tilde{U}_1 - U_1 HR^{-1} W \tilde{R} H')' \Sigma_x U_2\|_F \leq C \|(\tilde{U}_1 - U_1 HR^{-1} W \tilde{R} H')\|_F \leq C \|	ilde{R}\|_{op} \|	ilde{\Theta} - \Theta W\|_F,$$

where $\|	ilde{R}\|_{op} \leq \|	ilde{U}_1\|_{op} \leq \|((\Sigma_{xT_u})^{-1/2})_{op} \|((\tilde{\Sigma}_{xT_u})^{1/2} \tilde{U}_{1T_v})_{op} \leq C$ with probability at least $1 - \exp(-C'k_q^v \log(ep/k_q^u)) - \exp(-C'k_q^v \log(em/k_q^v))$ by Lemma 14. Hence, by Lemma 15, we have

$$\|\tilde{U}'_1 \Sigma_x U_2\|_F \leq C \inf_W \|	ilde{\Theta} - \Theta W\|_F \leq C \|\tilde{\Theta} \Theta' - \Theta \Theta'\|_F.$$
We note that both \( \widehat{\Theta}' \) and \( \Theta' \) are the projection matrices of the left singular spaces of \( \hat{U}_1 \hat{V}_1' \) and \( U_1 V_1' \) respectively and the eigen-gap is at constant level since the \( r \)-th singular value of \( U_1 V_1' \) is bounded below by some constant and the \( (r+1) \)-th singular value of \( \hat{U}_1 \hat{V}_1' \) is zero. Then a direct consequence of Wedin’s sin-theta theorem [35] gives

\[
\| \widehat{\Theta}' - \Theta' \|_F \leq C \| \hat{U}_1 \hat{V}_1' - U_1 V_1' \|_F.
\]

Combining (78) and (79), we have \( \| \hat{U}_1 \Sigma_x U_2 \|_F \leq C_1 \| \hat{U}_1 \hat{V}_1' - U_1 V_1' \|_F \). The same argument also implies \( \| V_2' \Sigma_y \hat{V}_1' \|_F \leq C_1 \| \hat{U}_1 \hat{V}_1' - U_1 V_1' \|_F \). Therefore,

\[
\left\langle \Sigma_x U_2 A_2 V_2' \Sigma_y, \hat{U}_1 \hat{V}_1' \right\rangle \leq C_2 \lambda_{r+1} \| \hat{U}_1 \hat{V}_1' - U_1 V_1' \|_F^2.
\]

Using the similar argument, we also obtain

\[
\left\langle \Sigma_x U_2 A_2 V_2' \Sigma_y, \hat{V}_1^* (\hat{V}_1^*)' \right\rangle \leq C_2 \lambda_{r+1} \| \hat{U}_1 \hat{V}_1' - U_1 V_1' \|_F^2.
\]

By triangle inequality, we complete the proof.

4.6. Proof of Lemma 5. Define

\[
W = \begin{bmatrix} 0 & \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \\ \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' & 0 \end{bmatrix}.
\]

Then simple algebra leads to

\[
\left\langle \hat{\Sigma}_{xy} - \hat{\Sigma}_{xy}, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \right\rangle = \frac{1}{2} \left\langle \Sigma - \hat{\Sigma}, W \right\rangle.
\]

In the rest of the proof, we bound \( \left\langle \Sigma - \hat{\Sigma}, W \right\rangle \) by using Lemma 8.

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

\[
T = T_u \cup (T_v + p)
\]

with \( T_v + p = \{ j + p : j \in T_v \} \). The cardinality of \( T \) is \( |T| = |T_u| + |T_v| \leq 2(k^u_q + k^v_q) \). Thus, we can rewrite (80) as

\[
\left\langle \hat{\Sigma}_{xy} - \hat{\Sigma}_{xy}, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \right\rangle = \frac{1}{2} \left\langle \Sigma - \hat{\Sigma}, W \right\rangle
\]

\[
= \frac{1}{2} \left\langle \Sigma_{TT} - \hat{\Sigma}_{TT}, W_{TT} \right\rangle
\]

\[
= \frac{1}{2} \left\langle I_{|T|} - \Sigma_{TT}^{-1/2} \hat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, \Sigma_{TT}^{1/2} W_{TT}^{1/2} \right\rangle
\]

\[
= \frac{1}{2} \| \Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2} \|_F \left\langle I_{|T|} - \Sigma_{TT}^{-1/2} \hat{\Sigma}_{TT} \Sigma_{TT}^{-1/2}, K^T \right\rangle,
\]
where $K^T = \frac{\Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2}}{\|\Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2}\|_F}$. Note that

$$\frac{1}{2} \|\Sigma_{TT}^{1/2} W_{TT} \Sigma_{TT}^{1/2}\|_F \leq C\|W_{TT}\|_F = C\|W\|_F = \sqrt{2}C\|\hat{U}_1 \hat{V}_1' - \hat{U}_1 \hat{V}_1'\|_F.$$ 

To obtain the desired bound, it suffices to show that

$$(81) \quad \left|\langle I_{[T]} - \Sigma_{TT}^{-1/2} \hat{S}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle\right|$$

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

To this end, we note that $T_u = S_u \cup \tilde{S}_u$ has at most $\binom{p}{k_q^u}$ different possible configurations since $S_u$ is deterministic and $\tilde{S}_u$ 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

$$K_{T_j} = \frac{\Sigma_{TT}^{1/2} W_{T_j} \Sigma_{TT}^{1/2}}{\|\Sigma_{TT}^{1/2} W_{T_j} \Sigma_{TT}^{1/2}\|_F}$$

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

$$\left|\langle I_{[T]} - \Sigma_{TT}^{-1/2} \hat{S}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle\right| \leq \max_{1 \leq j \leq N} \sup_{\|K\|_F \leq 1, \text{rank}(K) \leq 2r} \left|\langle I_{[T_j]} - \Sigma_{TT}^{-1/2} \hat{S}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle\right|.$$ 

Then the union bound leads to

$$\mathbb{P}_\Sigma(|(81)| > t) \leq \sum_{j=1}^N \mathbb{P} \left( \sup_{\|K\|_F \leq 1, \text{rank}(K) \leq 2r} \left|\langle I_{[T_j]} - \Sigma_{TT}^{-1/2} \hat{S}_{TT} \Sigma_{TT}^{-1/2}, K^T \rangle\right| > t \right)$$

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

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

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

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

\[
\langle \Sigma_x U_1^* \Lambda_1 \hat{V}_1^* \Sigma_y - \Sigma_x U_1^* \Lambda_1 V_1^* \Sigma_y, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \rangle \\
= \langle \Sigma_x U_1^* \Lambda_1 \hat{V}_1^* \Sigma_y - \Sigma_x U_1^* \Lambda_1 V_1^* \Sigma_y, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \rangle \\
+ \langle \Sigma_x U_1^* \Lambda_1 V_1^* \Sigma_y - \Sigma_x U_1^* \Lambda_1 V_1^* \Sigma_y, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \rangle \\
+ \langle \Sigma_x U_1^* \Lambda_1 \hat{V}_1^* \Sigma_y - \Sigma_x U_1^* \Lambda_1 \hat{V}_1^* \Sigma_y, \hat{U}_1^* \hat{V}_1' - \hat{U}_1 \hat{V}_1' \rangle.
\]

(83)

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 \(1 - \exp(-C'(k^u_n + \log(ep/k^u_n))) - \exp(-C'(k^v_n + \log(em/k^v_n)))\),

\[
| (83) | \leq C\| \hat{U}_1^* \Lambda_1 \hat{V}_1' - U_1^* \Lambda_1 V_1' \|_F \| \hat{U}_1 \hat{V}_1^* - \hat{U}_1 \hat{V}_1' \|_F \\
\leq C \frac{q}{2 - q} \lambda \epsilon_n \| \hat{U}_1 \hat{V}_1^* - \hat{U}_1 \hat{V}_1' \|_F.
\]

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

\[
| (84) | \leq C\| U_1^* \Lambda_1 V_1' \|_F \| \hat{U}_1 \hat{V}_1^* - \hat{U}_1 \hat{V}_1' \|_F \\
\leq C \frac{q}{2 - q} \lambda \epsilon_n \| \hat{U}_1 \hat{V}_1^* - \hat{U}_1 \hat{V}_1' \|_F.
\]

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

\[
H_{x}^{T_u} = (\Sigma x_{T_u} \Lambda_1)^{1/2} (\hat{U}_{1T_u}^* \Lambda_1 (\hat{V}_{1T_u}^* \Lambda_1)' - \hat{U}_{1T_u} (\hat{V}_{1T_u}^*)') \\
\times \hat{U}_{1T_u} \alpha_1 \Lambda_1 (\hat{V}_{1T_u}^*)' \alpha_1 (\Sigma x_{T_u} \Lambda_1)^{1/2},
\]

\[
H_{y}^{T_v} = (\Sigma y_{T_v} \Lambda_1)^{1/2} (\hat{U}_{1T_v}^* \Lambda_1 (\hat{V}_{1T_v}^*)' - \hat{U}_{1T_v} (\hat{V}_{1T_v}^*)') \alpha_1 (\Sigma y_{T_v} \Lambda_1)^{1/2},
\]
and $\mathbf{T}_x^{T_u} = H_x^{T_u} / \|H_x^{T_u}\|_F$, $\mathbf{T}_y^{T_v} = H_y^{T_v} / \|H_y^{T_v}\|_F$, we have

\begin{align}
&|\left(\tilde{\Sigma}_x - \Sigma_x, (\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1')\tilde{\Sigma}_y\tilde{V}^*_1\Lambda_1\tilde{U}_1'\right)\rangle \\
&\quad + \left(\tilde{\Sigma}_y - \Sigma_y, \tilde{V}^*_1\Lambda_1\tilde{U}_1'\Sigma_x(\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1')\right)\rangle \\
&\leq \left|\left(\tilde{\Sigma}_xT_uT_u - \Sigma_xT_uT_u, (\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1')\tilde{\Sigma}_yT_vT_v\tilde{V}^*_1T_{1u}\Lambda_1(\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1')\right)\right| \\
&\quad + \left|\left(\tilde{\Sigma}_yT_vT_v - \Sigma_yT_vT_v, \tilde{V}^*_1T_{1u}\Lambda_1(\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1')\right)\right| \\
&= \|H_x^{T_u}\|_F \left|\left(\Sigma_xT_uT_u - I_{|T_u|}\right)^{-1/2}\tilde{\Sigma}_xT_uT_u(\Sigma_xT_uT_u)^{-1/2} - I_{|T_u|}\right|, \\
&\quad + \|H_y^{T_v}\|_F \left|\left(\Sigma_yT_vT_v - I_{|T_v|}\right)^{-1/2}\tilde{\Sigma}_yT_vT_v(\Sigma_yT_vT_v)^{-1/2} - I_{|T_v|}\right|.
\end{align}

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

\begin{align}
&\left|\left(\Sigma_xT_uT_u - I_{|T_u|}\right)^{-1/2}\tilde{\Sigma}_xT_uT_u(\Sigma_xT_uT_u)^{-1/2} - I_{|T_u|}\right| \\
&\quad \leq C \sqrt{\frac{k_u}{n} \left( r + \log \frac{ep}{k_q^u} \right)}, \\
&\left|\left(\Sigma_yT_vT_v - I_{|T_v|}\right)^{-1/2}\tilde{\Sigma}_yT_vT_v(\Sigma_yT_vT_v)^{-1/2} - I_{|T_v|}\right| \\
&\quad \leq C \sqrt{\frac{k_v}{n} \left( r + \log \frac{em}{k_q^v} \right)}
\end{align}

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

To bound $\|H_x^{T_u}\|_F$ and $\|H_y^{T_v}\|_F$, we note that it follows from Lemma 14 that all eigenvalues of $\tilde{\Sigma}_xT_uT_u$ and $\tilde{\Sigma}_yT_vT_v$ are bounded from below and above by some universal positive constants with probability at least $1 - \exp(-C'k_u^u\log(ep/k_q^u)) - \exp(-C'k_v^v\log(em/k_q^v))$ under assumption (16). Thus, with the same probability we have

\begin{align}
\|H_x^{T_u}\|_F &\leq C\lambda \left\|\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1'\right\|_F \|\tilde{\Sigma}^{-1/2}_{xT_uT_u}\tilde{V}^*_1T_{1u}\|_{op} \\
&\quad + \|\tilde{\Sigma}^{-1/2}_{xT_uT_u}\|_{op} \|\tilde{\Sigma}^{1/2}_{xT_uT_u}\tilde{V}^*_1T_{1u}\|_{op} \|\tilde{\Sigma}^{-1/2}_{xT_uT_u}\|_{op} \\
&\leq C_1 \lambda \left\|\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1'\right\|_F,
\end{align}

and

\begin{align}
\|H_y^{T_v}\|_F &\leq C\lambda \left\|\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1'\right\|_F \|\tilde{\Sigma}^{-1/2}_{yT_vT_v}\tilde{V}^*_1T_{1u}\|_{op} \\
&\quad + \|\tilde{\Sigma}^{-1/2}_{yT_vT_v}\|_{op} \|\tilde{\Sigma}^{1/2}_{yT_vT_v}\tilde{V}^*_1T_{1u}\|_{op} \|\tilde{\Sigma}^{-1/2}_{xT_uT_u}\|_{op} \\
&\leq C_1 \lambda \left\|\tilde{U}_1^{*}\tilde{V}_1^{*} - \tilde{U}_1\tilde{V}_1'\right\|_F.
\end{align}
Combining (86), (87) and (88), we obtain
\[
|\langle 85 \rangle | \leq C \lambda^2 \epsilon_n \| \widehat{U}_1^* \widehat{V}_1^\nu - \widehat{U}_1 \widehat{V}_1 \|_F,
\]
with probability at least \(1 - \exp(-C' k_q^w \log(ep/k_q^w)) - \exp(-C' k_q^w \log(em/k_q^w))\).
Noting that \(\lambda < 1\), this completes the proof.

References.


