Bernoulli **26**(1), 2020, 432–470
https://doi.org/10.3150/19-BEJ1131

Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates

ZHUANG MA[1] and XIAODONG LI[2]

[1]*Department of Statistics, the Wharton School, University of Pennsylvania, 3730 Walnut street, Suite 400, Philadelphia, PA 19104, USA. E-mail: kop.mazhuang@gmail.com*
[2]*Department of Statistics, University of California, Davis, Davis, CA 95616, USA. E-mail: xdgli@ucdavis.edu*

Canonical correlation analysis (CCA) is a fundamental statistical tool for exploring the correlation structure between two sets of random variables. In this paper, motivated by the recent success of applying CCA to learn low dimensional representations of high dimensional objects, we propose two losses based on the principal angles between the model spaces spanned by the sample canonical variates and their population correspondents, respectively. We further characterize the non-asymptotic error bounds for the estimation risks under the proposed error metrics, which reveal how the performance of sample CCA depends adaptively on key quantities including the dimensions, the sample size, the condition number of the covariance matrices and particularly the population canonical correlation coefficients. The optimality of our uniform upper bounds is also justified by lower-bound analysis based on stringent and localized parameter spaces. To the best of our knowledge, for the first time our paper separates p_1 and p_2 for the first order term in the upper bounds without assuming the residual correlations are zeros. More significantly, our paper derives $(1-\lambda_k^2)(1-\lambda_{k+1}^2)/(\lambda_k-\lambda_{k+1})^2$ for the first time in the non-asymptotic CCA estimation convergence rates, which is essential to understand the behavior of CCA when the leading canonical correlation coefficients are close to 1.

Keywords: canonical correlation analysis; dimension reduction; minimax rates

1. Introduction

Canonical correlation analysis (CCA), first introduced in [17], is a fundamental statistical tool to characterize the relationship between two groups of random variables and finds a wide range of applications across many different fields. For example, in genome-wide association study (GWAS), CCA is used to discover the genetic associations between the genotype data of single nucleotide polymorphisms (SNPs) and the phenotype data of gene expression levels [7,30]. In information retrieval, CCA is used to embed both the search space (e.g., images) and the query space (e.g., text) into a shared low dimensional latent space such that the similarity between the queries and the candidates can be quantified [15,22]. In natural language processing, CCA is applied to the word co-occurrence matrix and learns vector representations of the words which capture the semantics [8,9]. Other applications, to name a few, include fMRI data analysis [11], computer vision [19] and speech recognition [2,27].

The enormous empirical success motivates us to revisit the estimation problem of canonical correlation analysis. Two theoretical questions are naturally posed: What are proper error metrics to quantify the discrepancy between population CCA and its sample estimates? And under such metrics, what are the quantities that characterize the fundamental statistical limits?

The justification of loss functions, in the context of CCA, has seldom appeared in the literature. From first principles that the proper metric to quantify the estimation loss should depend on the specific purpose of using CCA, we find that the applications discussed above mainly fall into two categories: identifying variables of interest and dimension reduction.

The first category, mostly in genomic research [7,30], treats one group of variables as responses and the other group of variables as covariates. The goal is to discover the specific subset of the covariates that are most correlated with the responses. Such applications are featured by low signal-to-noise ratio and the interpretability of the results is the major concern.

In contrast, the second category is investigated extensively in statistical machine learning and engineering community where CCA is used to learn low dimensional latent representations of complex objects such as images [22], text [8] and speeches [2]. These scenarios are usually accompanied with relatively high signal-to-noise ratio and the prediction accuracy, using the learned low dimensional embeddings as the new set of predictors, is of primary interest. In recent years, there has been a series of publications establishing fundamental theoretical guarantees for CCA to achieve sufficient dimension reduction ([6,10,12,18,24] and many others).

In this paper, we aim to address the problems raised above by treating CCA as a tool for dimension reduction.

1.1. Population and sample CCA

Let

$$\boldsymbol{x} = [X_1, \ldots, X_{p_1}]^\top \in \mathbb{R}^{p_1}, \qquad \boldsymbol{y} = [Y_1, \ldots, Y_{p_2}]^\top \in \mathbb{R}^{p_2} \tag{1.1}$$

be two sets of variates with the joint covariance matrix

$$\operatorname{Cov}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right) = \boldsymbol{\Sigma} := \begin{bmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{xy}^\top & \boldsymbol{\Sigma}_y\end{bmatrix}. \tag{1.2}$$

For simplicity, we assume

$$\mathbb{E}(X_i) = 0, \quad i = 1, \ldots, p_1, \qquad \mathbb{E}(Y_j) = 0, \quad j = 1, \ldots, p_2.$$

On the population level, CCA is designed to extract the most correlated linear combinations between two sets of random variables sequentially: The ith pair of *canonical variables*

$$U_i = \boldsymbol{\phi}_i^\top \boldsymbol{x}, \qquad V_i = \boldsymbol{\psi}_i^\top \boldsymbol{y}$$

maximizes

$$\lambda_i = \operatorname{Corr}(U_i, V_i)$$

such that U_i and V_i have unit variances and they are uncorrelated to all previous pairs of canonical variables. Here $(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i)$ is called the ith pair of *canonical loadings* and λ_i is the ith *canonical correlation*.

It is well known in multivariate statistical analysis that the canonical loadings can be found recursively by the following criterion:

$$\begin{aligned}(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i) = \arg\max \quad & \boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{\psi} \\ \text{subject to } & \boldsymbol{\phi}^\top \boldsymbol{\Sigma}_x \boldsymbol{\phi} = 1, \qquad \boldsymbol{\psi}^\top \boldsymbol{\Sigma}_y \boldsymbol{\psi} = 1; \\ & \boldsymbol{\phi}^\top \boldsymbol{\Sigma}_x \boldsymbol{\phi}_j = 0, \qquad \boldsymbol{\psi}^\top \boldsymbol{\Sigma}_y \boldsymbol{\psi}_j = 0, \quad \forall 1 \le j \le i-1.\end{aligned} \tag{1.3}$$

Although this criterion is a nonconvex optimization, it can be obtained easily by spectral methods: Define

$$\boldsymbol{\Phi} := [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{p_1 \wedge p_2}], \qquad \boldsymbol{\Psi} := [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{p_1 \wedge p_2}], \qquad \boldsymbol{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_{p_1 \wedge p_2}). \tag{1.4}$$

Then $\lambda_1, \ldots, \lambda_{p_1 \wedge p_2}$ are singular values of $\boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_y^{-1/2}$, and $\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_y^{1/2} \boldsymbol{\Psi}$ are actually left and right singular vectors of $\boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_y^{-1/2}$, respectively.

Let $(\boldsymbol{x}_i^\top, \boldsymbol{y}_i^\top) = (X_{i1}, \ldots, X_{ip_1}, Y_{i1}, \ldots, Y_{ip_2})$, $i = 1, \ldots, n$ be a random sample of $(\boldsymbol{x}^\top, \boldsymbol{y}^\top) = (X_1, \ldots, X_{p_1}, Y_1, \ldots, Y_{p_2})$. Moreover, denote the two data matrices

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \qquad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_n^\top \end{bmatrix}$$

Generally speaking, CCA estimation problems refer to how to estimate the canonical loadings $\{(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i)\}_{i=1}^{p_1 \wedge p_2}$ and the corresponding estimates for the canonical variables

$$\widehat{U}_i = \widehat{\boldsymbol{\phi}}_i^\top \boldsymbol{x}, \qquad \widehat{V}_i = \widehat{\boldsymbol{\psi}}_i^\top \boldsymbol{y}, \quad i = 1 \ldots, p_1 \wedge p_2, \tag{1.5}$$

from the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. Analogous to (1.4), we define the matrices of estimated canonical loadings

$$\widehat{\boldsymbol{\Phi}} := [\widehat{\boldsymbol{\phi}}_1, \ldots, \widehat{\boldsymbol{\phi}}_{p_1 \wedge p_2}], \qquad \widehat{\boldsymbol{\Psi}} := [\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_{p_1 \wedge p_2}]. \tag{1.6}$$

For example, when $n > p_1 + p_2$, the *sample canonical loadings* are defined recursively by

$$\begin{aligned}(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i) = \arg\max \quad & \boldsymbol{\phi}^\top \widehat{\boldsymbol{\Sigma}}_{xy} \boldsymbol{\psi} \\ \text{subject to } & \boldsymbol{\phi}^\top \widehat{\boldsymbol{\Sigma}}_x \boldsymbol{\phi} = 1, \qquad \boldsymbol{\psi}^\top \widehat{\boldsymbol{\Sigma}}_y \boldsymbol{\psi} = 1; \\ & \boldsymbol{\phi}^\top \widehat{\boldsymbol{\Sigma}}_x \boldsymbol{\phi}_j = 0, \qquad \boldsymbol{\psi}^\top \widehat{\boldsymbol{\Sigma}}_y \boldsymbol{\psi}_j = 0, \quad \forall 1 \le j \le i-1.\end{aligned} \tag{1.7}$$

where $\widehat{\boldsymbol{\Sigma}}_x = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{p_1 \times p_1}$, $\widehat{\boldsymbol{\Sigma}}_y = \frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{Y} \in \mathbb{R}^{p_2 \times p_2}$, $\widehat{\boldsymbol{\Sigma}}_{xy} = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{Y} \in \mathbb{R}^{p_1 \times p_2}$ are the sample covariance matrices. As with the population canonical loadings, the matrices of sample canonical loadings $\widehat{\boldsymbol{\Sigma}}_x^{1/2} \widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Sigma}}_y^{1/2} \widehat{\boldsymbol{\Psi}}$ are actually left and right singular vectors of $\widehat{\boldsymbol{\Sigma}}_x^{-1/2} \widehat{\boldsymbol{\Sigma}}_{xy} \widehat{\boldsymbol{\Sigma}}_y^{-1/2}$,

respectively. Subsequently, the *sample canonical variables* are the linear combinations by the sample canonical loadings as defined in (1.5).

1.2. Canonical variables versus canonical loadings

For any predetermined k, any estimated canonical loading $\{(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i)\}_{i=1}^k$ and the corresponding estimated canonical variables $\{(\widehat{U}_i, \widehat{V}_i)\}_{i=1}^k$ defined in (1.5), to quantify the estimation accuracy, generally speaking, we can either focus on measuring the differences between the canonical loadings $\{(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i)\}_{i=1}^k$ and $\{(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i)\}_{i=1}^k$ or measuring the differences between the canonical variables $\{(U_i, V_i)\}_{i=1}^k$ and $\{(\widehat{U}_i, \widehat{V}_i)\}_{i=1}^k$. Here $\boldsymbol{x}$, $\boldsymbol{y}$ in the definition of $\{(U_i, V_i)\}_{i=1}^k$ and $\{(\widehat{U}_i, \widehat{V}_i)\}_{i=1}^k$ are independent of the samples based on which $\{(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i)\}_{i=1}^k$ are constructed. Therefore, for the discrepancy between the canonical variables, there is an extra layer of randomness.

As discussed above, in modern machine learning applications, the leading sample canonical loadings are used for dimension reduction, i.e., for a new observation $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, ideally we hope to use the corresponding values of the canonical variables $(u_i = \boldsymbol{\phi}_i^\top \boldsymbol{x}_0)_{i=1}^k$ and $(v_i = \boldsymbol{\psi}_i^\top \boldsymbol{y}_0)_{i=1}^k$ to represent the observation in a low dimension space. Empirically, the actual low dimensional representations are $(\hat{u}_i = \widehat{\boldsymbol{\phi}}_i^\top \boldsymbol{x}_0)_{i=1}^k$ and $(\hat{v}_i = \widehat{\boldsymbol{\psi}}_i^\top \boldsymbol{y}_0)_{i=1}^k$. Therefore, the discrepancy between the ideal dimension reduction and actual dimension reduction should be explained by how well $\{(\widehat{U}_i, \widehat{V}_i)\}_{i=1}^k$ approximate $\{(U_i, V_i)\}_{i=1}^k$. Consequently, we choose to quantify the difference between the sample and population canonical variables instead of the canonical loadings.

1.3. Linear span

However, there are still many options to quantify how well the sample canonical variables approximate their population correspondents. To choose suitable losses, it is convenient to come back to specific applications to get some inspiration. Consider the model of multi-view sufficient dimension reduction [10], which studies how to predict Z using two sets of predictors denoted by $\boldsymbol{x} = [X_1, \ldots, X_{p_1}]^\top$ and $\boldsymbol{y} = [Y_1, \ldots, Y_{p_2}]^\top$, where the joint covariance of $(Z, \boldsymbol{x}, \boldsymbol{y})$ is

$$\operatorname{Cov}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ Z \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} & \boldsymbol{\sigma}_{xz} \\ \boldsymbol{\Sigma}_{xy}^\top & \boldsymbol{\Sigma}_y & \boldsymbol{\sigma}_{yz} \\ \boldsymbol{\sigma}_{xz}^\top & \boldsymbol{\sigma}_{yz}^\top & \sigma_z^2 \end{bmatrix}.$$

It was proven in [10] that under certain assumptions, the leading k canonical variables $U_1, \ldots, U_k$ are sufficient dimension reduction for the linear prediction of Z; That is, the best linear predictor of Z based on $X_1, \ldots, X_{p_1}$ is the same as the best linear predictor based on $U_1, \ldots, U_k$. (Similarly, the best linear predictor of Z based on $Y_1, \ldots, Y_{p_2}$ is the same as the best linear predictor based on $V_1, \ldots, V_k$.)

Notice that the best linear predictor is actually determined by the set of all linear combinations of $U_1, \ldots, U_k$ (referred to as the "model space" in the literature of linear prediction), which we denote as $\operatorname{span}(U_1, \ldots, U_k)$. Inspired by [10], we propose to quantify the discrepancy between

$\{U_i\}_{i=1}^k$ and $\{\widehat{U}_i\}_{i=1}^k$ by the discrepancy between the corresponding subspaces $\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$ and $\text{span}(U_1, \ldots, U_k)$ (and similarly measure the difference between $\{V_i\}_{i=1}^k$ and $\{\widehat{V}_i\}_{i=1}^k$ by the distance between $\text{span}(\widehat{V}_1, \ldots, \widehat{V}_k)$ and $\text{span}(V_1, \ldots, V_k)$).

1.4. Hilbert spaces and principal angles

In this section, we define the discrepancy between $\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$ and $\text{span}(U_1, \ldots, U_k)$ by introducing a Hilbert space. Conditional on the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, both $\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$ and $\text{span}(U_1, \ldots, U_k)$ are composed by linear combinations of $X_1, \ldots, X_{p_1}$. Denote the set of all possible linear combinations as

$$\mathcal{H} = \text{span}(X_1, \ldots, X_{p_1}). \tag{1.8}$$

Moreover, for any $X_1, X_2 \in \mathcal{H}$, we define a bilinear function $\langle X_1, X_2 \rangle := \text{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2)$. It is easy to show that $\langle \cdot, \cdot \rangle$ is an inner product and $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a p_1-dimensional Hilbert space, which is isomorphic to $\mathbb{R}^{p_1}$. With this covariance-based inner product, we know both $\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$ and $\text{span}(U_1, \ldots, U_k)$ are subspaces of $\mathcal{H}$, so it is natural to define their discrepancy based on their principal angles $\frac{\pi}{2} \geq \theta_1 \geq \cdots \geq \theta_k \geq 0$. In the literature of statistics and linear algebra, the following two loss functions for subspaces are usually used

$$\mathcal{L}_{\max}\big(\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k), \text{span}(U_1, \ldots, U_k)\big) = \sin^2(\theta_1)$$

and

$$\mathcal{L}_{\text{ave}}\big(\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k), \text{span}(U_1, \ldots, U_k)\big) = \frac{1}{k}\big(\sin^2(\theta_1) + \cdots + \sin^2(\theta_k)\big)$$

In spite of a somewhat abstract definition, we have the following clean formula for these two losses.

Theorem 1.1. *Suppose for any $p_1 \times k$ matrix $\boldsymbol{A}$, $\boldsymbol{P_A}$ represents the orthogonal projector onto the column span of $\boldsymbol{A}$. Assume the observed sample is fixed. Then*

$$\begin{aligned}
\mathcal{L}_{\text{ave}}\big(\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k), \text{span}(U_1, \ldots, U_k)\big) &= \frac{1}{2k}\|\boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Phi}_{1:k}}\|_{\text{F}}^2 \\
&= \frac{1}{k}\big\|(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_{\text{F}}^2 \\
&= \frac{1}{k}\min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big[\big\|\boldsymbol{u}^\top - \widehat{\boldsymbol{u}}^\top \boldsymbol{Q}\big\|_2^2 | \widehat{\boldsymbol{\Phi}}_{1:k}\big] \\
&:= \mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})
\end{aligned} \tag{1.9}$$

and

$$\begin{aligned}
\mathcal{L}_{\max}\big(\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k), \text{span}(U_1, \ldots, U_k)\big) &= \|\boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Phi}_{1:k}}\|^2 \\
&= \big\|(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2
\end{aligned}$$

$$= \max_{\|\boldsymbol{g}\|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big[\big((\boldsymbol{u}^\top - \widehat{\boldsymbol{u}}^\top \boldsymbol{Q})\boldsymbol{g}\big)^2 | \widehat{\boldsymbol{\Phi}}_{1:k}\big]$$

$$:= \mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}). \tag{1.10}$$

Here $\boldsymbol{\Phi}_{1:k} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k]$ is a $p_1 \times k$ matrix consisting of the leading k population canonical loadings for $\boldsymbol{x}$, and $\widehat{\boldsymbol{\Phi}}_{1:k}$ its estimate. Moreover $\boldsymbol{u}^\top := (U_1, \ldots, U_k)$ and $\hat{\boldsymbol{u}}^\top := (\widehat{U}_1, \ldots, \widehat{U}_k)$. By (1.5), we have $\boldsymbol{u}^\top = \boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k}$ and $\hat{\boldsymbol{u}}^\top = \boldsymbol{x}^\top \widehat{\boldsymbol{\Phi}}_{1:k}$.

1.5. Uniform upper bounds and minimax rates

The most important contribution of this paper is to establish sharp upper bounds for the estimation/prediction of CCA based on the proposed subspace losses $\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$ and $\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$. It is noteworthy that both upper bounds hold uniformly for all invertible $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ provided $n > C(p_1 + p_2)$ for some numerical constant C. Furthermore, in order to justify the sharpness of these bounds, we also establish minimax lower bounds under a family of stringent and localized parameter spaces. These results will be detailed in Section 2. Numerical simulations in Section 3 further validate our theoretical findings.

1.6. Notations and the organization

Throughout the paper, we use lower-case and upper-case non-bolded letters to represent fixed and random variables, respectively. We also use lower-case and upper-case bold letters to represent vectors (which could be either deterministic or random) and matrices, respectively. For any matrix $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ and vector $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{U}\|$, $\|\boldsymbol{U}\|_F$ denotes operator (spectral) norm and Frobenius norm respectively, $\|\boldsymbol{u}\|$ denotes the vector l_2 norm, $\boldsymbol{U}_{1:k}$ denotes the submatrix consisting of the first k columns of $\boldsymbol{U}$, and $\boldsymbol{P}_{\boldsymbol{U}}$ stands for the projection matrix onto the column space of $\boldsymbol{U}$. Moreover, we use $\sigma_{\max}(\boldsymbol{U})$ and $\sigma_{\min}(\boldsymbol{U})$ to represent the largest and smallest singular value of $\boldsymbol{U}$ respectively, and $\kappa(\boldsymbol{U}) = \sigma_{\max}(\boldsymbol{U})/\sigma_{\min}(\boldsymbol{U})$ to denote the condition number of the matrix. We use $\boldsymbol{I}_p$ for the identity matrix of dimension p and $\boldsymbol{I}_{p,k}$ for the submatrix composed of the first k columns of $\boldsymbol{I}_p$. Further, $\mathcal{O}(m, n)$ (and simply $\mathcal{O}(n)$ when $m = n$) stands for the set of $m \times n$ matrices with orthonormal columns and $\mathbb{S}_+^p$ denotes the set of $p \times p$ strictly positive definite matrices. For a random vector $\boldsymbol{x} \in \mathbb{R}^p$, $\text{span}(\boldsymbol{x}^\top) = \{\boldsymbol{x}^\top \boldsymbol{w}, \boldsymbol{w} \in \mathbb{R}^p\}$ denotes the subspace of all the linear combinations of $\boldsymbol{x}$. Other notations will be specified within the corresponding context.

In the following, we will introduce our main upper and lower bound results in Section 2. Various numerical simulations that illustrate our theoretical discoveries are demonstrated in Section 3. To highlight our contributions in the new loss functions and theoretical results, we will compare our results to existing work in the literature in Section 4. A summary of the significance of our theoretical results as well as future research topics are introduced in Section 5. All proofs are deferred to Section 6, Section 7 and the supplement article [20].

2. Theory

In this section, we introduce our main results on non-asymptotic upper and lower bounds for estimating CCA under the proposed loss functions. Recall that the sample canonical loadings are defined in (1.6) and the corresponding canonical variables are defined in (1.5). The following theorem provides upper bounds for the expected losses defined in (1.9) and (1.10).

Theorem 2.1 (Upper bound). *Suppose* $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *where* $\mathbf{\Sigma}$ *is defined as in* (1.2). *Assume* $\mathbf{\Sigma}_x$ *and* $\mathbf{\Sigma}_y$ *are invertible. Recall that the population canonical correlations* $\lambda_1, \ldots, \lambda_{p_1 \wedge p_2}$ *are singular values of* $\mathbf{\Sigma}_x^{-1/2} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_y^{-1/2}$, *and we assume* $\lambda_k > \lambda_{k+1}$ *for some predetermined* k. *Then there exist universal positive constants* γ, C, C_0 *such that if* $n \geq C(p_1 + p_2)$, *the top-*k *sample canonical loadings* $\widehat{\mathbf{\Phi}}_{1:k}$ *defined in* (1.7) *satisfy*

$$\mathbb{E}\big[\mathcal{L}_{\text{ave}}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \leq C_0 \left[\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1 - k}{n} + \frac{(p_1+p_2)^2}{n^2(\lambda_k - \lambda_{k+1})^4} + e^{-\gamma(p_1 \wedge p_2)} \right]$$

$$\mathbb{E}\big[\mathcal{L}_{\max}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \leq C_0 \left[\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1}{n} + \frac{(p_1+p_2)^2}{n^2(\lambda_k - \lambda_{k+1})^4} + e^{-\gamma(p_1 \wedge p_2)} \right].$$

In the special case $\lambda_k = 1$, *there holds*

$$\mathcal{L}_{\max}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k}) = \mathcal{L}_{\text{ave}}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k}) = 0, \quad a.s.$$

The upper bounds for $\widehat{\mathbf{\Psi}}_{1:k}$ *can be obtained by switching* p_1 *and* p_2.

Some features of the above upper bounds are worth highlighting

- For both the loss functions $\mathcal{L}_{\text{ave}}$ and $\mathcal{L}_{\max}$, we establish the upper bounds with the factor $(1-\lambda_k^2)(1-\lambda_{k+1}^2)/(\lambda_k - \lambda_{k+1})^2$ in the leading order term. This implies some interesting phenomena of the estimation of canonical variables, particularly when λ_k is close to 1: When λ_k is close to 1, the factor $(1-\lambda_k^2)(1-\lambda_{k+1}^2)/(\lambda_k - \lambda_{k+1})^2$ is close to zero even if λ_{k+1} is very close to λ_k. To the best of our knowledge, we first show these unique features of CCA estimates under the non-asymptotic setups. These properties will be further illustrated in Section 3 and explained in Section 4.2.
- We first decouple the estimation error bound of $\widehat{\mathbf{\Phi}}_{1:k}$ from p_2 without assuming the residual canonical correlations are zeros. More details will be given in Section 4.2.

Since we pursue a non-asymptotic theoretical framework for CCA estimates, and the loss functions we propose are nonstandard in the literature, the standard minimax lower bound results in parametric maximum likelihood estimates do not apply straightforwardly. Instead, we turn to the nonparametric minimax lower bound frameworks, particularly those in PCA and CCA; See, for example, [4,13,26]. Compared to these existing works, the technical novelties of our results and proofs are summarized in Sections 4.3 and the supplement article [20].

We define the parameter space $\mathcal{F}(p_1, p_2, k, \lambda_k, \lambda_{k+1}, \kappa_1, \kappa_2)$ as the collection of joint covariance matrices $\mathbf{\Sigma}$ satisfying

1. The condition numbers $\kappa(\mathbf{\Sigma}_x) = \kappa_1$ and $\kappa(\mathbf{\Sigma}_y) = \kappa_2$;
2. The kth and $(k+1)$th singular values of $\mathbf{\Sigma}_x^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1/2}$ are λ_k and λ_{k+1}, respectively, and $\lambda_k > \lambda_{k+1}$.

For the rest of the paper, we will use the shorthand $\mathcal{F}$ to represent this parameter space for simplicity.

Theorem 2.2 (Lower bound). *There exists a universal constant c independent of n, p_1, p_2 and $\mathbf{\Sigma}$ such that*

$$\inf_{\widehat{\mathbf{\Phi}}_{1:k}} \sup_{\mathbf{\Sigma}\in\mathcal{F}} \mathbb{E}\big[\mathcal{L}_{\max}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \geq c^2 \left\{ \left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\}$$

$$\inf_{\widehat{\mathbf{\Phi}}_{1:k}} \sup_{\mathbf{\Sigma}\in\mathcal{F}} \mathbb{E}\big[\mathcal{L}_{\mathrm{ave}}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \geq c^2 \left\{ \left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\}.$$

The lower bounds for $\widehat{\mathbf{\Psi}}_{1:k}$ can be obtained by replacing p_1 with p_2.

This theorem shows that the lower bound is independent of the condition numbers $\kappa(\mathbf{\Sigma}_x) = \kappa_1$ and $\kappa(\mathbf{\Sigma}_y) = \kappa_2$. By combining Theorem 2.1 and Theorem 2.2 together, as long as the sample size large is large enough, we can achieve the following results of minimax rates.

Corollary 2.3. *When $p_1, p_2 \geq (2k) \vee C(\log n)$ and*

$$n \geq C \frac{(p_1 + p_2)(1 + p_2/p_1)}{(\lambda_k - \lambda_{k+1})^2 (1-\lambda_k^2)(1-\lambda_{k+1}^2)} \tag{2.1}$$

for some universal positive constant c, the minimax rates can be characterized by

$$\inf_{\widehat{\mathbf{\Phi}}_{1:k}} \sup_{\mathbf{\Sigma}\in\mathcal{F}} \mathbb{E}\big[\mathcal{L}_{\max}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \asymp \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1}{n},$$

$$\inf_{\widehat{\mathbf{\Phi}}_{1:k}} \sup_{\mathbf{\Sigma}\in\mathcal{F}} \mathbb{E}\big[\mathcal{L}_{\mathrm{ave}}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})\big] \asymp \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1}{n}.$$

3. Numerical simulations

The purpose of this section is to illustrate Theorem 2.1 in understanding the empirical performances of $\mathbb{E}[\mathcal{L}_{\max}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})]$, $\mathbb{E}[\mathcal{L}_{\mathrm{ave}}(\mathbf{\Phi}_{1:k}, \widehat{\mathbf{\Phi}}_{1:k})]$, $\mathbb{E}[\mathcal{L}_{\max}(\mathbf{\Psi}_{1:k}, \widehat{\mathbf{\Psi}}_{1:k})]$ and $\mathbb{E}[\mathcal{L}_{\mathrm{ave}}(\mathbf{\Psi}_{1:k}, \widehat{\mathbf{\Psi}}_{1:k})]$, particularly their dependency on the canonical correlation coefficients, dimensions p_1 and p_2, the sample size n, and the choice of k.

In the first numerical experiment, we choose $p_1 = 100$ and $p_2 = 150$. The canonical correlation coefficients are chosen as $\lambda_1 = 1$, $\lambda_2 = 0.99$, $\lambda_3 = 0.98$, $\ldots$, $\lambda_{100} = 0.01$. As to the population covariance $\mathbf{\Sigma}$, we choose $\mathbf{\Sigma}_x = \boldsymbol{I}_{p_1}$, $\mathbf{\Sigma}_y = \boldsymbol{I}_{p_2}$, and $\mathbf{\Sigma}_{xy} = [\mathbf{\Lambda}, \mathbf{0}] \in \mathbb{R}^{100\times 150}$.

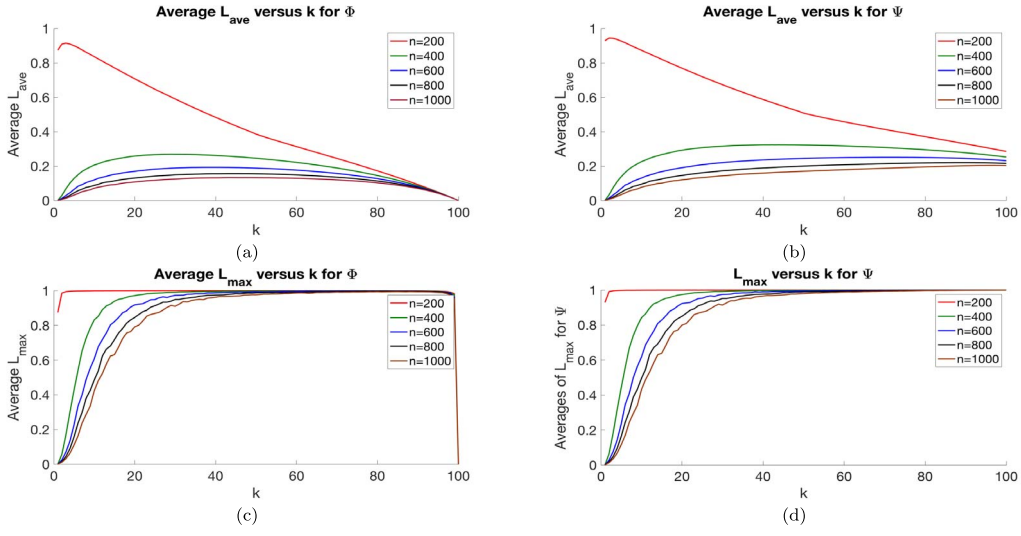

Figure 1. The problem parameters are set as $p_1 = 100$, $p_2 = 150$ and $\lambda_1 = 1$, $\lambda_2 = 0.99$, $\lambda_3 = 0.98$, $\ldots$, $\lambda_{100} = 0.01$. Expected losses are approximated by taking average over 100 independent Monte Carlo experiments. The sample sizes are chosen as $n = 200, 400, 600, 800, 1000$. (a) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ for $k = 1, 2, \ldots, 100$; (b) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$ for $k = 1, 2, \ldots, 100$; (c) Approximated $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ for $k = 1, 2, \ldots, 100$; (d) Approximated $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$ for $k = 1, 2, \ldots, 100$.

Here $\boldsymbol{\Lambda} = \text{diag}(1, 0.99, 0.98, \ldots, 0.01)$. We assume the i.i.d. sample of $(\boldsymbol{x}^\top, \boldsymbol{y}^\top)$ is generated from the distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. We study the empirical performances of $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$, $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$, $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ and $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$ in different choices of $n = 200, 400, 600, 800, 1000$ and $k = 1, 2, \ldots, 100$. The results are plotted in parts (a), (b), (c) and (d) in Figure 1. In order to approximate the expected losses, we implement 100 independent Monte Carlo experiments and take the average for each of the four losses. From these four figures, we can make the following observations that are consistent with our theoretical findings in Theorem 2.1:

- From figures (a), (b), (c) and (d), we see that as long as k gets close to 1, that is, λ_k and λ_{k+1} get close to 1, all four losses decrease to 0 for $n = 400, 600, 800, 1000$. In particular, in the case $k = 1$, the losses become exactly zero. This is consistent with Theorem 2.1 in which the leading term of upper bound is proportional to

$$\big(1 - \lambda_k^2\big)\big(1 - \lambda_{k+1}^2\big)/(\lambda_k - \lambda_{k+1})^2 \propto \big(1 - \lambda_k^2\big)\big(1 - \lambda_{k+1}^2\big) \tag{3.1}$$

  given $\lambda_k - \lambda_{k+1} = 0.01$ is independent of k. This phenomenon does not hold for the case $n = 200$ since $p_1 + p_2 > n$. It is perhaps surprising that even under the moderate sample size $n = 400$ compared to $p_1 + p_2 = 250$, all expected losses approach zero as long as k goes to 1.

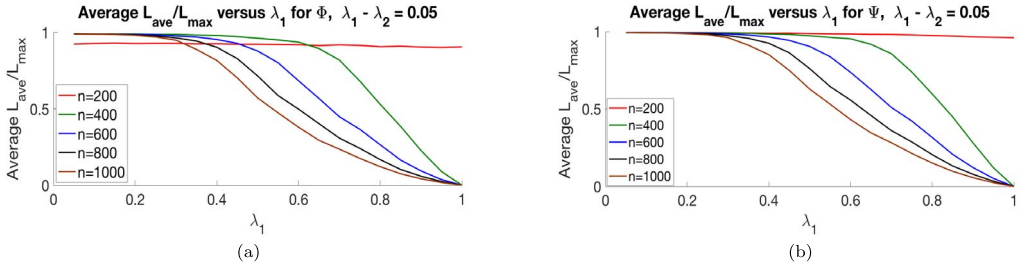

Figure 2. The problem parameters are set as $p_1 = 100$, $p_2 = 150$, $\lambda_2 = \lambda_1 - 0.05$ and $\lambda_3 = \cdots \lambda_{50} = 0$. We always set $k = 1$ and choose the sample sizes $n = 200, 400, 600, 800, 1000$. Expected losses are approximated by taking average over 100 independent Monte Carlo experiments. (a) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)]$ (or $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)]$) for $\lambda_1 = 1, 0.95, 0.9, \ldots, 0.05$; (b) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)]$ (or $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)]$) for $\lambda_1 = 1, 0.95, 0.9, \ldots, 0.05$.

- Figure (a) shows that when k approaches 100, $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ approaches 0 for all $n = 200, 400, 600, 800, 1000$. This fact can be partially explained by Theorem 2.1 in that the leading term of upper bound for $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ is proportional to $p_1 - k = 100 - k$.
- In each of figures (a), (b), (c) and (d), it is observed that the expected loss decreases if we fix k while increase the sample size n.
- By a careful comparison between Figures (a) and (b), we can conclude that $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ is in general no greater than $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$. This is also suggested by Theorem 2.1 in that the leading term of upper bound for $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})]$ is proportional to $p_1 - k = 100 - k$ while that for $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})]$ is proportional to $p_2 - k = 150 - k$.

In our second numerical experiment, we still set $p_1 = 100$ and $p_2 = 150$, but choose the canonical correlation coefficients as

$$\lambda_1 = 1, 0.95, 0.9, \ldots, 0.05, \qquad \lambda_2 = \lambda_1 - 0.05, \qquad \lambda_3 = \cdots \lambda_{50} = 0.$$

As for the population covariance $\boldsymbol{\Sigma}$, by setting $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, 0, \ldots, 0)$, we still let

$$\boldsymbol{\Sigma}_x = \boldsymbol{I}_{p_1}, \qquad \boldsymbol{\Sigma}_y = \boldsymbol{I}_{p_2}, \qquad \boldsymbol{\Sigma}_{xy} = [\boldsymbol{\Lambda}, \boldsymbol{0}] \in \mathbb{R}^{50 \times 200}.$$

In this case we set $k = 1$ and choose $n = 200, 400, 600, 800, 1000$. Notice that because $\lambda_1 - \lambda_2 = 0.05$, (3.1) still holds. We plot the approximated losses for different values of λ_1 in Figure 2.

As expected from Theorem 2.1, for $n = 400, 600, 800, 1000$, all four expected losses approach 0 as λ_1 approaches 1. When $n = 200 < p_1 + p_2$, our numerical result shows that sample CCA is never consistent no matter how λ_1 gets close to 1. Note that $\mathcal{L}_{\max}(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}_{\text{ave}}(\boldsymbol{u}, \boldsymbol{v})$ when $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors, so we have $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)] = \mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)]$ and $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)] = \mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)]$.

Our third numerical experiment is similar to the second, but we fix λ_1 instead of the eigengap. To be specific, we set $p_1 = 100$, $p_2 = 150$, $\lambda_1 = 0.99$, $\lambda_2 = 0.98, 0.96, 0.94, \ldots, 0$, $\lambda_3 = \cdots \lambda_{50} = 0$. The joint covariance $\boldsymbol{\Sigma}$ is defined the same as before. We still set $k = 1$ and $n = 400, 600, 800, 1000$. We plot the approximated losses for different values of λ_1 in Figure 3.

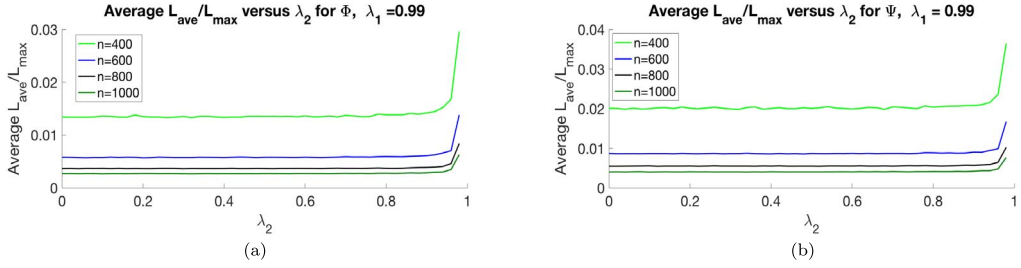

Figure 3. The problem parameters are set as $p_1 = 100$, $p_2 = 150$, $\lambda_1 = 0.99$ and $\lambda_3 = \cdots \lambda_{50} = 0$. We always set $k = 1$ and choose the sample sizes $n = 400, 600, 800, 1000$. Expected losses are approximated by taking average over 100 independent Monte Carlo experiments. (a) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)]$ (or $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Phi}_1, \widehat{\boldsymbol{\Phi}}_1)]$) for $\lambda_2 = 0.98, 0.96, 0.94, \ldots, 0$; (b) Approximated $\mathbb{E}[\mathcal{L}_{\text{ave}}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)]$ (or $\mathbb{E}[\mathcal{L}_{\max}(\boldsymbol{\Psi}_1, \widehat{\boldsymbol{\Psi}}_1)]$) for $\lambda_2 = 0.98, 0.96, 0.94, \ldots, 0$.

In the above setup, the factor

$$\big(1 - \lambda_1^2\big)\big(1 - \lambda_2^2\big)/(\lambda_1 - \lambda_2)^2 \approx 0.01\big(1 - \lambda_2^2\big)/(0.99 - \lambda_2)^2$$

is very small even if λ_2 is close to $\lambda_1 = 0.99$. Our numerical simulations are consistent with our theoretical results.

4. Related work and our contributions

Recently, the non-asymptotic rate of convergence of CCA has been studied by [13,14] under a sparse setup and by [5] under the non-sparse setup. The first version of [5] appeared on arXiv almost at the same time as the first version of our paper was posted. In this section, we state our contributions by detailed comparison with these works.

4.1. Novel loss functions

We proposed new loss functions based on the principal angles between the subspace spanned by the population canonical variates and that spanned by the estimated canonical variates. In contrast, [14] proposed and studied the loss $\overline{\mathcal{L}}_{\text{ave}}$; [5] proposed $\overline{\mathcal{L}}_{\max}$ and studied both $\overline{\mathcal{L}}_{\text{ave}}$ and $\overline{\mathcal{L}}_{\max}$, where

$$\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = \min_{\boldsymbol{Q} \in \mathcal{O}(k,k)} \mathbb{E}\big[\big\|\boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k} - \boldsymbol{x}^\top \widehat{\boldsymbol{\Phi}}_{1:k} \boldsymbol{Q}\big\|_2^2 \mid \widehat{\boldsymbol{\Phi}}_{1:k}\big],$$

$$\overline{\mathcal{L}}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = \max_{\boldsymbol{g} \in \mathbb{R}^k, |\boldsymbol{g}|=1} \min_{\boldsymbol{Q} \in \mathcal{O}(k,k)} \mathbb{E}\big[\big(\big(\boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k} - \boldsymbol{x}^\top \widehat{\boldsymbol{\Phi}}_{1:k} \boldsymbol{Q}\big)\boldsymbol{g}\big)^2 \mid \widehat{\boldsymbol{\Phi}}_{1:k}\big].$$

$\overline{\mathcal{L}}_{\text{ave}}$ and $\overline{\mathcal{L}}_{\max}$ resemble our loss functions $\mathcal{L}_{\text{ave}}$ and $\mathcal{L}_{\max}$ respectively. By Theorem 1.1, we also have

$$\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = 2 \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big[\big\|\boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k} - \boldsymbol{x}^\top \widehat{\boldsymbol{\Phi}}_{1:k}\boldsymbol{Q}\big\|_2^2 \mid \widehat{\boldsymbol{\Phi}}_{1:k}\big]$$

$$\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = \max_{\boldsymbol{g}\in\mathbb{R}^k, |\boldsymbol{g}|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big[\big(\big(\boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k} - \boldsymbol{x}^\top \widehat{\boldsymbol{\Phi}}_{1:k}\boldsymbol{Q}\big)\boldsymbol{g}\big)^2 \mid \widehat{\boldsymbol{\Phi}}_{1:k}\big]$$

Straightforward comparison between these two expressions yields

$$\begin{aligned} \mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) &\le 2\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) \\ \mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) &\le \overline{\mathcal{L}}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) \end{aligned} \tag{4.1}$$

However, $\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$ and $\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$ are not equivalent up to a constant, particularly when λ_k is close to 1, and neither are $\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$ and $\overline{\mathcal{L}}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})$. In fact, we can prove that as long as $n > \max(p_1, p_2)$, if $\lambda_k = 1 > \lambda_{k+1}$, then

$$\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = \mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = 0,$$

while almost surely $\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) \neq 0$ and $\overline{\mathcal{L}}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) \neq 0$.

To illustrate this comparison, we can consider the following very simple simulation: Suppose $p_1 = p_2 = 2$, $n = 3$ and $\boldsymbol{\Sigma}_x = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $\boldsymbol{\Sigma}_y = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $\boldsymbol{\Sigma}_{xy} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0.5 \end{smallmatrix}\right]$. In this setup, we know the population canonical correlation coefficients are $\lambda_1 = 1$ and $\lambda_2 = 0.5$, and the leading canonical loadings are $\boldsymbol{\phi}_1 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\boldsymbol{\psi}_1 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$. In our simulation, we generated the following data matrices

$$\boldsymbol{X} = \begin{bmatrix} 0.0736 & 1.5496 \\ 1.5390 & -0.0415 \\ 0.9331 & -0.4776 \end{bmatrix}, \qquad \boldsymbol{Y} = \begin{bmatrix} 0.0736 & 2.8982 \\ 1.5390 & -1.2214 \\ 0.9331 & 2.5931 \end{bmatrix}.$$

Furthermore, we can obtain the sample canonical correlations $\widehat{\lambda}_1 = 1$ and $\widehat{\lambda}_2 = 0.5210$, as well as the leading sample canonical loadings $\widehat{\boldsymbol{\phi}}_1 = \left[\begin{smallmatrix} -0.9616 \\ 0 \end{smallmatrix}\right]$ and $\widehat{\boldsymbol{\psi}}_1 = \left[\begin{smallmatrix} -0.9616 \\ 0 \end{smallmatrix}\right]$. Then $\mathcal{L}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) = \mathcal{L}_{\max}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) = 0$ while $\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \neq 0$, $\overline{\mathcal{L}}_{\max}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \neq 0$.

This numerical example clearly shows that the sample CCA can exactly identify that among all linear combinations of X_1 and X_2 and all linear combinations of Y_1 and Y_2, aX_1 and bY_1 are mostly correlated. Our loss functions $\mathcal{L}_{\text{ave}}$ and $\mathcal{L}_{\max}$ do characterize this exact identification, whereas $\overline{\mathcal{L}}_{\text{ave}}$ and $\overline{\mathcal{L}}_{\max}$ do not.

Moreover, the following joint loss was studied in [13]:

$$\overline{\mathcal{L}}_{\text{joint}}\big((\boldsymbol{\Phi}_{1:k}, \boldsymbol{\Psi}_{1:k}), (\widehat{\boldsymbol{\Phi}}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})\big) = \big\|\widehat{\boldsymbol{\Phi}}_{1:k}\widehat{\boldsymbol{\Psi}}_{1:k}^\top - \boldsymbol{\Phi}_{1:k}\boldsymbol{\Psi}_{1:k}^\top\big\|_{\text{F}}^2.$$

Similarly, $\overline{\mathcal{L}}_{\text{joint}}((\boldsymbol{\Phi}_{1:k}, \boldsymbol{\Psi}_{1:k}), (\widehat{\boldsymbol{\Phi}}_{1:k}, \widehat{\boldsymbol{\Psi}}_{1:k})) \neq 0$ almost surely under the special case $\lambda_k = 1 > \lambda_{k+1}$.

Finally, if we denote by $\boldsymbol{L}$ and $\widehat{\boldsymbol{L}}$ the first k singular vectors of $\boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1/2}$ and $\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2}$, respectively, then $\boldsymbol{\Phi}_{1:k} = \boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{L}$ and $\widehat{\boldsymbol{\Phi}}_{1:k} = \widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{L}}$. In Table 3 of [5], a

Table 1. Comparison between our results and that in [5]

	Cai and Zhang 2016	Our work
Loss function	$\overline{\mathcal{L}}_{\text{ave}}(\geq \mathcal{L}_{\text{ave}})$	$\mathcal{L}_{\text{ave}}$
Sample size	$n > C(\frac{p_1+\sqrt{p_1 p_2}}{\lambda_k^2} + \frac{p_2}{\lambda_k^{4/3}})$	$n > C(p_1 + p_2)$
$\lambda_{k+1} = \cdots = \lambda_{p_1} = 0$	Yes	No
Upper Bound Rates	$\frac{p_1}{n\lambda_k^2} + \frac{p_1 p_2}{n^2\lambda_k^4}$	$\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k-\lambda_{k+1})^2}\frac{p_1-k}{n} + \frac{(p_1+p_2)^2}{n^2(\lambda_k-\lambda_{k+1})^4} + e^{-\gamma(p_1\wedge p_2)}$

loss $\|\sin\boldsymbol{\Theta}(\boldsymbol{L}, \hat{\boldsymbol{L}})\|_{\text{F}}$ was proposed and it is equivalent to $\frac{1}{2}\|\boldsymbol{P}_{\boldsymbol{L}} - \boldsymbol{P}_{\hat{\boldsymbol{L}}}\|_{\text{F}}^2$. However, (1.9) in Theorem 1.1 gives

$$\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) = \frac{1}{2k}\|\boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Phi}_{1:k}}\|_{\text{F}}^2 = \frac{1}{2k}\|\boldsymbol{P}_{\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{L}}}\|_{\text{F}}^2.$$

which is different from $\|\sin\boldsymbol{\Theta}(\boldsymbol{L}, \hat{\boldsymbol{L}})\|_F$. This is not surprising since the loss functions discussed in Table 3 of [5] are regarding estimation of canonical loadings, while ours are regarding estimation of canonical variables.

4.2. Sharper upper bounds

Regardless of loss functions, we explain in the following why Theorem 2.1 implies sharper upper bounds than the existing rates in [13,14] and [5] under the nonsparse case. Our discussion is focused on $\mathcal{L}_{\text{ave}}$ in the following discussion while the discussion for $\mathcal{L}_{\max}$ is similar.

Notice that if we only apply Wedin's sin-theta law, that is, replacing the fine bound Lemma 7.4 with the rough bound Lemma 7.2 (also see [13] for similar ideas), we can obtain the following rough bound:

$$\mathbb{E}\big[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})\big] \leq C_0\bigg[\frac{p_1 + p_2}{n(\lambda_k - \lambda_{k+1})^2}\bigg]. \tag{4.2}$$

In order to decouple the estimation error bound of $\widehat{\boldsymbol{\Phi}}_{1:k}$ from p_2, both [14] and [5] assume the residual canonical correlations are zero, that is,

$$\lambda_{k+1} = \cdots = \lambda_{p_1\wedge p_2} = 0.$$

This assumption is essential for proofs in both [14] and [5] under certain sample size conditions. We got rid of this assumption by developing new proof techniques and these techniques actually work for $\overline{\mathcal{L}}_{\text{ave}}$, $\overline{\mathcal{L}}_{\max}$ as well. A detailed comparison between our result and that in [5] is summarized in Table 1 (The results of [14] in the non-sparse regime can be implied by [5] under milder sample size conditions).

Perhaps the most striking contribution of our upper bound is that we first derive the factors $(1 - \lambda_k^2)$ and $(1 - \lambda_{k+1}^2)$ in the literature of non-asymptotic CCA estimate. We now explain why these factors are essential when leading canonical correlation coefficients are close to 1.

Example 1: $\lambda_k = 1$ *and* $\lambda_{k+1} = 0$

Consider the example that $k = 1$, $p_1 = p_2 := p \gg \log n$, $\lambda_1 = 1$ and $\lambda_2 = 0$. Then Theorem 2.1 implies that $\mathbb{E}\mathcal{L}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) = 0$, while the rates in [5][1] imply that

$$\mathbb{E}\mathcal{L}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \le 2\mathbb{E}\overline{\mathcal{L}}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \le C\frac{p}{n}.$$

Example 2: *Both* λ_k *and* λ_{k+1} *are close to* 1

Consider the example that $k = 1$, $p_1 = p_2 := p \gg \log n$, $\lambda_1 = 1 - \sqrt[4]{\frac{p}{n}}$ and $\lambda_2 = 1 - 2\sqrt[4]{\frac{p}{n}}$. Then our bound rates $\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2}\frac{p_1 - k}{n} + \frac{(p_1+p_2)^2}{n^2(\lambda_k - \lambda_{k+1})^4} + e^{-\gamma(p_1 \wedge p_2)}$ actually imply that

$$\mathbb{E}\mathcal{L}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \le C\frac{p}{n},$$

while the rough rates (4.2) by Wedin's sin-theta law implies

$$\mathbb{E}\mathcal{L}_{\text{ave}}(\boldsymbol{\phi}_1, \widehat{\boldsymbol{\phi}}_1) \le C\sqrt{\frac{p}{n}}.$$

This shows that our upper bound rates could be much sharper than the rough rates (4.2) when both λ_k and λ_{k+1} are close to 1.

New proof techniques and connection to asymptotic theory

To the best of our knowledge, none of the analysis in [5,13,14] can be used to obtain the multiplicative factor $(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)/(\lambda_k - \lambda_{k+1})^2$ in the first order term of the upper bound, even under the strong condition that $\lambda_{k+1} = \cdots = \lambda_{p_1 \wedge p_2} = 0$.

Following a different path, we do careful non-asymptotic entry-wise perturbation analysis of the estimating equations of CCA to avoid the loss of precision caused by applying matrix inequalities in the early stage of the proof. The main challenge is to analyze the properties of matrix Hadamard products, especially to derive tight operator norm bounds for certain Hadamard products. We are lucky to find a divide-and-conquer approach ($\lambda_k \ge \frac{1}{2}$ and $\lambda_k < \frac{1}{2}$ in the proof of Lemma 7.4) to decompose the target matrices into simple-structure matrices where we can apply the tools developed in Lemma 7.6.

The asymptotic distribution of the canonical loadings $\{(\widehat{\boldsymbol{\phi}}_i, \widehat{\boldsymbol{\psi}}_i)\}_{i=1}^{p_1 \wedge p_2}$ has been studied in [1] under the assumption that all the canonical correlations are distinct and $\lambda_1 \ne 1$. Since we focus on subspaces, we only require $\lambda_k > \lambda_{k+1}$ for the given k. Both [1] and our work are based on analyzing the estimating equations (7.5) of CCA. Our analysis is more involved because completely novel techniques are required to obtain the factor $(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)$ in the non-asymptotic framework.

[1]The result was also effectively proven in [14], since their assumption that λ_1 is bounded away from 1 is not necessary for the derivation of their upper bounds.

4.3. Sharper lower bounds under parameter spaces with fixed λ_k and λ_{k+1}

The minimax lower bounds for the estimation rates of CCA were first established in [13,14] under the losses $\overline{\mathcal{L}}_{\text{joint}}$ and $\overline{\mathcal{L}}_{\text{ave}}$. However, the parameter space discussed in [14] requires $\lambda_{k+1} = 0$. Moreover, the parameter space in [13] is parameterized by λ satisfying $\lambda_k \geq \lambda$, but λ_{k+1} is not specified. In fact, they also constructed the hypothesis class with $\lambda_{k+1} = 0$ and the resulting minimax lower bound is proportional to $\frac{1}{\lambda^2}$.

However, this minimax lower bound is not sharp when λ_k and λ_{k+1} are close. Suppose $p_1 = p_2 := p$, $k = 1$, $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{2} - \sqrt{\frac{p}{n}}$. Our minimax lower bound in Theorem 2.2 leads to

$$\inf_{\widehat{\boldsymbol{\Phi}}_{1:k}} \sup_{\boldsymbol{\Sigma} \in \mathcal{F}} \mathbb{E}\big[\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})\big] \geq O(1).$$

In contrast, to capture the fundamental limit of CCA estimates in this scenario under the framework of [13], one needs to choose λ to capture both λ_k and λ_{k+1}, i.e., $\lambda_{k+1} \leq \lambda \leq \lambda_k$ and hence $\lambda \approx 1/2$. Then the resulting minimax lower bound rate will be $\frac{p}{n\lambda^2} = O(\frac{p}{n})$, which is much looser than $O(1)$.

Technically speaking, we follow the analytical framework of [13] and [14], but the hypothesis classes construction requires any given $\lambda_{k+1} > 0$ instead of $\lambda_{k+1} = 0$, and this brings in new technical challenges. More detailed technical discussions are deferred to the supplement article [20].

5. Discussion

In this paper, we study the theoretical properties of canonical correlation analysis by investigating the estimation of the canonical variables. Two losses are proposed based on the principal angles between the linear spans determined by the sample canonical variates and those by the population correspondents. The estimation risks are upper bounded non-asymptotically, and these upper bounds illustrate how the population canonical correlation coefficients affect the estimation accuracy in a nontrivial manner. In particular, we derive $(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)/(\lambda_k - \lambda_{k+1})^2$ in the leading term for non-asymptotic CCA estimation, and this implies that when the leading canonical correlation is close to 1, the estimation of canonical variates can be significantly accurate even with small eigen-gaps. Various numerical simulations are conducted to illustrate our theoretical findings, and the optimality of upper bounds are also justified by our derivation of the same factor in the minimax lower bounds.

We leave several theoretical questions for future research: First, in Theorem 2.1, we discuss the case $\lambda_k = 1$ separately since this exact recovery result cannot be directly implied by the general result, and we are particularly interested how to improve the general upper bound in order to include this special case. Second, we are particularly interested in figuring out whether the factor $(1-\lambda_k^2)(1-\lambda_{k+1}^2)/(\lambda_k-\lambda_{k+1})^2$ should also appear in the second order term in the upper bounds as shown in Theorem 2.1. Actually we are also interested in whether the second order term can be removed but if this is true it must rely on techniques totally different ours. Third, we are interested in removing the implicit absolute constants in Theorem 2.1, that is, those for the upper

bounds and that for the sample size. In particular, we hope to establish similar upper bound results with the only assumption that $n > p_1 + p_2$. Fourth, it would be interesting to extend the current results to other CCA problems, such as kernel CCA and sparse CCA. Finally, we hope that our techniques such as operator norm bounds for Hadamard products could be useful for other multivariate statistical problems.

6. Proof of Theorem 1.1

Suppose the observed sample of $(\boldsymbol{x}, \boldsymbol{y})$ is fixed and consider the correlation between the two subspaces of $\mathcal{H}$ (defined in (1.8)): $\operatorname{span}(U_1, \ldots, U_k)$ and $\operatorname{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$. Let $(W_1, \widehat{W}_1)$, $(W_2, \widehat{W}_2)$, $\ldots$, $(W_k, \widehat{W}_k)$ be the first, second, $\ldots$, and kth pair of canonical variates between $U_1, \ldots, U_k$ and $\widehat{U}_1, \ldots, \widehat{U}_k$. Then $\operatorname{span}(W_1, \ldots, W_k) = \operatorname{span}(U_1, \ldots, U_k)$, $\operatorname{span}(\widehat{W}_1, \ldots, \widehat{W}_k) = \operatorname{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$ and $\langle W_i, W_j \rangle = \langle W_i, \widehat{W}_j \rangle = \langle \widehat{W}_i, \widehat{W}_j \rangle = 0$, for any $i \neq j$ and $\operatorname{Var}(W_i) = \operatorname{Var}(\widehat{W}_i) = 1$, for $i = 1, \ldots, k$.

By the definition of principal angles, we know $\angle(W_i, \widehat{W}_i)$ is actually the ith principal angle between $\operatorname{span}(U_1, \ldots, U_k)$ and $\operatorname{span}(\widehat{U}_1, \ldots, \widehat{U}_k)$, that is, $\theta_i := \angle(W_i, \widehat{W}_i)$. This implies that

$$k\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) := \sum_{i=1}^{k} \sin^2 \theta_i = \sum_{i=1}^{k} \bigl(1 - \bigl|\langle W_i, \widehat{W}_i \rangle\bigr|^2\bigr).$$

Since $U_1, \ldots, U_k, \widehat{U}_1, \ldots, \widehat{U}_k$ are linear combinations of $X_1, \ldots, X_{p_1}$, we can denote

$$\boldsymbol{w}^\top := (W_1, \ldots, W_k) = \boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{B}, \quad \text{and} \quad \hat{\boldsymbol{w}}^\top := (\widehat{W}_1, \ldots, \widehat{W}_k) = \boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \widehat{\boldsymbol{B}},$$

where $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k]$, $\widehat{\boldsymbol{B}} := [\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_k] \in \mathbb{R}^{p \times k}$.

By the definition of $\boldsymbol{w}$, we have

$$\boldsymbol{I}_k = \operatorname{Cov}(\boldsymbol{w}) = \boldsymbol{B}^\top \boldsymbol{\Sigma}_x^{-1/2} \operatorname{Cov}(\boldsymbol{x}) \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{B} = \boldsymbol{B}^\top \boldsymbol{B}$$

and similarly $\boldsymbol{I}_k = \widehat{\boldsymbol{B}}^\top \widehat{\boldsymbol{B}}$. Then $\boldsymbol{B}$, $\widehat{\boldsymbol{B}}$ are $p \times k$ basis matrices. Moreover, we have $\boldsymbol{b}_i^\top \widehat{\boldsymbol{b}}_j = \langle W_i, \widehat{W}_j \rangle = 0$, for all $i \neq j$. Moreover, we have

$$\operatorname{Diag}\bigl(\cos(\theta_1), \ldots, \cos(\theta_k)\bigr) = \operatorname{Cov}(\boldsymbol{w}, \hat{\boldsymbol{w}}) = \boldsymbol{B}^\top \boldsymbol{\Sigma}_x^{-1/2} \operatorname{Cov}(\boldsymbol{x}) \boldsymbol{\Sigma}_x^{-1/2} \widehat{\boldsymbol{B}} = \boldsymbol{B}^\top \widehat{\boldsymbol{B}}.$$

Notice that $\operatorname{span}(U_1, \ldots, U_k) = \operatorname{span}(W_1, \ldots, W_k)$, $(U_1, \ldots, U_k) = \boldsymbol{x}^\top \boldsymbol{\Phi}_{1:k}$, and $(W_1, \ldots, W_k) = \boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{B}$. Then

$$\boldsymbol{\Phi}_{1:k} = \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{B} \boldsymbol{C} \quad \Rightarrow \quad \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}_{1:k} = \boldsymbol{B} \boldsymbol{C}$$

for some nonsingular $k \times k$ matrix C. This implies that $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}_{1:k}$ have the same column space. Since $\boldsymbol{B} \in \mathbb{R}^{p \times k}$ is a basis matrix, we have

$$\boldsymbol{B} \boldsymbol{B}^\top = \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}_{1:k}}.$$

Similarly, we have

$$\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top} = \boldsymbol{P}_{\boldsymbol{\Sigma}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}}.$$

Straightforward calculation gives

$$\begin{aligned}\big\|\boldsymbol{B}\boldsymbol{B}^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\big\|_F^2 &= \operatorname{trace}\big(\boldsymbol{B}\boldsymbol{B}^{\top}\boldsymbol{B}\boldsymbol{B}^{\top} - \boldsymbol{B}\boldsymbol{B}^{\top}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\boldsymbol{B}\boldsymbol{B}^{\top} + \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\big)\\ &= 2k - 2\operatorname{trace}\big(\boldsymbol{B}^{\top}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\boldsymbol{B}\big)\\ &= 2k - 2\operatorname{trace}\big(\operatorname{Diag}\big(\cos^2(\theta_1),\ldots,\cos^2(\theta_k)\big)\big)\\ &= 2\big(\sin^2(\theta_1)+\cdots+\sin^2(\theta_k)\big) = 2k\mathcal{L}_{\mathrm{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k})\end{aligned}$$

and

$$\begin{aligned}\big\|\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^{\top}\big)\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\big\|_F^2 &= \operatorname{trace}\big(\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^{\top}\big)\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^{\top}\big)\big)\\ &= k - \operatorname{trace}\big(\boldsymbol{B}^{\top}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\boldsymbol{B}\big)\\ &= k\mathcal{L}_{\mathrm{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}).\end{aligned}$$

The above equalities yield the first two equalities in (1.9).

Notice that both $U_1,\ldots,U_k$ and $W_1,\ldots W_k$ are both orthonormal bases of $\operatorname{span}(U_1,\ldots,U_k)$. (Similarly, $\widehat{U}_1,\ldots,\widehat{U}_k$ and $\widehat{W}_1,\ldots\widehat{W}_k$ are both orthonormal bases of $\operatorname{span}(\widehat{U}_1,\ldots,\widehat{U}_k)$).) Then we have $\boldsymbol{u}^{\top} = \boldsymbol{w}^{\top}\boldsymbol{R}$ where $\boldsymbol{R}$ is a $k\times k$ orthogonal matrix. Then

$$\begin{aligned}\min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{u}^{\top} - \hat{\boldsymbol{u}}^{\top}\boldsymbol{Q}\big\|_2^2 &= \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{u}^{\top} - \hat{\boldsymbol{w}}^{\top}\boldsymbol{Q}\big\|_2^2 = \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{w}^{\top}\boldsymbol{R} - \hat{\boldsymbol{w}}^{\top}\boldsymbol{Q}\big\|_2^2\\ &= \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{w}^{\top} - \hat{\boldsymbol{w}}^{\top}\boldsymbol{Q}\boldsymbol{R}^{\top}\big\|_2^2 = \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{w}^{\top} - \hat{\boldsymbol{w}}^{\top}\boldsymbol{Q}\big\|_2^2\\ &= \min_{\boldsymbol{q}_i\in\mathbb{R}^k, i=1,\ldots,k} \mathbb{E}\sum_{i=1}^{k}\big(W_i - \hat{\boldsymbol{w}}^{\top}\boldsymbol{q}_i\big)^2\\ &= \min_{\boldsymbol{q}_i\in\mathbb{R}^k, i=1,\ldots,k} \sum_{i=1}^{k}\mathbb{E}\big(W_i - \hat{\boldsymbol{w}}^{\top}\boldsymbol{q}_i\big)^2\\ &= \sum_{i=1}^{k}\min_{\boldsymbol{q}_i\in\mathbb{R}^k} \mathbb{E}\big(W_i - \hat{\boldsymbol{w}}^{\top}\boldsymbol{q}_i\big)^2\end{aligned}$$

Notice that $\min_{\boldsymbol{q}_i\in\mathbb{R}^k}\mathbb{E}(W_i - \hat{\boldsymbol{w}}^{\top}\boldsymbol{q}_i)^2$ is obtained by the best linear predictor, so

$$\begin{aligned}\min_{\boldsymbol{q}_i\in\mathbb{R}^k}\mathbb{E}\big(W_i - \hat{\boldsymbol{w}}^{\top}\boldsymbol{q}_i\big)^2 &= \operatorname{Var}(W_i) - \operatorname{Cov}(\hat{\boldsymbol{w}}, W_i)^{\top}\operatorname{Cov}^{-1}(\hat{\boldsymbol{w}})\operatorname{Cov}(\hat{\boldsymbol{w}}, W_i)\\ &= 1 - \cos^2\theta_i = \sin^2\theta_i.\end{aligned}$$

Therefore,

$$\min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big\|\boldsymbol{u}^\top - \hat{\boldsymbol{u}}^\top \boldsymbol{Q}\big\|_2^2 = \sum_{i=1}^{k} \sin^2\theta_i = k\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}),$$

which implies the third equality in (1.9). Similarly,

$$\begin{aligned}
&\max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big(\big(\boldsymbol{u}^\top - \hat{\boldsymbol{u}}^\top \boldsymbol{Q}\big)\boldsymbol{g}\big)^2 \\
&\quad= \max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big(\big(\boldsymbol{u}^\top - \hat{\boldsymbol{w}}^\top \boldsymbol{Q}\big)\boldsymbol{g}\big)^2 \\
&\quad= \max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big(\big(\boldsymbol{w}^\top \boldsymbol{R} - \hat{\boldsymbol{w}}^\top \boldsymbol{Q}\big)\boldsymbol{R}^\top \boldsymbol{g}\big)^2 \\
&\quad= \max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \min_{\boldsymbol{Q}\in\mathbb{R}^{k\times k}} \mathbb{E}\big(\big(\boldsymbol{w}^\top - \hat{\boldsymbol{w}}^\top \boldsymbol{Q}\big)\boldsymbol{g}\big)^2 \\
&\quad= \max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \min_{\boldsymbol{q}_i\in\mathbb{R}^k, i=1,\dots,k} \mathbb{E}\sum_{i=1}^{k} g_i^2\big(W_i - \hat{\boldsymbol{w}}^\top \boldsymbol{q}_i\big)^2 \\
&\quad= \max_{\boldsymbol{g}\in\mathbb{R}^k,\|\boldsymbol{g}\|=1} \sum_{i=1}^{k} g_i^2 \sin^2\theta_i \\
&\quad= \sin^2\theta_1
\end{aligned}$$

Finally, we prove (1.10). By [29], we have

$$\begin{aligned}
\big\|\boldsymbol{B}\boldsymbol{B}^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\big\|^2 &= \big\|\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^\top\big)\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\big\|^2 = \big\|\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^\top\big)\widehat{\boldsymbol{B}}\big\|^2 \\
&= \lambda_{\max}\big(\widehat{\boldsymbol{B}}^\top\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^\top\big)^\top\big(\boldsymbol{I}_{p_1} - \boldsymbol{B}\boldsymbol{B}^\top\big)\widehat{\boldsymbol{B}}\big) \\
&= \lambda_{\max}\big(\boldsymbol{I}_k - \operatorname{Diag}\big(\cos^2(\theta_1),\dots,\cos^2(\theta_k)\big)\big) \\
&= 1 - \cos^2(\theta_1) = \sin^2(\theta_1) = \mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}),
\end{aligned}$$

which implies the the equalities in (1.10).

7. Upper bound: Proof of Theorem 2.1

Throughout this proof, we denote $\Delta := \lambda_k - \lambda_{k+1}$. Also recall that the two samples of $\boldsymbol{x}$ and $\boldsymbol{y}$ are $\boldsymbol{X} = [X_{ij}] \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{Y} = [Y_{ij}] \in \mathbb{R}^{n\times p_2}$, respectively.

7.1. Linear invariance

Without loss of generality, we assume $p_2 \geq p_1 := p$. By the definition of canonical variables, we know that $U_1, \ldots, U_p$ and $V_1, \ldots, V_p$ are only determined by $\text{span}(X_1, \ldots, X_{p_1})$ and $\text{span}(Y_1, \ldots, Y_{p_2})$. In other words, for any invertible $\boldsymbol{C}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{C}_2 \in \mathbb{R}^{p_2 \times p_2}$, the canonical pairs of $(X_1, \ldots, X_{p_1})\boldsymbol{C_1}$ and $(Y_1, \ldots, Y_{p_2})\boldsymbol{C_2}$ are still $(U_1, V_1), \ldots, (U_{p_1}, V_{p_1})$. Therefore, we can consider the following orthonormal bases

$$U_1, \ldots, U_{p_1} \in \text{span}(X_1, \ldots, X_{p_1})$$

and

$$V_1, \ldots, V_{p_1}, V_{p_1+1}, \ldots, V_{p_2} \in \text{span}(Y_1, \ldots, Y_{p_2}).$$

Here $(V_1, \ldots, V_{p_1}, V_{p_1+1}, \ldots, V_{p_2})$ is an orthonormal extension of $V_1, \ldots, V_{p_1}$. Therefore, we know that $(U_1, V_1), \ldots, (U_{p_1}, V_{p_1})$ are also the the canonical pairs between $U_1, \ldots, U_{p_1}$ and $V_1, \ldots, V_{p_2}$.

Similarly, for a fixed sample of the variables of $\boldsymbol{x}$ and $\boldsymbol{y}$, the sample canonical pairs $(\widehat{U}_1, \widehat{V}_1), \ldots, (\widehat{U}_{p_1}, \widehat{V}_{p_1})$ are also sample canonical pairs of the corresponding sample of $(X_1, \ldots, X_{p_1})\boldsymbol{C_1}$ and $(Y_1, \ldots, Y_{p_2})\boldsymbol{C_2}$. This can be easily seen from the concept of sample canonical variables. For example, $\widehat{U}_1$ and $\widehat{V}_1$ are respectively, the linear combinations of $X_1, \ldots, X_{p_1}$ and $Y_1, \ldots, Y_{p_1}$, such that their corresponding sample variance are both 1 and sample correlation is maximized. If we replace $(X_1, \ldots, X_{p_1})$ and $(Y_1, \ldots, Y_{p_1})$ with $(X_1, \ldots, X_{p_1})\boldsymbol{C_1}$ and $(Y_1, \ldots, Y_{p_2})\boldsymbol{C_2}$, respectively and seek for the first sample canonical pair, the constraints (linear combinations of the two sets of variables and unit sample variances) and the objective (sample correlation is maximized) are the same as before, so $(\widehat{U}_1, \widehat{V}_1)$ is still the answer. Similarly, $(\widehat{U}_1, \widehat{V}_1), \ldots, (\widehat{U}_{p_1}, \widehat{V}_{p_1})$ are the sample canonical pairs of $(X_1, \ldots, X_{p_1})\boldsymbol{C_1}$ and $(Y_1, \ldots, Y_{p_2})\boldsymbol{C_2}$. In particular, they are the sample canonical pairs of $U_1, \ldots, U_{p_1}$ and $V_1, \ldots, V_{p_2}$.

The above argument gives the following convenient fact: In order to bound

$$\mathcal{L}_{\text{ave/max}}\big(\text{span}(\widehat{U}_1, \ldots, \widehat{U}_k), \text{span}(U_1, \ldots, U_k)\big)$$

we can replace $X_1, \ldots, X_{p_1}, Y_1, \ldots, Y_{p_2}$ with $U_1, \ldots, U_{p_1}, V_1, \ldots, V_{p_2}$. In other words, we can assume $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy the standard form

$$\boldsymbol{\Sigma}_x = \boldsymbol{I}_{p_1}, \qquad \boldsymbol{\Sigma}_y = \boldsymbol{I}_{p_2}, \qquad \boldsymbol{\Sigma}_{xy} = [\boldsymbol{\Lambda}, \boldsymbol{0}_{p_1 \times (p_2 - p_1)}] := \widetilde{\boldsymbol{\Lambda}}$$

where $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p_1}) \in \mathbb{R}^{p_1 \times p_1}$. Moreover,

$$\boldsymbol{\Phi}_{1:p_1} = \boldsymbol{I}_{p_1}, \qquad \boldsymbol{\Psi}_{1:p_1} = \begin{bmatrix} \boldsymbol{I}_{p_1} \\ \boldsymbol{0}_{(p_2 - p_1) \times p_1} \end{bmatrix},$$

which implies that

$$\boldsymbol{\Phi}_{1:k} = \begin{bmatrix} \boldsymbol{I}_k \\ \boldsymbol{0}_{(p_1 - k) \times k} \end{bmatrix}, \qquad \boldsymbol{\Psi}_{1:k} = \begin{bmatrix} \boldsymbol{I}_k \\ \boldsymbol{0}_{(p_2 - k) \times k} \end{bmatrix}.$$

7.2. Upper bound under the standard form

Under the standard form, by (1.9) and (1.10), we have

$$\mathcal{L}_{\text{ave}}\big(\text{span}(\widehat{U}_1,\ldots,\widehat{U}_k),\text{span}(U_1,\ldots,U_k)\big)=\frac{1}{k}\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_{\text{F}}^2 \tag{7.1}$$

and

$$\mathcal{L}_{\max}\big(\text{span}(\widehat{U}_1,\ldots,\widehat{U}_k),\text{span}(U_1,\ldots,U_k)\big)=\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2. \tag{7.2}$$

Denote $\widehat{\boldsymbol{\Phi}}_{1:k}=\left[\begin{smallmatrix}\widehat{\boldsymbol{\Phi}}^u_{1:k}\\ \widehat{\boldsymbol{\Phi}}^l_{1:k}\end{smallmatrix}\right]$ where $\widehat{\boldsymbol{\Phi}}^u_{1:k}$ and $\widehat{\boldsymbol{\Phi}}^l_{1:k}$ are the upper $k\times k$ and lower $(p_1-k)\times k$ submatrices of $\widehat{\boldsymbol{\Phi}}_{1:k}$, respectively. Then

$$\begin{aligned}\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_{\text{F}}^2&=\text{trace}\big((\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\widehat{\boldsymbol{\Phi}}_{1:k}\big(\widehat{\boldsymbol{\Phi}}_{1:k}^\top\widehat{\boldsymbol{\Phi}}_{1:k}\big)^{-1}\widehat{\boldsymbol{\Phi}}_{1:k}^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\big),\\ \big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2&=\lambda_{\max}\big((\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\widehat{\boldsymbol{\Phi}}_{1:k}\big(\widehat{\boldsymbol{\Phi}}_{1:k}^\top\widehat{\boldsymbol{\Phi}}_{1:k}\big)^{-1}\widehat{\boldsymbol{\Phi}}_{1:k}^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\big)\end{aligned}$$

Since

$$\begin{aligned}&(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\widehat{\boldsymbol{\Phi}}_{1:k}\big(\widehat{\boldsymbol{\Phi}}_{1:k}^\top\widehat{\boldsymbol{\Phi}}_{1:k}\big)^{-1}\widehat{\boldsymbol{\Phi}}_{1:k}^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\\ &\quad\preceq\frac{1}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\widehat{\boldsymbol{\Phi}}_{1:k}\widehat{\boldsymbol{\Phi}}_{1:k}^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})=\frac{1}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}\begin{bmatrix}\boldsymbol{0}_{k\times k}\\ \widehat{\boldsymbol{\Phi}}^l_{1:k}\end{bmatrix}\begin{bmatrix}\boldsymbol{0}_{k\times k} & \big(\widehat{\boldsymbol{\Phi}}^l_{1:k}\big)^\top\end{bmatrix},\end{aligned}$$

we have

$$\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_F^2\le\text{trace}\left(\frac{1}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}\begin{bmatrix}\boldsymbol{0}_{k\times k}\\ \widehat{\boldsymbol{\Phi}}^l_{1:k}\end{bmatrix}\begin{bmatrix}\boldsymbol{0}_{k\times k} & \widehat{\boldsymbol{\Phi}}_{1:k}^\top\end{bmatrix}\right)=\frac{\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\|_F^2}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}, \tag{7.3}$$

and

$$\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2\le\lambda_{\max}\left(\frac{1}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}\begin{bmatrix}\boldsymbol{0}_{k\times k}\\ \widehat{\boldsymbol{\Phi}}^l_{1:k}\end{bmatrix}\begin{bmatrix}\boldsymbol{0}_{k\times k} & \widehat{\boldsymbol{\Phi}}_{1:k}^\top\end{bmatrix}\right)=\frac{\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\|^2}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}. \tag{7.4}$$

Therefore, it suffices to give upper bounds of $\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\|_F^2$ and $\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\|^2$, as well as a lower bound of $\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})$.

7.3. Basic bounds

Recall that

$$\boldsymbol{\Sigma}_x=\boldsymbol{I}_{p_1},\qquad\boldsymbol{\Sigma}_y=\boldsymbol{I}_{p_2},\qquad\boldsymbol{\Sigma}_{xy}=[\boldsymbol{\Lambda},\boldsymbol{0}_{p_1\times(p_2-p_1)}]:=\widetilde{\boldsymbol{\Lambda}}.$$

Then

$$\operatorname{Cov}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right) := \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{I}_{p_1} & \widetilde{\boldsymbol{\Lambda}}\\ \widetilde{\boldsymbol{\Lambda}}^{\top} & \boldsymbol{I}_{p_2}\end{bmatrix}$$

and

$$\widehat{\operatorname{Cov}}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right) := \widehat{\boldsymbol{\Sigma}} = \begin{bmatrix}\widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xy}\\ \widehat{\boldsymbol{\Sigma}}_{yx} & \widehat{\boldsymbol{\Sigma}}_y\end{bmatrix}.$$

Moreover, we can define $\widehat{\boldsymbol{\Sigma}}_{2p_1}$ as the left upper $(2p_1)\times(2p_1)$ principal submatrix of $\widehat{\boldsymbol{\Sigma}}$. We can similarly define $\boldsymbol{\Sigma}_{2p_1}$.

Lemma 7.1. *There exist universal constants γ, C and C_0 such that when $n \geq C_0 p_1$, then with probability at least $1 - e^{-\gamma p_1}$, the following inequalities hold*

$$\|\boldsymbol{\Sigma}_{2p_1} - \widehat{\boldsymbol{\Sigma}}_{2p_1}\|, \|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x\|, \|\widehat{\boldsymbol{\Sigma}}_x^{1/2} - \boldsymbol{I}_{p_1}\| \leq C\sqrt{\frac{p_1}{n}}.$$

Proof. It is obvious that $\|\boldsymbol{\Sigma}_{2p_1}\| \leq 2$. By Lemma 7.9, there exist constants γ, C_0 and C_1, such that when $n \geq C_0 p_1$, with probability at least $1 - e^{-\gamma p_1}$ there holds

$$\|\widehat{\boldsymbol{\Sigma}}_{2p_1} - \boldsymbol{\Sigma}_{2p_1}\| \leq C_1\sqrt{\frac{p_1}{n}}.$$

As submatrices, we have $\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x\| \leq C_1\sqrt{\frac{p_1}{n}}$. Moreover,

$$\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x\| = \left\|\left(\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right)\left(\boldsymbol{I}_{p_1} + \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right)\right\| \geq \sigma_{\min}\left(\boldsymbol{I}_{p_1} + \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right)\left\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right\| \geq \left\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right\|,$$

which implies $\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\| \leq C_1\sqrt{\frac{p_1+p_2}{n}}$. □

Lemma 7.2. *There exist universal constants c, C and C_0 such that when $n \geq C_0(p_1+p_2)$, then with probability at least $1 - e^{-c(p_1+p_2)}$, the following inequalities hold*

$$\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|, \|\boldsymbol{I}_{p_2} - \widehat{\boldsymbol{\Sigma}}_y\|, \|\boldsymbol{\Sigma}_{xy} - \widehat{\boldsymbol{\Sigma}}_{xy}\|, \left\|\widehat{\boldsymbol{\Sigma}}_y^{1/2} - \boldsymbol{I}_{p_2}\right\| \leq C\sqrt{\frac{p_1+p_2}{n}},$$

$$\|\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\| \leq \left\|\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2} - \boldsymbol{\Sigma}_{xy}\right\| \leq C\sqrt{\frac{p_1+p_2}{n}},$$

$$\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k}) \geq \frac{1}{2}, \qquad \|\widehat{\boldsymbol{\Phi}}_{1:k}\|^2 \leq \frac{3}{2}, \qquad \sigma_k^2(\widehat{\boldsymbol{\Psi}}_{1:k}) \geq \frac{1}{2}, \qquad \|\widehat{\boldsymbol{\Psi}}_{1:k}\|^2 \leq \frac{3}{2},$$

$$\left\|\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\right\|, \left\|\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\right\| \leq \frac{C}{\Delta}\sqrt{\frac{p_1+p_2}{n}},$$

where $\Delta = \lambda_k - \lambda_{k+1}$ is the eigen-gap.

The proof is deferred to Section 7.7.

7.4. Estimating equations and upper bound of $\|\widehat{\mathbf{\Phi}}^{l}_{1:k}\|^2$

In this section, we aim to give a sharp upper bound for $\|\widehat{\mathbf{\Phi}}^{l}_{1:k}\|^2$. Notice that we have already established an upper bound in Lemma 7.2, where Wedin's $\sin\theta$ law plays the essential role. However, this bound is actually too loose for our purpose. Therefore, we need to develop new techniques to sharpen the results.

Recall that $\widehat{\mathbf{\Phi}} \in \mathbb{R}^{p_1 \times p_1}$, $\widehat{\mathbf{\Psi}} \in \mathbb{R}^{p_2 \times p_1}$ consist of the sample canonical coefficients. By definition, the sample canonical coefficients satisfy the following two estimating equations (because $\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}$ and $\widehat{\mathbf{\Sigma}}_y^{1/2}\widehat{\mathbf{\Psi}}$ are left and right singular vectors of $\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2}$, respectively),

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Psi}} &= \widehat{\mathbf{\Sigma}}_x\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Lambda}} \\ \widehat{\mathbf{\Sigma}}_{yx}\widehat{\mathbf{\Phi}} &= \widehat{\mathbf{\Sigma}}_y\widehat{\mathbf{\Psi}}\widehat{\mathbf{\Lambda}}. \end{aligned} \tag{7.5}$$

If we define

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & \\ & \mathbf{\Lambda}_2 \end{bmatrix} \in \mathbb{R}^{p_1 \times p_1}, \qquad \widehat{\mathbf{\Lambda}} = \begin{bmatrix} \widehat{\mathbf{\Lambda}}_1 & \\ & \widehat{\mathbf{\Lambda}}_2 \end{bmatrix} \in \mathbb{R}^{p_1 \times p_1}, \tag{7.6}$$

where $\mathbf{\Lambda}_1$, $\widehat{\mathbf{\Lambda}}_1$ are $k \times k$ diagonal matrices while $\mathbf{\Lambda}_2$, $\widehat{\mathbf{\Lambda}}_2$ are $(p_1 - k) \times (p_1 - k)$ diagonal matrices. Then (7.5) imply

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Psi}}_{1:k} &= \widehat{\mathbf{\Sigma}}_x\widehat{\mathbf{\Phi}}_{1:k}\widehat{\mathbf{\Lambda}}_1 \\ \widehat{\mathbf{\Sigma}}_{yx}\widehat{\mathbf{\Phi}}_{1:k} &= \widehat{\mathbf{\Sigma}}_y\widehat{\mathbf{\Psi}}_{1:k}\widehat{\mathbf{\Lambda}}_1. \end{aligned} \tag{7.7}$$

Divide the matrices into blocks,

$$\widehat{\mathbf{\Sigma}}_x = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_x^{11} & \widehat{\mathbf{\Sigma}}_x^{12} \\ \widehat{\mathbf{\Sigma}}_x^{21} & \widehat{\mathbf{\Sigma}}_x^{22} \end{bmatrix}, \qquad \widehat{\mathbf{\Sigma}}_y = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_y^{11} & \widehat{\mathbf{\Sigma}}_y^{12} \\ \widehat{\mathbf{\Sigma}}_y^{21} & \widehat{\mathbf{\Sigma}}_y^{22} \end{bmatrix},$$

$$\widehat{\mathbf{\Sigma}}_{xy} = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_{xy}^{11} & \widehat{\mathbf{\Sigma}}_{xy}^{12} \\ \widehat{\mathbf{\Sigma}}_{xy}^{21} & \widehat{\mathbf{\Sigma}}_{xy}^{22} \end{bmatrix}, \qquad \widehat{\mathbf{\Sigma}}_{yx} = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_{yx}^{11} & \widehat{\mathbf{\Sigma}}_{yx}^{12} \\ \widehat{\mathbf{\Sigma}}_{yx}^{21} & \widehat{\mathbf{\Sigma}}_{yx}^{22} \end{bmatrix}$$

where $\widehat{\mathbf{\Sigma}}_x^{11}$, $\widehat{\mathbf{\Sigma}}_y^{11}$, $\widehat{\mathbf{\Sigma}}_{xy}^{11}$, $\widehat{\mathbf{\Sigma}}_{yx}^{11}$ are $k \times k$ matrices. Finally, we define $\widehat{\mathbf{\Psi}}^{u}_{1:k} \in \mathbb{R}^{k \times k}$, $\widehat{\mathbf{\Psi}}^{l}_{1:k} \in \mathbb{R}^{(p_2-k) \times k}$ in the same way as $\widehat{\mathbf{\Phi}}^{u}_{1:k}$, $\widehat{\mathbf{\Phi}}^{l}_{1:k}$. With these blocks, (7.7) can be rewritten as

$$\widehat{\mathbf{\Sigma}}_{xy}^{21}\widehat{\mathbf{\Psi}}^{u}_{1:k} + \widehat{\mathbf{\Sigma}}_{xy}^{22}\widehat{\mathbf{\Psi}}^{l}_{1:k} = \widehat{\mathbf{\Sigma}}_x^{21}\widehat{\mathbf{\Phi}}^{u}_{1:k}\widehat{\mathbf{\Lambda}}_1 + \widehat{\mathbf{\Sigma}}_x^{22}\widehat{\mathbf{\Phi}}^{l}_{1:k}\widehat{\mathbf{\Lambda}}_1, \tag{7.8}$$

$$\widehat{\mathbf{\Sigma}}_{yx}^{21}\widehat{\mathbf{\Phi}}^{u}_{1:k} + \widehat{\mathbf{\Sigma}}_{yx}^{22}\widehat{\mathbf{\Phi}}^{l}_{1:k} = \widehat{\mathbf{\Sigma}}_y^{21}\widehat{\mathbf{\Psi}}^{u}_{1:k}\widehat{\mathbf{\Lambda}}_1 + \widehat{\mathbf{\Sigma}}_y^{22}\widehat{\mathbf{\Psi}}^{l}_{1:k}\widehat{\mathbf{\Lambda}}_1, \tag{7.9}$$

$$\widehat{\mathbf{\Sigma}}_{xy}^{11}\widehat{\mathbf{\Psi}}^{u}_{1:k} + \widehat{\mathbf{\Sigma}}_{xy}^{12}\widehat{\mathbf{\Psi}}^{l}_{1:k} = \widehat{\mathbf{\Sigma}}_x^{11}\widehat{\mathbf{\Phi}}^{u}_{1:k}\widehat{\mathbf{\Lambda}}_1 + \widehat{\mathbf{\Sigma}}_x^{12}\widehat{\mathbf{\Phi}}^{l}_{1:k}\widehat{\mathbf{\Lambda}}_1, \tag{7.10}$$

$$\widehat{\boldsymbol{\Sigma}}_{yx}^{11}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\widehat{\boldsymbol{\Sigma}}_{yx}^{12}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}=\widehat{\boldsymbol{\Sigma}}_{y}^{11}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\widehat{\boldsymbol{\Lambda}}_{1}+\widehat{\boldsymbol{\Sigma}}_{y}^{12}\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}. \tag{7.11}$$

Define the zero-padding of $\boldsymbol{\Lambda}_2$:

$$\widetilde{\boldsymbol{\Lambda}}_{2}:=[\boldsymbol{\Lambda}_{2},\mathbf{0}]=\boldsymbol{\Sigma}_{xy}^{22}\in\mathbb{R}^{(p_1-k)\times(p_2-k)}.$$

The above equations imply the following lemma.

Lemma 7.3. *The equality* (7.7) *gives the following result*

$$\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_{1}^{2}-\boldsymbol{\Lambda}_{2}^{2}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}=\boldsymbol{B}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\boldsymbol{R} \tag{7.12}$$

$$=\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_{1}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_{1}+\widetilde{\boldsymbol{\Lambda}}_{2}\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{21}-\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{\Lambda}_{1}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\widetilde{\boldsymbol{R}} \tag{7.13}$$

where

$$\begin{aligned}
\boldsymbol{B}&:=\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\boldsymbol{\Lambda}_{1}+\widetilde{\boldsymbol{\Lambda}}_{2}\widehat{\boldsymbol{\Sigma}}_{yx}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_{1}^{2}-\widetilde{\boldsymbol{\Lambda}}_{2}\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{\Lambda}_{1},\\
\widetilde{\boldsymbol{R}}&:=\big(\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{R}_{1}-\boldsymbol{R}_{3}\big)\boldsymbol{\Lambda}_{1}-\widetilde{\boldsymbol{\Lambda}}_{2}\big(\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{R}_{2}+\boldsymbol{R}_{4}\big),\\
\boldsymbol{R}&:=\widetilde{\boldsymbol{R}}-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_{1}\big)\boldsymbol{R}_{2}.
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{R}_{1}&:=\widehat{\boldsymbol{\Phi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})+\big(\widehat{\boldsymbol{\Sigma}}_{x}^{11}-\boldsymbol{I}_{k}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\widehat{\boldsymbol{\Lambda}}_{1}+\widehat{\boldsymbol{\Sigma}}_{x}^{12}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{11}-\boldsymbol{\Lambda}_{1}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{u}-\widehat{\boldsymbol{\Sigma}}_{xy}^{12}\widehat{\boldsymbol{\Psi}}_{1:k}^{l},\\
\boldsymbol{R}_{2}&:=\widehat{\boldsymbol{\Psi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})+\big(\widehat{\boldsymbol{\Sigma}}_{y}^{11}-\boldsymbol{I}_{k}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\widehat{\boldsymbol{\Lambda}}_{1}+\widehat{\boldsymbol{\Sigma}}_{y}^{12}\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}-\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{11}-\boldsymbol{\Lambda}_{1}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{u}-\widehat{\boldsymbol{\Sigma}}_{yx}^{12}\widehat{\boldsymbol{\Phi}}_{1:k}^{l},\\
\boldsymbol{R}_{3}&:=\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})+\big(\widehat{\boldsymbol{\Sigma}}_{x}^{22}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}-\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_{1}\big)-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{22}-\widetilde{\boldsymbol{\Lambda}}_{2}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{l},\\
\boldsymbol{R}_{4}&:=\widehat{\boldsymbol{\Sigma}}_{y}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})+\big(\widehat{\boldsymbol{\Sigma}}_{y}^{22}\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}-\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\boldsymbol{\Lambda}_{1}\big)-\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{22}-\widetilde{\boldsymbol{\Lambda}}_{2}^{\top}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{l}.
\end{aligned}$$

The proof is deferred to Section 7.7.

By Lemma 7.2, one can easily obtain that

$$\|\boldsymbol{R}_{1}\|,\|\boldsymbol{R}_{2}\|\leq C\sqrt{\frac{p_1+p_2}{n}}.$$

Recall that

$$\boldsymbol{R}_{3}:=\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})+\big(\widehat{\boldsymbol{\Sigma}}_{x}^{22}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_{1}-\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_{1}\big)-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{22}-\widetilde{\boldsymbol{\Lambda}}_{2}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{l}$$

By Lemma 7.2, we have

$$\big\|\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_{1}-\boldsymbol{\Lambda}_{1})\big\|\leq C\frac{p_1+p_2}{n},\qquad\big\|\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{22}-\widetilde{\boldsymbol{\Lambda}}_{2}\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\big\|\leq C\frac{p_1+p_2}{\Delta n},$$

and

$$\begin{aligned}\big\|\widehat{\boldsymbol{\Sigma}}_x^{22}\widehat{\boldsymbol{\Phi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1-\widehat{\boldsymbol{\Phi}}_{1:k}^l\boldsymbol{\Lambda}_1\big\| &\le \big\|\big(\widehat{\boldsymbol{\Sigma}}_x^{22}-\boldsymbol{I}_{p_1-k}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1+\widehat{\boldsymbol{\Phi}}_{1:k}^l(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1)\big\|\\ &\le \big\|\big(\widehat{\boldsymbol{\Sigma}}_x^{22}-\boldsymbol{I}_{p_1-k}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1\big\|+\big\|\widehat{\boldsymbol{\Phi}}_{1:k}^l(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1)\big\|\le C\frac{p_1+p_2}{\Delta n}.\end{aligned}$$

Therefore, we get $\|\boldsymbol{R}_3\|\le C\frac{p_1+p_2}{\Delta n}$. Similarly, $\|\boldsymbol{R}_4\|\le C\frac{p_1+p_2}{\Delta n}$.

Combined with Lemma 7.2, we have

$$\|\widetilde{\boldsymbol{R}}\|=\big\|\big(\widehat{\boldsymbol{\Sigma}}_x^{21}\boldsymbol{R}_1-\boldsymbol{R}_3\big)\boldsymbol{\Lambda}_1-\widetilde{\boldsymbol{\Lambda}}_2\big(\widehat{\boldsymbol{\Sigma}}_y^{21}\boldsymbol{R}_2+\boldsymbol{R}_4\big)\big\|\le C\frac{p_1+p_2}{\Delta n}$$

and

$$\|\boldsymbol{R}\|\le\|\widetilde{\boldsymbol{R}}\|+\big\|\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_x^{21}\boldsymbol{\Lambda}_1\big\|\|\boldsymbol{R}_2\|\le C\frac{p_1+p_2}{\Delta n}.$$

The proof of the following lemma is deferred to Section 7.7:

Lemma 7.4. *If $n\ge C_0(p_1+p_2)$, then with probability $1-c_0\exp(-\gamma p_1)$,*

$$\big\|\widehat{\boldsymbol{\Phi}}_{1:k}^l\big\|\le C\bigg[\sqrt{\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2}}+\frac{(p_1+p_2)}{n\Delta^2}\bigg].$$

7.5. Upper bounds of risks

Notice that the inequality (7.4) yields

$$\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2\le\frac{\|\widehat{\boldsymbol{\Phi}}_{1:k}^l\|^2}{\sigma_k^2(\widehat{\boldsymbol{\Phi}}_{1:k})}.$$

By Lemma 7.4 and Lemma 7.2, we know on an event G with probability at least $1-Ce^{-\gamma p_1}$,

$$\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2\le C\bigg[\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2}+\frac{(p_1+p_2)^2}{n^2\Delta^4}\bigg].$$

Moreover, since $\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\|^2\le 1$, by (7.2), we have

$$\begin{aligned}\mathbb{E}\mathcal{L}_{\max}(\boldsymbol{\Phi}_{1:k},\widehat{\boldsymbol{\Phi}}_{1:k})&=\mathbb{E}\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2\\ &\le C\bigg[\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2}+\frac{(p_1+p_2)^2}{n^2\Delta^4}+e^{-\gamma p_1}\bigg].\end{aligned}$$

Since $(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}$ is of at most rank-k, we have

$$\frac{1}{k}\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_F^2\le\big\|(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2$$

Then by (7.1) and the previous inequality, we have

$$\begin{aligned}\mathbb{E}\mathcal{L}_{\text{ave}}(\boldsymbol{\Phi}_{1:k}, \widehat{\boldsymbol{\Phi}}_{1:k}) &= \mathbb{E}\big\|(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2 \\ &= \mathbb{E}\frac{1}{k}\big\|(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|_{\text{F}}^2 \\ &\le \mathbb{E}\big\|(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{\Phi}}_{1:k}}\big\|^2 \\ &\le C\left[\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2} + \frac{(p_1+p_2)^2}{n^2\Delta^4} + e^{-\gamma p_1}\right].\end{aligned}$$

In fact, the factor p_1 in the main term can be reduced to $p_1 - k$ by similar arguments as done for the operator norm. The Frobenius norm version of Lemma 7.4 is actually much simpler. We omit the proof to avoid unnecessary redundancy and repetition.

7.6. Supporting lemmas in linear algebra and probability

Definition 7.5 (Hadamard Operator Norm). For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, define the Hadamard operator norm as

$$|||\boldsymbol{A}||| = \sup\big\{\|\boldsymbol{A}\circ\boldsymbol{B}\| : \|\boldsymbol{B}\| \le 1, \boldsymbol{B} \in \mathbb{R}^{m\times n}\big\}$$

Lemma 7.6. *Let $\{\alpha_i\}_{i=1}^m$ and $\{\beta_i\}_{i=1}^n$ be two sequences of positive numbers. for any $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, there hold*

$$\left\|\left[\frac{\sqrt{\alpha_i\beta_j}}{\alpha_i+\beta_j}\right]\circ\boldsymbol{X}\right\| \le \frac{1}{2}\|\boldsymbol{X}\|, \tag{7.14}$$

and

$$\left\|\left[\frac{\min(\alpha_i,\beta_j)}{\alpha_i+\beta_j}\right]\circ\boldsymbol{X}\right\| \le \frac{1}{2}\|\boldsymbol{X}\|, \qquad \left\|\left[\frac{\max(\alpha_i,\beta_j)}{\alpha_i+\beta_j}\right]\circ\boldsymbol{X}\right\| \le \frac{3}{2}\|\boldsymbol{X}\|. \tag{7.15}$$

Proof. The proof of (7.14) can be found in "Norm Bounds for Hadamard Products and an Arithmetic-Geometric Mean Inequality for Unitarily Invariant Norms" by Horn.

Denote

$$\boldsymbol{G}_1 = \left[\frac{\max(\alpha_i,\beta_j)}{\alpha_i+\beta_j}\right], \qquad \boldsymbol{G}_2 = \left[\frac{\min(\alpha_i,\beta_j)}{\alpha_i+\beta_j}\right]$$

The proof of (7.15) relies on the following two results.

Lemma 7.7 (Theorem 5.5.18 of [16]). *If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{A}$ is positive semidefinite. Then,*

$$\|\boldsymbol{A}\circ\boldsymbol{B}\| \le \Big(\max_{1\le i\le n}\boldsymbol{A}_{ii}\Big)\|\boldsymbol{B}\|,$$

where $\|\cdot\|$ is the operator norm.

Lemma 7.8 (Theorem 3.2 of [21]). *The symmetric matrix*

$$\left(\frac{\min(a_i, a_j)}{a_i + a_j}\right)_{1 \le i,j \le n}$$

is positive semidefinite if $a_i > 0$, $1 \le i \le n$.

Define $\gamma_i = \beta_i$, $1 \le i \le n$ and $\gamma_i = \alpha_{i-n}$, $n+1 \le i \le m+n$. Define $\boldsymbol{M} \in \mathbb{R}^{(m+n)\times(m+n)}$ by

$$M_{ij} = \frac{\min\{\gamma_i, \gamma_j\}}{\gamma_i + \gamma_j}.$$

By Lemma 7.8, $\boldsymbol{M}$ is also positive semidefinite. Again, apply Lemma 7.7 and notice that $\boldsymbol{G}_2$ is the lower left sub-matrix of $\boldsymbol{M}$, It is easy to obtain

$$|||\boldsymbol{G}_2||| \le |||\boldsymbol{M}||| \le \frac{1}{2}.$$

Finally, since $\boldsymbol{G}_1 \circ \boldsymbol{B} = \boldsymbol{B} - \boldsymbol{G}_2 \circ \boldsymbol{B}$ for any $\boldsymbol{B}$, we have

$$\|\boldsymbol{G}_1 \circ \boldsymbol{B}\| \le \|\boldsymbol{B}\| + \|\boldsymbol{G}_2 \circ \boldsymbol{B}\|,$$

which implies,

$$|||\boldsymbol{G}_1||| \le 1 + |||\boldsymbol{G}_2||| \le \frac{3}{2}. \qquad \square$$

Lemma 7.9 (Covariance Matrix Estimation, Remark 5.40 of [25]). *Assume* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *has independent sub-gaussian random rows with second moment matrix* $\boldsymbol{\Sigma}$. *Then there exists universal constant* C *such that for every* $t \ge 0$, *the following inequality holds with probability at least* $1 - e^{-ct^2}$,

$$\left\| \frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{\Sigma} \right\| \le \max\{\delta, \delta^2\} \|\boldsymbol{\Sigma}\| \qquad \delta = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}.$$

Lemma 7.10 (Bernstein inequality, Proposition 5.16 of [25]). *Let* $X_1, \ldots, X_n$ *be independent centered sub-exponential random variables and* $K = \max_i \|X_i\|_{\psi_1}$. *Then for every* $\boldsymbol{a} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) \in \mathbb{R}^n$ *and every* $t \ge 0$, *we have*

$$P\left\{ \left| \sum_{i=1}^n \boldsymbol{a}_i X_i \right| \ge t \right\} \le 2\exp\left\{ -c \min\left(\frac{t^2}{K^2 \|\boldsymbol{a}\|_2^2}, \frac{t}{K\|\boldsymbol{a}\|_\infty} \right) \right\}.$$

Lemma 7.11 (Hanson-Wright inequality, Theorem 1.1 of [23]). *Let* $\boldsymbol{x} = (x_1, \ldots, x_p)$ *be a random vector with independent components* x_i *which satisfy* $\mathbb{E}x_i = 0$ *and* $\|x_i\|_{\psi_2} \le K$, *Let* $\boldsymbol{A} \in \mathbb{R}^{p\times p}$. *Then there exists universal constant* c *such that for every* $t \ge 0$,

$$P\left\{ \left| \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} - \mathbb{E}\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \right| \ge t \right\} \le 2\exp\left\{ -c \min\left(\frac{t^2}{K^4 \|\boldsymbol{A}\|_F^2}, \frac{t}{K^2 \|\boldsymbol{A}\|} \right) \right\}.$$

Lemma 7.12 (Covering Number of the Sphere, Lemma 5.2 of [25]). *The unit Euclidean sphere $\mathbb{S}^{n-1}$ equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that*

$$\left|\mathcal{N}\left(\mathbb{S}^{n-1}, \epsilon\right)\right| \leq \left(1 + \frac{2}{\epsilon}\right)^n,$$

where $\mathcal{N}(\mathbb{S}^{n-1}, \epsilon)$ is the ϵ-net of $\mathbb{S}^{n-1}$ with minimal cardinality.

The following variant of Wedin's $\sin\theta$ law [28] is proved in Proposition 1 of [3].

Lemma 7.13. *For $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{m\times n}$ and $\widehat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$, define the singular value decompositions of $\boldsymbol{A}$ and $\widehat{\boldsymbol{A}}$ as*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \qquad \widehat{\boldsymbol{A}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{V}}^\top.$$

Then the following perturbation bound holds,

$$\left\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{U}_{1:k}})\boldsymbol{P}_{\widehat{\boldsymbol{U}}_{1:k}}\right\| = \|\boldsymbol{P}_{\boldsymbol{U}_{1:k}} - \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{1:k}}\| \leq \frac{2\|\boldsymbol{E}\|}{\sigma_k(\boldsymbol{A}) - \sigma_{k+1}(\boldsymbol{A})},$$

where $\sigma_k(\boldsymbol{A})$, $\sigma_{k+1}(\boldsymbol{A})$ are the kth and $(k+1)$th singular values of $\boldsymbol{A}$.

7.7. Proofs of key lemmas

7.7.1. *Proof of Lemma 7.2*

(1) The proof of

$$\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|, \|\boldsymbol{I}_{p_2} - \widehat{\boldsymbol{\Sigma}}_y\|, \|\boldsymbol{\Sigma}_{xy} - \widehat{\boldsymbol{\Sigma}}_{xy}\|, \left\|\widehat{\boldsymbol{\Sigma}}_y^{1/2} - \boldsymbol{I}_{p_2}\right\| \leq C\sqrt{\frac{p_1 + p_2}{n}}$$

is exactly the same as that of Lemma 7.1.

(2) Observe that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2} - \boldsymbol{\Sigma}_{xy} &= \left(\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right)\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2} \\ &\quad + \widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2}\left(\boldsymbol{I}_{p_2} - \widehat{\boldsymbol{\Sigma}}_y^{1/2}\right) + (\widehat{\boldsymbol{\Sigma}}_{xy} - \boldsymbol{\Sigma}_{xy}),\end{aligned}$$

and $\|\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2}\| = \widehat{\lambda}_1 \leq 1$. Then

$$\left\|\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2} - \boldsymbol{\Sigma}_{xy}\right\| \leq \left\|\boldsymbol{I}_{p_1} - \widehat{\boldsymbol{\Sigma}}_x^{1/2}\right\| + \|\widehat{\boldsymbol{\Sigma}}_x\|\left\|\boldsymbol{I}_{p_2} - \widehat{\boldsymbol{\Sigma}}_y^{1/2}\right\| + \|\widehat{\boldsymbol{\Sigma}}_{xy} - \boldsymbol{\Sigma}_{xy}\|.$$

Notice that $\widehat{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}$ are singular values of $\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2}$ and $\mathbf{\Sigma}_{xy}$ respectively. Hence by the famous Weyl's inequality for singular values,

$$\begin{aligned}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\| &\le \big\|\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2} - \mathbf{\Sigma}_{xy}\big\| \\ &\le \|\boldsymbol{I}_{p_1} - \widehat{\mathbf{\Sigma}}_x\| + \|\widehat{\mathbf{\Sigma}}_x\|\big\|\boldsymbol{I}_{p_2} - \widehat{\mathbf{\Sigma}}_y^{1/2}\big\| + \|\widehat{\mathbf{\Sigma}}_{xy} - \mathbf{\Sigma}_{xy}\| \\ &\le \left(3 + C_1\sqrt{\frac{p_1+p_2}{n}}\right)C_1\sqrt{\frac{p_1+p_2}{n}} \le C_2\sqrt{\frac{p_1+p_2}{n}}.\end{aligned}$$

(3) Since $\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}$ are left singular vectors of $\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2}$, we have $\|\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}\| = 1$, $\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Sigma}}_x\widehat{\mathbf{\Phi}} = \boldsymbol{I}_{p_1}$ and $\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Phi}} - \boldsymbol{I}_{p_1} = -\widehat{\mathbf{\Phi}}^\top(\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1})\widehat{\mathbf{\Phi}}$. Then we have,

$$\begin{aligned}\big\|\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Phi}} - \boldsymbol{I}_{p_1}\big\| &= \big\|\widehat{\mathbf{\Phi}}^\top(\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1})\widehat{\mathbf{\Phi}}\big\| \le \big\|\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Sigma}}_x^{1/2}\big\|\big\|\widehat{\mathbf{\Sigma}}_x^{-1/2}(\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1})\widehat{\mathbf{\Sigma}}_x^{-1/2}\big\|\big\|\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}\big\| \\ &= \big\|\widehat{\mathbf{\Sigma}}_x^{-1/2}(\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1})\widehat{\mathbf{\Sigma}}_x^{-1/2}\big\|.\end{aligned}$$

As a submatrix,

$$\begin{aligned}\big\|\widehat{\mathbf{\Phi}}_{1:k}^\top\widehat{\mathbf{\Phi}}_{1:k} - \boldsymbol{I}_k\big\| &\le \big\|\widehat{\mathbf{\Sigma}}_x^{-1/2}(\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1})\widehat{\mathbf{\Sigma}}_x^{-1/2}\big\| \le \big\|\widehat{\mathbf{\Sigma}}_x^{-1}\big\|\|\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1}\| \\ &\le \frac{1}{1 - \|\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1}\|}\|\widehat{\mathbf{\Sigma}}_x - \boldsymbol{I}_{p_1}\| \le \frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{1 - \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|} \le \frac{1}{2}\end{aligned}$$

as long as $n \ge C_0(p_1 + p_2)$ for sufficiently large C_0. In this case,

$$\sigma_k^2(\widehat{\mathbf{\Phi}}_{1:k}) \ge 1/2, \qquad \|\widehat{\mathbf{\Phi}}_{1:k}\|^2 \le 3/2.$$

By the same argument,

$$\sigma_k^2(\widehat{\mathbf{\Psi}}_{1:k}) \ge 1/2, \qquad \|\widehat{\mathbf{\Psi}}_{1:k}\|^2 \le 3/2.$$

(4) Recall that

$$\mathbf{\Phi}_{1:k} = \begin{bmatrix}\boldsymbol{I}_k \\ \mathbf{0}_{(p_1-k)\times k}\end{bmatrix}, \qquad \mathbf{\Psi}_{1:k} = \begin{bmatrix}\boldsymbol{I}_k \\ \mathbf{0}_{(p_2-k)\times k}\end{bmatrix}.$$

The last inequality in the lemma relies on the fact that $\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}_{1:k}$ and $\mathbf{\Phi}_{1:k}$ are leading k singular vectors of $\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2}$ and $\mathbf{\Sigma}_{xy}$ respectively. By a variant of Wedin's $\sin\theta$ law as stated in Lemma 7.13,

$$\big\|\boldsymbol{P}_{\widehat{\mathbf{\Sigma}}_x^{1/2}\widehat{\mathbf{\Phi}}_{1:k}}(\boldsymbol{I}_{p_1} - \boldsymbol{P}_{\mathbf{\Phi}_{1:k}})\big\| \le \frac{2\|\widehat{\mathbf{\Sigma}}_x^{-1/2}\widehat{\mathbf{\Sigma}}_{xy}\widehat{\mathbf{\Sigma}}_y^{-1/2} - \mathbf{\Sigma}_{xy}\|}{\Delta} \le \frac{2C_2}{\Delta}\sqrt{\frac{p_1+p_2}{n}}.$$

On the other hand,

$$\begin{aligned}\left\|\boldsymbol{P}_{\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}}(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\right\| &= \left\|\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big)^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\right\| \\ &= \left\|\big(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big)^\top(\boldsymbol{I}_{p_1}-\boldsymbol{P}_{\boldsymbol{\Phi}_{1:k}})\right\| \\ &= \left\|\big(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big)^l\right\|.\end{aligned}$$

Here the second equality is due to the fact that $\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}$ has orthonormal columns. Moreover, $(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k})^l$ denotes the lower $(p_1-k)\times k$ sub-matrix of $\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}$. Again, by triangle inequality,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Phi}}_{1:k}^l\right\| &= \left\|\big(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big)^l - \big((\widehat{\boldsymbol{\Sigma}}_x^{1/2}-\boldsymbol{I}_{p_1})\widehat{\boldsymbol{\Phi}}_{1:k}\big)^l\right\| \\ &\le \left\|\big(\widehat{\boldsymbol{\Sigma}}_x^{1/2}\widehat{\boldsymbol{\Phi}}_{1:k}\big)^l\right\| + \left\|(\widehat{\boldsymbol{\Sigma}}_x^{1/2}-\boldsymbol{I}_{p_1})\right\|\left\|\widehat{\boldsymbol{\Phi}}_{1:k}\right\| \\ &\le \frac{2C_2}{\Delta}\sqrt{\frac{p_1+p_2}{n}} + \sqrt{\frac{3}{2}}C_1\sqrt{\frac{p_1+p_2}{n}} \le \frac{C_3}{\Delta}\sqrt{\frac{p_1+p_2}{n}}.\end{aligned}$$

The last inequality is due to $\Delta \le 1$. Let $C = \max(C_1, C_2, C_3)$, the proof is done.

7.7.2. *Proof of Lemma 7.3*

The equality (7.10) implies

$$\begin{aligned}\boldsymbol{\Lambda}_1\widehat{\boldsymbol{\Psi}}_{1:k}^u - \widehat{\boldsymbol{\Phi}}_{1:k}^u\boldsymbol{\Lambda}_1 &= \widehat{\boldsymbol{\Phi}}_{1:k}^u(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1) + \big(\widehat{\boldsymbol{\Sigma}}_x^{11}-\boldsymbol{I}_k\big)\widehat{\boldsymbol{\Phi}}_{1:k}^u\widehat{\boldsymbol{\Lambda}}_1 + \widehat{\boldsymbol{\Sigma}}_x^{12}\widehat{\boldsymbol{\Phi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1 \\ &\quad - \big(\widehat{\boldsymbol{\Sigma}}_{xy}^{11}-\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Psi}}_{1:k}^u - \widehat{\boldsymbol{\Sigma}}_{xy}^{12}\widehat{\boldsymbol{\Psi}}_{1:k}^l \\ &:= \boldsymbol{R}_1. \end{aligned} \tag{7.16}$$

Similarly, (7.11) implies

$$\begin{aligned}\boldsymbol{\Lambda}_1\widehat{\boldsymbol{\Phi}}_{1:k}^u - \widehat{\boldsymbol{\Psi}}_{1:k}^u\boldsymbol{\Lambda}_1 &= \widehat{\boldsymbol{\Psi}}_{1:k}^u(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1) + \big(\widehat{\boldsymbol{\Sigma}}_y^{11}-\boldsymbol{I}_k\big)\widehat{\boldsymbol{\Psi}}_{1:k}^u\widehat{\boldsymbol{\Lambda}}_1 + \widehat{\boldsymbol{\Sigma}}_y^{12}\widehat{\boldsymbol{\Psi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1 \\ &\quad - \big(\widehat{\boldsymbol{\Sigma}}_{yx}^{11}-\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Phi}}_{1:k}^u - \widehat{\boldsymbol{\Sigma}}_{yx}^{12}\widehat{\boldsymbol{\Phi}}_{1:k}^l \\ &:= \boldsymbol{R}_2. \end{aligned} \tag{7.17}$$

The equality (7.8) is equivalent to

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^u + \widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Psi}}_{1:k}^l + \big(\widehat{\boldsymbol{\Sigma}}_{xy}^{22}-\widetilde{\boldsymbol{\Lambda}}_2\big)\widehat{\boldsymbol{\Psi}}_{1:k}^l &= \widehat{\boldsymbol{\Sigma}}_x^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^u\boldsymbol{\Lambda}_1 + \widehat{\boldsymbol{\Sigma}}_x^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^u(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1) \\ &\quad + \widehat{\boldsymbol{\Phi}}_{1:k}^l\boldsymbol{\Lambda}_1 + \big(\widehat{\boldsymbol{\Sigma}}_x^{22}\widehat{\boldsymbol{\Phi}}_{1:k}^l\widehat{\boldsymbol{\Lambda}}_1 - \widehat{\boldsymbol{\Phi}}_{1:k}^l\boldsymbol{\Lambda}_1\big),\end{aligned}$$

which can be written as

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}+\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Psi}}_{1:k}^{l}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\boldsymbol{\Lambda}_1-\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1\\
&\quad=\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1)+\big(\widehat{\boldsymbol{\Sigma}}_{x}^{22}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_1-\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1\big)-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{22}-\widetilde{\boldsymbol{\Lambda}}_2\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\\
&\quad:=\boldsymbol{R}_3. \qquad (7.18)
\end{aligned}$$

Apply the same argument to (7.9), we obtain

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{yx}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\widetilde{\boldsymbol{\Lambda}}_2^{\top}\widehat{\boldsymbol{\Phi}}_{1:k}^{l}-\widehat{\boldsymbol{\Sigma}}_{y}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1-\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\boldsymbol{\Lambda}_1\\
&\quad=\widehat{\boldsymbol{\Sigma}}_{y}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}(\widehat{\boldsymbol{\Lambda}}_1-\boldsymbol{\Lambda}_1)+\big(\widehat{\boldsymbol{\Sigma}}_{y}^{22}\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\widehat{\boldsymbol{\Lambda}}_1-\widehat{\boldsymbol{\Psi}}_{1:k}^{l}\boldsymbol{\Lambda}_1\big)-\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{22}-\widetilde{\boldsymbol{\Lambda}}_2^{\top}\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\\
&\quad:=\boldsymbol{R}_4. \qquad (7.19)
\end{aligned}$$

Consider (7.18) $\times(-\boldsymbol{\Lambda}_1)-\widetilde{\boldsymbol{\Lambda}}_2\times$ (7.19), then

$$\begin{aligned}
&\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1^2-\boldsymbol{\Lambda}_2^2\widehat{\boldsymbol{\Phi}}_{1:k}^{l}+\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\boldsymbol{\Lambda}_1^2-\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1-\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{yx}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{y}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1\\
&\quad=-(\boldsymbol{R}_3\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{R}_4),
\end{aligned}$$

that is

$$\begin{aligned}
\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1^2-\boldsymbol{\Lambda}_2^2\widehat{\boldsymbol{\Phi}}_{1:k}^{l}&=\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{yx}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\\
&\quad-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\boldsymbol{\Lambda}_1^2-\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{y}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1-(\boldsymbol{R}_3\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{R}_4). \qquad (7.20)
\end{aligned}$$

Combined with (7.16) and (7.17),

$$\begin{aligned}
\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1^2-\boldsymbol{\Lambda}_2^2\widehat{\boldsymbol{\Phi}}_{1:k}^{l}&=\widehat{\boldsymbol{\Sigma}}_{xy}^{21}\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{yx}^{21}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_1\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1+\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{R}_1\boldsymbol{\Lambda}_1\\
&\quad-\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{\Lambda}_1\widehat{\boldsymbol{\Phi}}_{1:k}^{u}-\widetilde{\boldsymbol{\Lambda}}_2\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{R}_2-(\boldsymbol{R}_3\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{R}_4)\\
&=\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Psi}}_{1:k}^{u}\boldsymbol{\Lambda}_1+\widetilde{\boldsymbol{\Lambda}}_2\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{21}-\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{u}\\
&\quad+\big(\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{R}_1-\boldsymbol{R}_3\big)\boldsymbol{\Lambda}_1-\widetilde{\boldsymbol{\Lambda}}_2\big(\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{R}_2+\boldsymbol{R}_4\big). \qquad (7.21)
\end{aligned}$$

This finishes the proof of (7.13).

Plug (7.17) into (7.21), we get

$$\begin{aligned}
\widehat{\boldsymbol{\Phi}}_{1:k}^{l}\boldsymbol{\Lambda}_1^2-\widetilde{\boldsymbol{\Lambda}}_2^2\widehat{\boldsymbol{\Phi}}_{1:k}^{l}&=\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_1\big)\big(\boldsymbol{\Lambda}_1\widehat{\boldsymbol{\Phi}}_{1:k}^{u}-\boldsymbol{R}_2\big)+\widetilde{\boldsymbol{\Lambda}}_2\big(\widehat{\boldsymbol{\Sigma}}_{yx}^{21}-\widehat{\boldsymbol{\Sigma}}_{y}^{21}\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\widetilde{\boldsymbol{R}}\\
&=\boldsymbol{B}\widehat{\boldsymbol{\Phi}}_{1:k}^{u}+\big(\widetilde{\boldsymbol{R}}-\big(\widehat{\boldsymbol{\Sigma}}_{xy}^{21}-\widehat{\boldsymbol{\Sigma}}_{x}^{21}\boldsymbol{\Lambda}_1\big)\boldsymbol{R}_2\big).
\end{aligned}$$

This finishes the proof of (7.12).

7.7.3. *Proof of Lemma 7.4*

First, we discuss two quite different cases: $\lambda_k \geq \frac{1}{2}$ and $\lambda_k < \frac{1}{2}$.

Case 1: $\lambda_k \geq \frac{1}{2}$

Let

$$\delta := \lambda_k^2 - \lambda_{k+1}^2 = (\lambda_k - \lambda_{k+1})(\lambda_k + \lambda_{k+1}) \geq \frac{1}{2}\Delta.$$

Define the $(p_1 - k) \times k$ matrices $\boldsymbol{A}$ by

$$A_{ij} = \frac{\sqrt{\lambda_j^2 - \lambda_k^2 + \frac{\delta}{2}}\sqrt{\lambda_{k+1}^2 - \lambda_{k+i}^2 + \frac{\delta}{2}}}{\lambda_j^2 - \lambda_{k+i}^2}, \quad 1 \leq i \leq p_1 - k, 1 \leq j \leq k$$

By (7.12) in Lemma 7.3, there holds

$$\widehat{\boldsymbol{\Phi}}_{1:k}^l = \boldsymbol{A} \circ \left(\boldsymbol{D}_1 \boldsymbol{B} \widehat{\boldsymbol{\Phi}}_{1:k}^u \boldsymbol{D}_2\right) + \boldsymbol{A} \circ (\boldsymbol{D}_1 \boldsymbol{R} \boldsymbol{D}_2),$$

where

$$\boldsymbol{D}_1 = \operatorname{diag}\left(\frac{1}{\sqrt{\frac{\delta}{2}}}, \ldots, \frac{1}{\sqrt{\lambda_{k+1}^2 - \lambda_{p_1}^2 + \frac{\delta}{2}}}\right)$$

and

$$\boldsymbol{D}_2 = \operatorname{diag}\left(\frac{1}{\sqrt{\lambda_1^2 - \lambda_k^2 + \frac{\delta}{2}}}, \ldots, \frac{1}{\sqrt{\frac{\delta}{2}}}\right).$$

By Lemma 7.6, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Phi}}_{1:k}^l\right\| &\leq \frac{1}{2}\left\|\boldsymbol{D}_1 \boldsymbol{B} \widehat{\boldsymbol{\Phi}}_{1:k}^u \boldsymbol{D}_2\right\| + \frac{1}{2}\left\|(\boldsymbol{D}_1 \boldsymbol{R} \boldsymbol{D}_2)\right\| \\ &\leq \frac{1}{2}\|\boldsymbol{D}_1 \boldsymbol{B}\|\left\|\widehat{\boldsymbol{\Phi}}_{1:k}^u\right\|\|\boldsymbol{D}_2\| + \frac{1}{2}\|\boldsymbol{D}_1\|\|\boldsymbol{R}\|\|\boldsymbol{D}_2\|.\end{aligned}$$

Recall that $\|\widehat{\boldsymbol{\Phi}}_{1:k}^u\| \leq \|\widehat{\boldsymbol{\Phi}}_{1:k}\| \leq \sqrt{\frac{3}{2}}$ and it is obvious that $\|\boldsymbol{D}_1\|, \|\boldsymbol{D}_2\| \leq \sqrt{\frac{2}{\delta}}$. Moreover, in the previous section, we also have shown that $\|\boldsymbol{R}\| \leq \frac{C(p_1+p_2)}{n\Delta}$. It suffices to bound $\|\boldsymbol{D}_1 \boldsymbol{B}\|$ and to this end we apply the standard covering argument.

Step 1. Reduction. Denote by $\mathcal{N}_\epsilon(\mathbb{S}^d)$ the d-dimensional unit ball surface. For $\epsilon > 0$ and any pair of vectors $\boldsymbol{u} \in \mathbb{R}^{p_1-k}$, $\boldsymbol{v} \in \mathbb{R}^k$, we can choose $\boldsymbol{u}_\epsilon \in \mathcal{N}_\epsilon(\mathbb{S}^{p_1-k-1})$, $\boldsymbol{v}_\epsilon \in \mathcal{N}_\epsilon(\mathbb{S}^{k-1})$ such that $\|\boldsymbol{u} - \boldsymbol{u}_\epsilon\|, \|\boldsymbol{v} - \boldsymbol{v}_\epsilon\| \leq \epsilon$. Then

$$\begin{aligned}\boldsymbol{u}^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v} &= \boldsymbol{u}^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v} - \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v} + \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v} - \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}_\epsilon + \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}_\epsilon \\ &\leq \|\boldsymbol{u} - \boldsymbol{u}_\epsilon\|\|\boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}\| + \left\|\boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B}\right\|\|\boldsymbol{v} - \boldsymbol{v}_\epsilon\| + \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}_\epsilon\end{aligned}$$

$$\le 2\epsilon\|\boldsymbol{D}_1\boldsymbol{B}\| + \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v}_\epsilon$$
$$\le 2\epsilon\|\boldsymbol{D}_1\boldsymbol{B}\| + \max_{\boldsymbol{u}_\epsilon,\boldsymbol{v}_\epsilon} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v}_\epsilon.$$

Maximize over $\boldsymbol{u}$ and $\boldsymbol{v}$, we obtain

$$\|\boldsymbol{D}_1\boldsymbol{B}\| \le 2\epsilon\|\boldsymbol{D}_1\boldsymbol{B}\| + \max_{\boldsymbol{u}_\epsilon,\boldsymbol{v}_\epsilon} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v}_\epsilon.$$

Therefore, $\|\boldsymbol{D}_1\boldsymbol{B}\| \le (1-2\epsilon)^{-1}\max_{\boldsymbol{u}_\epsilon,\boldsymbol{v}_\epsilon} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v}_\epsilon$. Let $\epsilon = 1/4$. Then it suffices to give an upper bound $\max_{\boldsymbol{u}_\epsilon,\boldsymbol{v}_\epsilon} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v}_\epsilon$ with high probability.

Step 2. Concentration. Let $Z_{\alpha,l} = \frac{Y_{\alpha,l}-\lambda_l X_l}{\sqrt{1-\lambda_l^2}}$ for all $1 \le \alpha \le n$ and $1 \le l \le p_1$. Then for $1 \le i \le p_1 - k$ and $1 \le j \le k$

$$\begin{aligned}
&[\boldsymbol{D}_1\boldsymbol{B}]_{i,j} \\
&= \frac{1}{\sqrt{\lambda_{k+1}^2 - \lambda_{k+i}^2 + \frac{\delta}{2}}} \frac{1}{n}\sum_{\alpha=1}^n \big(\lambda_j X_{\alpha,k+i} Y_{\alpha,j} \\
&\quad - \lambda_j^2 X_{\alpha,k+i} X_{\alpha,j} + \lambda_{k+i} Y_{\alpha,k+i} X_{\alpha,j} - \lambda_{k+i}\lambda_j Y_{\alpha,k+i} Y_{\alpha,j}\big) \\
&= \frac{1}{\sqrt{\lambda_{k+1}^2 - \lambda_{k+i}^2 + \frac{\delta}{2}}} \frac{1}{n}\sum_{\alpha=1}^n \big\{\big(1-\lambda_j^2\big)\lambda_{k+i}\lambda_j X_{\alpha,k+i} X_{\alpha,j} \\
&\quad - \lambda_j^2 (Y_{\alpha,k+i} - \lambda_{k+i} X_{\alpha,k+i})(Y_{\alpha,j} - \lambda_j X_{\alpha,j}) \\
&\quad + \big(1-\lambda_j^2\big)\lambda_j (Y_{\alpha,k+i} - \lambda_{k+i} X_{\alpha,k+i}) X_{\alpha,j} + \big(1-\lambda_j^2\big)\lambda_{k+i}(Y_{\alpha,j} - \lambda_j X_{\alpha,j}) X_{\alpha,k+i}\big\} \\
&= \frac{1}{\sqrt{\lambda_{k+1}^2 - \lambda_{k+i}^2 + \frac{\delta}{2}}} \frac{1}{n}\sum_{\alpha=1}^n \big\{\big(1-\lambda_j^2\big)\lambda_{k+i}\lambda_j X_{\alpha,k+i} X_{\alpha,j} \\
&\quad - \lambda_j^2\sqrt{1-\lambda_{k+i}^2}\sqrt{1-\lambda_j^2} Z_{\alpha,k+i} Z_{\alpha,j} + \big(1-\lambda_j^2\big)\lambda_j\sqrt{1-\lambda_{k+i}^2} Z_{\alpha,k+i} X_{\alpha,j} \\
&\quad + \big(1-\lambda_j^2\big)\lambda_{k+i}\sqrt{1-\lambda_{k+i}^2} X_{\alpha,k+i} Z_{\alpha,j}\big\}.
\end{aligned}$$

In this way, $\{X_{\alpha,k+i}, Z_{\alpha,k+i}, 1 \le i \le p_1, 1 \le \alpha \le n\}$ are mutually independent standard gaussian random variables. For any given pair of vectors $\boldsymbol{u} \in \mathbb{R}^{p_1-k}$, $\boldsymbol{v} \in \mathbb{R}^k$,

$$\begin{aligned}
\boldsymbol{u}^\top \boldsymbol{D}_1\boldsymbol{B}\boldsymbol{v} = \frac{1}{n}\sum_{\alpha=1}^n \sum_{i=1}^{p_1-k}\sum_{j=1}^k \frac{u_i v_j}{\sqrt{\lambda_{k+1}^2 - \lambda_{k+i}^2 + \frac{\delta}{2}}} \big\{\big(1-\lambda_j^2\big)\lambda_{k+i}\lambda_j X_{\alpha,k+i} X_{\alpha,j} \\
- \lambda_j^2\sqrt{1-\lambda_{k+i}^2}\sqrt{1-\lambda_j^2} Z_{\alpha,k+i} Z_{\alpha,j} + \big(1-\lambda_j^2\big)\lambda_j\sqrt{1-\lambda_{k+i}^2} Z_{\alpha,k+i} X_{\alpha,j}
\end{aligned}$$

$$
\begin{aligned}
&+ \big(1-\lambda_j^2\big)\lambda_{k+i}\sqrt{1-\lambda_{k+i}^2}\,X_{\alpha,k+i}Z_{\alpha,j}\big\} \\
&\doteq \frac{1}{n}\sum_{\alpha=1}^{n} \boldsymbol{w}_\alpha^\top \boldsymbol{A}_\alpha \boldsymbol{w}_\alpha,
\end{aligned}
$$

where

$$
\boldsymbol{w}_\alpha^\top = \big[\boldsymbol{x}_\alpha^\top, \boldsymbol{z}_\alpha^\top\big] = [X_{\alpha,1},\ldots,X_{\alpha,p_1},Z_{\alpha,1},\ldots,Z_{\alpha,p_1}]
$$

and $\boldsymbol{A}_\alpha \in \mathbb{R}^{(2p_1)\times(2p_1)}$ is symmetric and determined by the corresponding quadratic form. This yields

$$
\begin{aligned}
\|\boldsymbol{A}_\alpha\|_F^2 &= \frac{1}{2}\sum_{i=1}^{p_1-k}\sum_{j=1}^{k}\frac{u_i^2 v_j^2}{\lambda_{k+1}^2-\lambda_{k+i}^2+\frac{\delta}{2}}\big\{\big(1-\lambda_j^2\big)^2\lambda_{k+i}^2\lambda_j^2+\lambda_j^4\big(1-\lambda_{k+i}^2\big)\big(1-\lambda_j^2\big) \\
&\quad + \big(1-\lambda_j^2\big)^2\lambda_j^2\big(1-\lambda_{k+i}^2\big)+\big(1-\lambda_j^2\big)^2\lambda_{k+i}^2\big(1-\lambda_{k+i}^2\big)\big\} \\
&= \frac{1}{2}\sum_{i=1}^{p_1-k}\sum_{j=1}^{k}\frac{u_i^2 v_j^2}{\lambda_{k+1}^2-\lambda_{k+i}^2+\frac{\delta}{2}}\big(1-\lambda_j^2\big)\big(\lambda_{k+i}^2+\lambda_j^2-2\lambda_{k+i}^2\lambda_j^2\big) \\
&\le \frac{1}{2}\left(\sum_{i=1}^{p_1-k}u_i^2\right)\left(\sum_{j=1}^{k}v_j^2\right)\max_{\substack{1\le i\le p_1-k\\ 1\le j\le k}}\frac{(1-\lambda_j^2)(\lambda_{k+i}^2+\lambda_j^2-2\lambda_{k+i}^2\lambda_j^2)}{\lambda_{k+1}^2-\lambda_{k+i}^2+\frac{\delta}{2}} \\
&\le \frac{1}{2}\max_{\substack{1\le i\le p_1-k\\ 1\le j\le k}}\frac{(1-\lambda_k^2)(2\lambda_j^2-2\lambda_{k+i}^2\lambda_j^2)}{\lambda_{k+1}^2-\lambda_{k+i}^2+\frac{\delta}{2}} \\
&\le \big(1-\lambda_k^2\big)\max_{\substack{1\le i\le p_1-k\\ 1\le j\le k}}\frac{\lambda_j^2(1-\lambda_{k+i}^2)}{\frac{\delta}{2}+\lambda_{k+1}^2-\lambda_{i+k}^2} \\
&\le \big(1-\lambda_k^2\big)\max_{\substack{1\le i\le p_1-k\\ 1\le j\le k}}\frac{(1-\lambda_{k+1}^2)}{\frac{\delta}{2}} \\
&\le \frac{2(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\delta} \doteq K^2,
\end{aligned}
$$

where the second last inequality is due to the facts that $\lambda_j \le 1$ and

$$
\frac{(1-\lambda_{k+i}^2)}{\frac{\delta}{2}+\lambda_{k+1}^2-\lambda_{i+k}^2} \le \frac{(1-\lambda_{k+1}^2)}{\frac{\delta}{2}} \quad \left(\because \frac{\delta}{2}+\lambda_{k+1}^2 < \lambda_k^2 \le 1\right).
$$

Moreover, $\|\boldsymbol{A}_\alpha\|_2^2 \le \|\boldsymbol{A}_\alpha\|_F^2 \le K^2$.

Now define $\boldsymbol{w}^\top := [\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_n^\top]$ and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_n \end{bmatrix}.$$

Then we have

$$\|\boldsymbol{A}\| \le \max_{1\le\alpha\le n} \|\boldsymbol{A}_\alpha\| \le K, \qquad \|\boldsymbol{A}\|_F^2 \le \sum_{\alpha=1}^{n} \|\boldsymbol{A}_\alpha\|_F^2 \le nK^2$$

and

$$\boldsymbol{u}^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v} = \frac{1}{n} \boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w}, \quad \text{where } \boldsymbol{w} \in \mathcal{N}_{2p_1 n}(\boldsymbol{0}, \boldsymbol{I}_{2p_1 n}).$$

Therefore, by the classic Hanson–Wright inequality (Lemma 7.11), there holds

$$P\left\{n\left|\boldsymbol{u}^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}\right| \ge t\right\} \le 2\exp\left\{-c_0 \min\left(\frac{t^2}{nK^2}, \frac{t}{K}\right)\right\}$$

for some numerical constant $c_0 > 0$. Without loss of generality, we can also assume $c_0 \le 1$. Let $t = \frac{4}{c_0}\sqrt{np_1}K$. By $n \ge p_1$, straightforward calculation gives

$$P\left\{n\left|\boldsymbol{u}^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}\right| \ge \frac{4}{c_0}\sqrt{np_1}K\right\} \le 2e^{-4p_1}.$$

Step 3. Union Bound. By Lemma 7.12, we choose $1/4$-net such that

$$P\left\{\max_{\substack{\boldsymbol{u}_\epsilon \in \mathcal{N}_\epsilon(\mathbb{S}^{p_1-k-1}) \\ \boldsymbol{v}_\epsilon \in \mathcal{N}_\epsilon(\mathbb{S}^{k-1})}} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}_\epsilon \ge \left(\frac{4\sqrt{2}}{c_0}\right)\sqrt{\frac{p_1}{n}}\sqrt{\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\delta}}\right\}$$
$$\le 9^{p_1-k}9^k \times 2e^{-4p_1} \le 2e^{-\frac{3}{2}p_1}.$$

In other words, with probability at least $1 - 2e^{-\frac{3}{2}p_1}$, we have

$$\|\boldsymbol{D}_1 \boldsymbol{B}\| \le (1-2\epsilon)^{-1} \max_{\boldsymbol{u}_\epsilon, \boldsymbol{v}_\epsilon} \boldsymbol{u}_\epsilon^\top \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{v}_\epsilon \le \left(\frac{8\sqrt{2}}{c_0}\right)\sqrt{\frac{p_1}{n}}\sqrt{\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\delta}}.$$

In summary, we have as long as $n \geq C_0(p_1+p_2)$, with probability $1-c_0\exp(-\gamma p_1)$,

$$\begin{aligned}\big\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\big\| &\leq C\left[\sqrt{\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\delta^2}}+\frac{(p_1+p_2)}{n\Delta\delta}\right]\\ &\leq C\left[\sqrt{\frac{p_1(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2}}+\frac{(p_1+p_2)}{n\Delta^2}\right].\end{aligned}$$

Here the last inequality is due to $\delta=(\lambda_k+\lambda_{k+1})\Delta\geq\frac{1}{2}\Delta$. Here C_0, C, c_0, γ are absolute constants.

Case 2: $\lambda_k \leq \frac{1}{2}$

By (7.13), we have

$$\widehat{\boldsymbol{\Phi}}^l_{1:k}\boldsymbol{\Lambda}_1^2-\boldsymbol{\Lambda}_2^2\widehat{\boldsymbol{\Phi}}^l_{1:k}=\boldsymbol{G}\boldsymbol{\Lambda}_1+\boldsymbol{\Lambda}_2\boldsymbol{F},$$

where

$$\boldsymbol{G}:=\big(\widehat{\boldsymbol{\Sigma}}^{21}_{xy}-\widehat{\boldsymbol{\Sigma}}^{21}_{x}\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Psi}}^u_{1:k}+\big(\widehat{\boldsymbol{\Sigma}}^{21}_{x}\boldsymbol{R}_1-\boldsymbol{R}_3\big)$$

and

$$\boldsymbol{F}:=[\boldsymbol{I}_{p_1},\boldsymbol{0}_{p_1\times(p_2-p_1)}]\big[\big(\widehat{\boldsymbol{\Sigma}}^{21}_{yx}-\widehat{\boldsymbol{\Sigma}}^{21}_{y}\boldsymbol{\Lambda}_1\big)\widehat{\boldsymbol{\Phi}}^u_{1:k}-\big(\widehat{\boldsymbol{\Sigma}}^{21}_{y}\boldsymbol{R}_2+\boldsymbol{R}_4\big)\big].$$

Notice that $\widehat{\boldsymbol{\Sigma}}^{21}_{xy}$ and $\widehat{\boldsymbol{\Sigma}}^{21}_{x}$ are submatrices of $\widehat{\boldsymbol{\Sigma}}_{2p_1}$. By Lemma 7.1, we have

$$\big\|\widehat{\boldsymbol{\Sigma}}^{21}_{xy}-\widehat{\boldsymbol{\Sigma}}^{21}_{x}\boldsymbol{\Lambda}_1\big\|\leq C\sqrt{\frac{p_1}{n}}.$$

Moreover, by $\|\boldsymbol{R}_1\|\leq C\sqrt{\frac{p_1+p_2}{n}}$, $\|\boldsymbol{R}_3\|\leq C\frac{p_1+p_2}{n\Delta}$ and Lemma 7.2, there holds

$$\|\boldsymbol{G}\|\leq C\left(\sqrt{\frac{p_1}{n}}+\frac{p_1+p_2}{n\Delta}\right).$$

Similarly, $[\boldsymbol{I}_{p_1},\boldsymbol{0}_{p_1\times(p_2-p_1)}]\widehat{\boldsymbol{\Sigma}}^{21}_{yx}$ and $[\boldsymbol{I}_{p_1},\boldsymbol{0}_{p_1\times(p_2-p_1)}]\widehat{\boldsymbol{\Sigma}}^{21}_{x}$ are submatrices of $\widehat{\boldsymbol{\Sigma}}_{2p_1}$. By a similar argument,

$$\|\boldsymbol{F}\|\leq C\left(\sqrt{\frac{p_1}{n}}+\frac{p_1+p_2}{n\Delta}\right).$$

Then

$$\widehat{\boldsymbol{\Phi}}^l_{1:k}=\left[\frac{\lambda_j}{\lambda_{k+i}+\lambda_j}\right]\circ\left[\frac{1}{\lambda_j-\lambda_{k+i}}\right]\circ\boldsymbol{G}+\left[\frac{\lambda_{k+i}}{\lambda_{k+i}+\lambda_j}\right]\circ\left[\frac{1}{\lambda_j-\lambda_{k+i}}\right]\circ\boldsymbol{F}$$

Here $1 \le i \le p_1 - k$ and $1 \le j \le k$. By Lemma 7.6, there holds for any $\boldsymbol{X}$,

$$\left\| \left[\frac{\lambda_j}{\lambda_{k+i} + \lambda_j} \right] \boldsymbol{X} \right\| = \left\| \left[\frac{\max(\lambda_{k+i}, \lambda_j)}{\lambda_{k+i} + \lambda_j} \right] \boldsymbol{X} \right\| \le \frac{3}{2} \|\boldsymbol{X}\|$$

and

$$\left\| \left[\frac{\lambda_{k+i}}{\lambda_{k+i} + \lambda_j} \right] \boldsymbol{X} \right\| = \left\| \left[\frac{\min(\lambda_{k+i}, \lambda_j)}{\lambda_{k+i} + \lambda_j} \right] \boldsymbol{X} \right\| \le \frac{1}{2} \|\boldsymbol{X}\|.$$

Finally, for any $\boldsymbol{X}$,

$$\left[\frac{1}{\lambda_j - \lambda_{k+i}} \right] \boldsymbol{X} = \boldsymbol{A} \circ (\boldsymbol{D}_1 \boldsymbol{X} \boldsymbol{D}_2)$$

where

$$\boldsymbol{A} := \left[\frac{\sqrt{\lambda_j - \lambda_k + \frac{\Delta}{2}} \sqrt{\lambda_{k+1} - \lambda_{k+i} + \frac{\Delta}{2}}}{\lambda_j - \lambda_{k+i}} \right],$$

$$\boldsymbol{D}_1 = \operatorname{diag}\left(\frac{1}{\sqrt{\frac{\Delta}{2}}}, \ldots, \frac{1}{\sqrt{\lambda_{k+1} - \lambda_{p_1} + \frac{\Delta}{2}}} \right),$$

and

$$\boldsymbol{D}_2 = \operatorname{diag}\left(\frac{1}{\sqrt{\lambda_1 - \lambda_k + \frac{\Delta}{2}}}, \ldots, \frac{1}{\sqrt{\frac{\Delta}{2}}} \right).$$

Since $\|\boldsymbol{D}_1\|, \|\boldsymbol{D}_2\| \le \sqrt{\frac{2}{\Delta}}$, by Lemma 7.6,

$$\left\| \left[\frac{1}{\lambda_j - \lambda_{k+i}} \right] \boldsymbol{X} \right\| \le \frac{1}{2} \|\boldsymbol{D}_1 \boldsymbol{X} \boldsymbol{D}_2\| \le \frac{1}{\Delta} \|\boldsymbol{X}\|.$$

In summary, we have

$$\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\| \le C \left(\sqrt{\frac{p_1}{n\Delta^2}} + \frac{p_1 + p_2}{n\Delta^2} \right).$$

Since $\frac{1}{2} \ge \lambda_k \ge \lambda_{k+1}$, there holds

$$\|\widehat{\boldsymbol{\Phi}}^l_{1:k}\| \le C \left[\sqrt{\frac{p_1(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2}} + \frac{(p_1 + p_2)}{n\Delta^2} \right].$$

7.8. Special case: $\lambda_k = 1$

Recall that

$$\boldsymbol{\Sigma}_x = \boldsymbol{I}_{p_1}, \qquad \boldsymbol{\Sigma}_y = \boldsymbol{I}_{p_2}, \qquad \boldsymbol{\Sigma}_{xy} = [\boldsymbol{\Lambda}, \boldsymbol{0}_{p_1\times(p_2-p_1)}] := \widetilde{\boldsymbol{\Lambda}},$$
$$\boldsymbol{\Sigma} := \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^\top & \boldsymbol{\Sigma}_y \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{p_1} & \widetilde{\boldsymbol{\Lambda}} \\ \widetilde{\boldsymbol{\Lambda}}^\top & \boldsymbol{I}_{p_2} \end{bmatrix}.$$

If $\lambda_1 = \cdots = \lambda_k = 1 > \lambda_{k+1}$, we have $\operatorname{rank}(\boldsymbol{\Sigma}) = p_1 + p_2 - k$. Moreover, since the joint distribution of $\boldsymbol{x}^\top = [X_1, \ldots, X_{p_1}]$ and $\boldsymbol{y}^\top = [Y_1, \ldots, Y_{p_2}]$ is multivariate normal, there must hold

$$X_i = Y_i, \quad 1 \le i \le k.$$

Then the first k columns of $\boldsymbol{X} \in \mathbb{R}^{n\times p_1}$ and those of $\boldsymbol{Y} \in \mathbb{R}^{n\times p_2}$ are identical, respectively, which implies

$$(\widehat{\boldsymbol{\Sigma}}_x)_{1:k,1:k} = (\widehat{\boldsymbol{\Sigma}}_y)_{1:k,1:k} = (\widehat{\boldsymbol{\Sigma}}_{xy})_{1:k,1:k},$$

where the subscript represents the upper left $k \times k$ submatrices. Moreover, as long as $n > p_1 + p_2$, we know with probability one there holds

$$\operatorname{rank}(\boldsymbol{X}) = p_1, \qquad \operatorname{rank}(\boldsymbol{Y}) = p_2, \qquad \operatorname{rank}\big([\boldsymbol{X}, \boldsymbol{Y}]\big) = p_1 + p_2 - k,$$

which also implies that with probability one $\hat{\lambda}_{k+1} < 1$.

Define $\widehat{\boldsymbol{U}}_k$, $\widehat{\boldsymbol{V}}_k$ as

$$\widehat{\boldsymbol{U}}_k = \boldsymbol{I}_{p_1,k}(\widehat{\boldsymbol{\Sigma}}_x)^{-\frac{1}{2}}_{1:k,1:k}, \qquad \widehat{\boldsymbol{V}}_k = \boldsymbol{I}_{p_2,k}(\widehat{\boldsymbol{\Sigma}}_y)^{-\frac{1}{2}}_{1:k,1:k},$$

where $\boldsymbol{I}_{p,k}$ denotes the first k columns of the $p \times p$ dimensional identity matrix $\boldsymbol{I}_p$. It is straightforward to verify that

$$\widehat{\boldsymbol{U}}_k^\top \widehat{\boldsymbol{\Sigma}}_x \widehat{\boldsymbol{U}}_k = \boldsymbol{I}_k, \qquad \widehat{\boldsymbol{V}}_k^\top \widehat{\boldsymbol{\Sigma}}_y \widehat{\boldsymbol{V}}_k = \boldsymbol{I}_k$$

and

$$\widehat{\boldsymbol{U}}_k^\top \widehat{\boldsymbol{\Sigma}}_{xy} \widehat{\boldsymbol{V}}_k = \boldsymbol{I}_k.$$

Notice that the sample canonical correlations are at most 1, by the definition in equation (1.7), we have $\hat{\lambda}_1 = \cdots = \hat{\lambda}_k = 1$ and $(\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k)$ is one of the solutions for the leading k pair of sample canonical loadings. The fact $\hat{\lambda}_{k+1} < 1$ implies that the subspace spanned by the top k left/right singular vectors of $\widehat{\boldsymbol{\Sigma}}_x^{-1/2}\widehat{\boldsymbol{\Sigma}}_{xy}\widehat{\boldsymbol{\Sigma}}_y^{-1/2}$ is unique, which further implies that the subspace spanned by the top k left/right sample canonical loadings is unique. Then for any top k sample canonical loading matrix $\widehat{\boldsymbol{\Phi}}_{1:k}$, the column space of $\widehat{\boldsymbol{\Phi}}_{1:k}$ must equal to the column space of $\widehat{\boldsymbol{U}}_k$, namely, the column space of $\boldsymbol{I}_{p_1,k}$. Thus, $\widehat{\boldsymbol{\Phi}}^l_{1:k} = \boldsymbol{0}$. Substituting this into equation (7.3) and (7.4) reveals that both loss functions are zero.

Supplementary Material

Supplement to "Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates" (DOI: 10.3150/19-BEJ1131SUPP; .pdf). We give a complete proof of Theorem 2.2 in [20].

References

[1] Anderson, T.W. (1999). Asymptotic theory for canonical correlation analysis. *J. Multivariate Anal.* **70** 1–29. MR1701396 https://doi.org/10.1006/jmva.1999.1810

[2] Arora, R. and Livescu, K. (2013). Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *Acoustics*, *Speech and Signal Processing* (*ICASSP*), 2013 *IEEE International Conference on* 7135–7139. IEEE.

[3] Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281 https://doi.org/10.1007/s00440-014-0562-z

[4] Cai, T.T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

[5] Cai, T.T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. MR3766946 https://doi.org/10.1214/17-AOS1541

[6] Chaudhuri, K., Kakade, S.M., Livescu, K. and Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 129–136. ACM.

[7] Chen, X., Liu, H. and Carbonell, J.G. (2012). Structured sparse canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics* 199–207.

[8] Dhillon, P.S., Foster, D. and Ungar, L. (2011). Multi-view learning of word embeddings via cca. In *Advances in Neural Information Processing Systems* (*NIPS*), *Vol.* 24.

[9] Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. Association for Computational Linguistics.

[10] Foster, D.P., Johnson, R., Kakade, S.M. and Zhang, T. (2008). Multi-view dimensionality reduction via canonical correlation analysis. Technical report.

[11] Friman, O., Borga, M., Lundberg, P. and Knutsson, H. (2003). Adaptive analysis of fmri data. *NeuroImage* **19** 837–845.

[12] Fukumizu, K., Bach, F.R. and Jordan, M.I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37** 1871–1905. MR2533474 https://doi.org/10.1214/08-AOS637

[13] Gao, C., Ma, Z., Ren, Z. and Zhou, H.H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982 https://doi.org/10.1214/15-AOS1332

[14] Gao, C., Ma, Z. and Zhou, H.H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. MR3718162 https://doi.org/10.1214/16-AOS1519

[15] Gong, Y., Ke, Q., Isard, M. and Lazebnik, S. (2014). A multi-view embedding space for modeling Internet images, tags, and their semantics. *Int. J. Comput. Vis.* **106** 210–233.

[16] Horn, R.A. and Johnson, C.R. (1991). *Topics in Matrix Analysis*. Cambridge: Cambridge Univ. Press. MR1091716 https://doi.org/10.1017/CBO9780511840371

[17] Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28** 312–377.

[18] Kakade, S.M. and Foster, D.P. (2007). Multi-view regression via canonical correlation analysis. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 82–96. Berlin: Springer. MR2397580 https://doi.org/10.1007/978-3-540-72927-3_8

[19] Kim, T.-K., Wong, S.-F. and Cipolla, R. (2007). Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition*, 2007. *CVPR'*07. *IEEE Conference on* 1–8. IEEE.

[20] Ma, Z. and Li, X. (2020). Supplement to "Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates." https://doi.org/10.3150/19-BEJ1131SUPP.

[21] Mathias, R. (1993). The Hadamard operator norm of a circulant and applications. *SIAM J. Matrix Anal. Appl.* **14** 1152–1167. MR1238930 https://doi.org/10.1137/0614080

[22] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R. and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the* 18*th ACM International Conference on Multimedia* 251–260. ACM.

[23] Rudelson, M. and Vershynin, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258 https://doi.org/10.1214/ECP.v18-2865

[24] Sridharan, K. and Kakade, S.M. (2008). An information theoretic framework for multi-view learning. In *COLT* (R.A. Servedio and T. Zhang, eds.) 403–414. Omnipress.

[25] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[26] Vu, V.Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452 https://doi.org/10.1214/13-AOS1151

[27] Wang, W., Arora, R., Livescu, K. and Bilmes, J. (2015). On deep multi-view representation learning. In *Proceedings of the* 32*nd International Conference on Machine Learning* (*ICML*-15) 1083–1092.

[28] Wedin, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968 https://doi.org/10.1007/bf01932678

[29] Wedin, P.Å. (1983). On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils* 263–285. Springer.

[30] Witten, D.M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* kxp008.

Received February 2018 and revised December 2018