D-CCA: A Decomposition-based Canonical Correlation Analysis for High-Dimensional Datasets

Hai Shu[a], Xiao Wang[b], and Hongtu Zhu[a,c*]
[a]Department of Biostatistics, The University of Texas MD Anderson Cancer Center
[b]Department of Statistics, Purdue University
[c]Department of Biostatistics, The University of North Carolina at Chapel Hill

Abstract

A typical approach to the joint analysis of two high-dimensional datasets is to decompose each data matrix into three parts: a low-rank common matrix that captures the shared information across datasets, a low-rank distinctive matrix that characterizes the individual information within a single dataset, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common and distinctive matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive matrices. The latter guarantees that no more shared information is extractable from the distinctive matrices. We propose decomposition-based canonical correlation analysis (D-CCA), a novel decomposition method that defines the common and distinctive matrices from the $\mathcal{L}^2$ space of random variables rather than the conventionally used Euclidean space, with a careful construction of the orthogonal relationship between distinctive matrices. D-CCA represents a natural generalization of the traditional canonical correlation analysis. The proposed estimators of common and distinctive matrices are shown to be consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and the real data analysis of breast cancer data obtained from The Cancer Genome Atlas.

Keywords: approximate factor model, canonical variable, common structure, distinctive structure, soft thresholding.

*Address for correspondence and reprints: Hongtu Zhu, Ph.D., E-mail: htzhu@email.unc.edu; Phone No: 919-929-9010. Dr. Zhu's work was partially supported by NIH grants MH086633 and MH116527, NSF grants SES-1357666 and DMS-1407655, a grant from the Cancer Prevention Research Institute of Texas, and the endowed Bao-Shan Jing Professorship in Diagnostic Imaging. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

1 Introduction

Many large biomedical studies have collected high-dimensional genetic and/or imaging data and associated data (e.g., clinical data) from increasingly large cohorts to delineate the complex genetic and environmental contributors to many diseases, such as cancer and Alzheimer's disease. For example, The Cancer Genome Atlas (TCGA; Koboldt et al., 2012) project collected human tumor specimens and derived different types of large-scale genomic data such as mRNA expression and DNA methylation to enhance the understanding of cancer biology and therapy. The Human Connectome Project (Van Essen et al., 2013) acquired imaging datasets from multiple modalities (HARDI, R-fMRI, T-fMRI, MEG) across a large cohort to build a "network map" (connectome) of the anatomical and functional connectivity within the healthy human brain. These cross-platform datasets share some common information, but individually contain distinctive patterns. Disentangling the underlying common and distinctive patterns is critically important for facilitating the integrative and discriminative analysis of these cross-platform datasets (van der Kloet et al., 2016; Smilde et al., 2017).

Throughout this paper, we focus on disentangling the common and distinctive patterns of two high-dimensional datasets written as matrices $\mathbf{Y}_k \in \mathbb{R}^{p_k \times n}$ for $k = 1, 2$ on a common set of n objects, where each of the p_k rows corresponds to a mean-zero variable. A popular approach to such an analysis is to decompose each data matrix into three parts:

$$\mathbf{Y}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k \quad \text{for} \quad k = 1, 2, \tag{1}$$

where $\mathbf{C}_k$'s are low-rank "common" matrices that capture the shared structure between datasets, $\mathbf{D}_k$'s are low-rank "distinctive" matrices that capture the individual structure within each dataset, and $\mathbf{E}_k$'s are additive noise matrices. Model (1) has been widely used in genomics (Lock et al., 2013; O'Connell and Lock, 2016), metabolomics (Kuligowski et al., 2015), and neuroscience (Yu et al., 2017), among other areas of research. Ideally, the common and distinctive matrices should provide different "views" for each individual dataset, while borrowing information from the other. A fundamental question for model (1) is how to decompose $\mathbf{Y}_k$'s into the common and distinctive matrices within each dataset and across datasets.

Most decomposition methods for model (1) are based on the Euclidean space $(\mathbb{R}^n, \cdot)$ endowed with the dot product. Such methods include JIVE (Lock et al., 2013), angle-based JIVE (AJIVE;

Feng et al., 2018), OnPLS (Trygg, 2002; Löfstedt and Trygg, 2011), COBE (Zhou et al., 2016), and DISCO-SCA (Schouteden et al., 2014). A common characteristic among all these methods is to enforce the row-space orthogonality between the common and distinctive matrices within each dataset, that is, $\mathbf{C}_k\mathbf{D}_k^\top = \mathbf{0}$ for $k = 1, 2$. With the exception of OnPLS, these methods impose additional orthogonality across the datasets, that is, $\mathbf{C}_k\mathbf{D}_\ell^\top = \mathbf{0}$ for all k and ℓ. A potential issue associated with these methods is that they inadequately consider the more desired orthogonality between the distinctive matrices $\mathbf{D}_1$ and $\mathbf{D}_2$, which guarantees that no common structure is retained therein. Specifically, the first four methods do not impose any orthogonality constraint between $\mathbf{D}_1$ and $\mathbf{D}_2$. Although DISCO-SCA and a modified JIVE (O'Connell and Lock (2016); denoted as R.JIVE) have considered the row-space orthogonality between the distinctive matrices, it may be incompatible with their orthogonal condition that $\mathbf{C}_k\mathbf{D}_\ell^\top = \mathbf{0}$ for all $k, \ell = 1, 2$ even as $p_1 = p_2 = 1$.

Rather than the conventionally used Euclidean space $(\mathbb{R}^n, \cdot)$, the aim of this paper is to develop a new decomposition method for model (1) based on the inner product space $(\mathcal{L}_0^2, \mathrm{cov})$, which is the vector space composed of all zero-mean and finite-variance real-valued random variables and endowed with the covariance operator as the inner product. Specifically, model (1) is a sample-matrix version of the prototype given by

$$\boldsymbol{y}_k = \boldsymbol{c}_k + \boldsymbol{d}_k + \boldsymbol{e}_k \in \mathbb{R}^{p_k} \quad \text{for} \quad k = 1, 2. \tag{2}$$

The Euclidean space $(\mathbb{R}^n, \cdot)$ is hence not an appropriate space for defining the common matrices $\mathbf{C}_k$'s and the distinctive matrices $\mathbf{D}_k$'s, because two uncorrelated non-constant random variables will almost never have zero sample correlation, i.e., the orthogonality in $(\mathbb{R}^n, \cdot)$. The matrices $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^2$ defined by the aforementioned methods based on $(\mathbb{R}^n, \cdot)$ are, in fact, estimators of the counterparts defined through model (2) on $(\mathcal{L}_0^2, \mathrm{cov})$. Instead, for model (2), we introduce a common-space constraint for the common vectors $\{\boldsymbol{c}_k\}_{k=1}^2$, an orthogonal-space constraint for the distinctive vectors $\{\boldsymbol{d}_k\}_{k=1}^2$, and a parsimonious-representation constraint for the signal vectors $\boldsymbol{x}_k \coloneqq \boldsymbol{y}_k - \boldsymbol{e}_k, k = 1, 2$ as follows:

$$\mathrm{span}(\boldsymbol{c}_1^\top) = \mathrm{span}(\boldsymbol{c}_2^\top), \tag{3}$$

$$\mathrm{span}(\boldsymbol{d}_1^\top) \perp \mathrm{span}(\boldsymbol{d}_2^\top), \tag{4}$$

$$\text{span}((\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top)) = \text{span}((\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \boldsymbol{d}_1^\top, \boldsymbol{d}_2^\top)), \tag{5}$$

where $\text{span}(\boldsymbol{v}^\top) = \text{span}(\{v_j\}_{j=1}^p) = \left\{ \sum_{j=1}^p a_j v_j : \forall a_j \in \mathbb{R} \right\}$ is the vector space spanned by entries of any random vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$, and $\perp$ denotes the orthogonality between two subspaces and/or random variables in $(\mathcal{L}_0^2, \text{cov})$. The orthogonal relationship between distinctive matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ is now described by (4).

To illustrate the advantage of our proposed constraints over those imposed by the six existing methods mentioned above, we consider a toy example based on model (2) with $p_1 = p_2 = 1$. Suppose z_1 and z_2 are two standardized signal random variables with the same distribution and $\text{corr}(z_1, z_2) \in (0, 1)$, i.e., their angle on $(\mathcal{L}_0^2, \text{cov})$, denoted as θ, in $(0, \pi/2)$ (see Figure 1). We want to decompose them as $z_k = c_k + d_k$ for $k = 1, 2$. The constraints of JIVE, AJIVE, OnPLS, and COBE translated into space $(\mathcal{L}_0^2, \text{cov})$ do not guarantee $d_1 \perp d_2$, i.e., $\text{corr}(d_1, d_2) = 0$. DISCO-SCA and R.JIVE impose $d_1 \perp d_2$ and $c_j \perp d_k$ for all $j, k = 1, 2$. Restrict $\text{span}(\{z_1, z_2\}) = \text{span}(\{c_k, d_k\}_{k=1}^2)$ as in our (5) to avoid the signal space being represented by a higher dimensional space. Then their orthogonal constraints result in either (i) $d_1 = d_2 = 0$ or (ii) that only one of d_1 and d_2 is a zero constant, since a two-dimensional space does not tolerate three nonzero orthogonal elements. Scenario (i) indicates $z_1 = c_1 \neq z_2 = c_2$ and fails to reveal the distinctive patterns of z_1 and z_2. Scenario (ii) implies unequal distributions of d_1 and d_2, which contradicts the symmetry of z_1 and z_2 about $0.5(z_1 + z_2)$. However, our proposed constraints and developed method will achieve the desirable decomposition shown in Figure 1, where $d_1 \perp d_2$, $c_1 = c_2 = c \propto 0.5(z_1 + z_2)$, and moreover, $\|c\|$ indicates the extent of $1/\theta$ or $\text{corr}(z_1, z_2)$.

Motivated by the toy example above, we introduce a novel method, decomposition-based canonical correlation analysis (D-CCA), which generalizes the classical canonical correlation analysis (CCA; Hotelling, 1936) by further separating common vectors $\{\boldsymbol{c}_k\}_{k=1}^2$ and distinctive vectors $\{\boldsymbol{d}_k\}_{k=1}^2$ between signal vectors $\{\boldsymbol{x}_k\}_{k=1}^2$ subject to constraints (3)-(5). In contrast, classical CCA only seeks the association between two random vectors by sequentially determining the mutually orthogonal pairs of canonical variables that have maximal correlations between the vector spaces respectively spanned by entries of the two random vectors. Another related but different method, the sparse CCA (Chen et al., 2013; Gao et al., 2015, 2017), focuses on the sparse linear combinations of original variables for representing canonical variables with improved in-

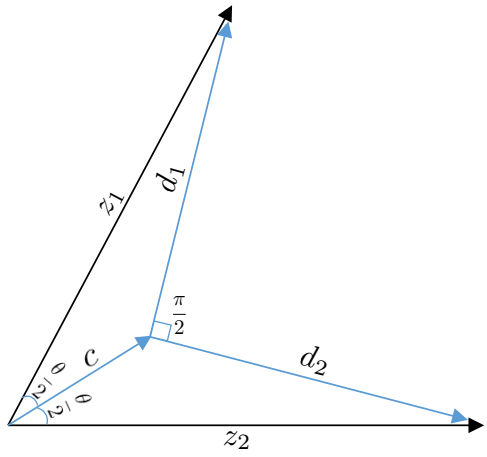


Figure 1: The geometry of D-CCA for two standardized random variables.

terpretability, which is neither required nor pursued by our D-CCA.

The "low-rank plus noise" model $\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k$ for each single k can be naturally formulated by a factor model as $\boldsymbol{y}_k = \mathbf{B}_k \boldsymbol{f}_k + \boldsymbol{e}_k$, where the latent factor $\boldsymbol{f}_k^\top$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$ with $\mathbf{B}_k$ being the coefficient matrix. In factor model analysis (Bai and Ng, 2008), $\boldsymbol{x}_k = \mathbf{B}_k \boldsymbol{f}_k$ is called the "common component", and $\boldsymbol{e}_k$ the "idiosyncratic error". These two terms should not be confused with our considered common vectors $\{\boldsymbol{c}_k\}_{k=1}^2$ and distinctive vectors $\{\boldsymbol{d}_k\}_{k=1}^2$ that are solely based on signals $\{\boldsymbol{x}_k\}_{k=1}^2$ excluding noises $\{\boldsymbol{e}_k\}_{k=1}^2$. For general dynamic factor models (Forni et al., 2000), Hallin and Liška (2011) proposed a joint decomposition method, which divides each dataset into strongly common, weakly common, weakly idiosyncratic, and strongly idiosyncratic components (also see Forni et al. (2017) and Barigozzi et al. (2018)). Applying their method to our considered scenarios with no temporal dependence, and additionally assuming no correlations between signals $\{\boldsymbol{x}_k\}_{k=1}^2$ and noises $\{\boldsymbol{e}_k\}_{k=1}^2$, then for each $\boldsymbol{y}_k$, $\boldsymbol{x}_k$ is the sum of strongly common and weakly common components, $\boldsymbol{e}_k$ is the strongly idiosyncratic component, and no weakly idiosyncratic component exists. One may treat their strongly common and weakly common components as the common vector $\boldsymbol{c}_k$ and the distinctive vector $\boldsymbol{d}_k$, respectively, but the desired orthogonality (4) is still not imposed. Especially when $\mathrm{span}(\boldsymbol{x}_1^\top) \cap \mathrm{span}(\boldsymbol{x}_2^\top) = \{0\}$, $\boldsymbol{x}_k$ is entirely a weakly common component, and thus the orthogonality (4) fails for the toy example shown in Figure 1. See Remark S.1 in the supplementary material for more detailed discussions.

Our major contributions of this paper are as follows. The proposed D-CCA method appro-

priately decomposes each paired canonical variables of signal vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ into a common variable and two orthogonal distinctive variables, and then collects all of them to form the common vector $\boldsymbol{c}_k$ and the distinctive vector $\boldsymbol{d}_k$ for each $\boldsymbol{x}_k$. The common matrix $\mathbf{C}_k$ and the distinctive matrix $\mathbf{D}_k$ are defined with columns as n realizations of $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$, respectively. Three challenging issues that arise in estimating the low-rank matrices defined by D-CCA are high dimensionality, the corruption of signal random vectors by unobserved noises, and the unknown signal covariance and cross-covariance matrices that are needed in CCA. To address these issues, we study the considered "low-rank plus noise" model under the framework of approximate factor models (Wang and Fan, 2017), and develop a novel estimation approach by integrating the S-POET method for spiked covariance matrix estimation (Wang and Fan, 2017) and the construction of principal vectors (Björck and Golub, 1973). Under some mild conditions, we systematically investigate the consistency and convergence rates of the proposed matrix estimators under a high-dimensional setting with $\min(p_1, p_2) > \kappa_0 n$ for a positive constant κ_0.

The rest of this paper is organized as follows. Section 2 introduces the D-CCA method that appropriately defines the common and distinctive matrices from the inner product space $(\mathcal{L}_0^2, \mathrm{cov})$. A soft-thresholding approach is then proposed for estimating the matrices defined by D-CCA. Section 3 is devoted to the theoretical results of the proposed matrix estimators under a high-dimensional setting. The performance of D-CCA and the associated estimation approach is compared to that of the aforementioned state-of-the-art methods through simulations in Section 4 and through the analysis of TCGA breast cancer data in Section 5. Possible future extensions of D-CCA are discussed in Section 6. All technical proofs are provided in the supplementary material.

Here, we introduce some notation. For a real matrix $\mathbf{M} = (M_{ij})_{1\le i\le p, 1\le j\le n}$, the ℓ-th largest singular value and the ℓ-th largest eigenvalue (if $p = n$) are respectively denoted by $\sigma_\ell(\mathbf{M})$ and $\lambda_\ell(\mathbf{M})$, the spectral norm $\|\mathbf{M}\|_2 = \sigma_1(\mathbf{M})$, the Frobenius norm $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^n M_{ij}^2}$, and the matrix $\mathcal{L}^\infty$ norm $\|\mathbf{M}\|_\infty = \max_{1\le i\le p} \sum_{j=1}^n |M_{ij}|$. We use $\mathbf{M}^{[s:t,u:v]}$, $\mathbf{M}^{[s:t,:]}$ and $\mathbf{M}^{[:,u:v]}$ to represent the submatrices $(M_{ij})_{s\le i\le t, u\le j\le v}$, $(M_{ij})_{s\le i\le t, 1\le j\le n}$ and $(M_{ij})_{1\le i\le p, u\le j\le v}$ of the $p\times n$ matrix $\mathbf{M}$, respectively. Denote the Moore-Penrose pseudoinverse of matrix $\mathbf{M}$ by $\mathbf{M}^\dagger$. Define $\mathbf{0}_{p\times n}$ to be the $p\times n$ zero matrix and $\mathbf{I}_{p\times p}$ to be the $p\times p$ identity matrix. Denote $\mathrm{diag}(\mathbf{M}_1, \ldots, \mathbf{M}_m)$ to be a block diagonal matrix with $\mathbf{M}_1, \ldots, \mathbf{M}_m$ as its main diagonal blocks. For signal vec-

tors $\boldsymbol{x}_k$'s, denote $\boldsymbol{\Sigma}_k = \mathrm{cov}(\boldsymbol{x}_k)$, $\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, $r_k = \mathrm{rank}(\boldsymbol{\Sigma}_k)$, $r_{\min} = \min(r_1, r_2)$, $r_{\max} = \max(r_1, r_2)$ and $r_{12} = \mathrm{rank}(\boldsymbol{\Sigma}_{12})$. For a subspace B of a vector space A, denote its orthogonal complement in A by $A \setminus B$. We write $a \propto b$ if a is proportional to b, i.e., $a = \kappa b$ for some constant κ. Throughout the paper, our asymptotic arguments are by default under $n \to \infty$. We reserve $\{c, c_\ell\}, \{\boldsymbol{c}_k\}$ and $\{\mathbf{C}_k\}$ for the common variables, common vectors and common matrices, respectively, and use other notation for constants, e.g., κ_0.

2 The D-CCA Method

Suppose the columns of matrices $\mathbf{Y}_k$, $\mathbf{X}_k$ and $\mathbf{E}_k$ are, respectively, n independent and identically distributed (i.i.d.) copies of mean-zero random vectors $\boldsymbol{y}_k$, $\boldsymbol{x}_k$ and $\boldsymbol{e}_k$ for $k = 1, 2$. We consider the "low-rank plus noise" model for the observable random vector $\boldsymbol{y}_k$ as follows:

$$\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \mathbf{B}_k \boldsymbol{f}_k + \boldsymbol{e}_k, \tag{6}$$

where $\mathbf{B}_k \in \mathbb{R}^{p_k \times r_k}$ is a real deterministic matrix, $\boldsymbol{f}_k \in \mathbb{R}^{r_k}$ is a mean-zero random vector of r_k latent factors such that $\mathrm{cov}(\boldsymbol{f}_k) = \mathbf{I}_{r_k \times r_k}$ and $\mathrm{cov}(\boldsymbol{f}_k, \boldsymbol{e}_k) = \mathbf{0}_{r_k \times p_k}$, and r_k is a fixed number independent of $\{n, p_1, p_2\}$. Write the model in a sample-matrix form by

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{B}_k \mathbf{F}_k + \mathbf{E}_k, \tag{7}$$

where the columns of $\mathbf{F}_k$ are assumed to be i.i.d. copies of $\boldsymbol{f}_k$. We assume that the model given in (6) and (7) is an approximate factor model (Wang and Fan, 2017) that allows for correlations among entries of $\boldsymbol{e}_k$ in contrast with the strict factor model (Ross, 1976) and has $\mathrm{cov}(\boldsymbol{y}_k) = \mathbf{B}_k \mathbf{B}_k^\top + \mathrm{cov}(\boldsymbol{e}_k)$ be a spiked covariance matrix for which the top r_k eigenvalues are significantly larger than the rest (i.e., signals are stronger than noises). Detailed conditions for consistent estimation will be given later in Assumption 1. Although approximate factor models are often used in econometric literature (Chamberlain and Rothschild, 1983; Bai and Ng, 2002; Stock and Watson, 2002; Bai, 2003) with temporal dependence on $\{\mathbf{F}_k^{[:,t]}, \mathbf{E}_k^{[:,t]}\}$ across t's, we assume independence across the n samples as in Wang and Fan (2017) since no temporal dependence is quite natural in our motivating TCGA datasets and considered in the six competing methods mentioned in Section 1.

2.1 Definition of common and distinctive matrices

We define the common and distinctive matrices of two datasets based on the inner product space $(\mathcal{L}_0^2, \mathrm{cov})$. The low-rank structure of $\boldsymbol{x}_k$ in (6) indicates that the dimension of $\mathrm{span}(\boldsymbol{x}_k^\top)$ is r_k.

One natural way to construct the decomposition of $\mathbf{X}_k = \mathbf{C}_k + \mathbf{D}_k$ for $k = 1, 2$ is to decompose the signal vectors as

$$\boldsymbol{x}_k = \sum_{\ell=1}^{r_k} \boldsymbol{\beta}_{k\ell} z_{k\ell} = \boldsymbol{c}_k + \boldsymbol{d}_k \coloneqq \sum_{\ell=1}^{L_{12}} \boldsymbol{\beta}_{k\ell}^{(C)} c_\ell + \sum_{\ell=1}^{L_k} \boldsymbol{\beta}_{k\ell}^{(D)} d_{k\ell}, \tag{8}$$

subject to the constraints (3)-(5) with space dimensions $L_{12} \le r_{\min}$ and $L_k \le r_k$, where $\boldsymbol{\beta}_{k\ell}$, $\boldsymbol{\beta}_{k\ell}^{(C)}$ and $\boldsymbol{\beta}_{k\ell}^{(D)}$ are real deterministic vectors, and random variables $\{z_{k\ell}\}_{\ell=1}^{r_k}$, $\{c_\ell\}_{\ell=1}^{L_{12}}$ and $\{d_{k\ell}\}_{\ell=1}^{L_k}$ are, respectively, the orthogonal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$, $\mathrm{span}(\boldsymbol{c}_1^\top) = \mathrm{span}(\boldsymbol{c}_2^\top)$ and $\mathrm{span}(\boldsymbol{d}_k^\top)$. The desirable contraints (3)-(5) are now equivalent to

$$\begin{cases} \mathrm{span}(\{z_{1\ell}\}_{\ell=1}^{r_1} \cup \{z_{2\ell}\}_{\ell=1}^{r_2}) = \mathrm{span}(\{c_\ell\}_{\ell=1}^{L_{12}} \cup \{d_{1\ell}\}_{\ell=1}^{L_1} \cup \{d_{2\ell}\}_{\ell=1}^{L_2}), \\ d_{su} \perp d_{tv} \quad \text{for} \quad s \neq t \quad \text{or} \quad u \neq v. \end{cases} \tag{9}$$

We call $\{c_\ell\}_{\ell=1}^{L_{12}}$ the common variables of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\{d_{k\ell}\}_{\ell=1}^{L_k}$ the distinctive variables of $\boldsymbol{x}_k$. The columns of common matrix $\mathbf{C}_k$ are defined as the i.i.d. copies of $\boldsymbol{c}_k$, and those of distinctive matrix $\mathbf{D}_k$ are the ones of $\boldsymbol{d}_k$. The space $\mathrm{span}(\{c_\ell\}_{\ell=1}^{L_{12}})$ represents the common structure of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, or datasets $\mathbf{X}_1$ and $\mathbf{X}_2$, and the spaces $\{\mathrm{span}(\{d_{k\ell}\}_{\ell=1}^{L_k})\}_{k=1}^2$ correspond to their distinctive structures.

To achieve a decomposition of form (8), our D-CCA method adopts a two-step optimization strategy given in (10) and (11) below. The first step uses the classical CCA to recursively find the most correlated variables between signal spaces $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^2$ as follows: For $\ell = 1, \ldots, r_{12}$,

$$\begin{aligned} \{z_{1\ell}, z_{2\ell}\} \in \underset{\{z_k\}_{k=1}^2}{\arg\max}\ \mathrm{corr}(z_1, z_2) \quad \text{subject to} \\ \mathrm{var}(z_k) = 1 \text{ and } z_k \in \mathrm{span}(\boldsymbol{x}_k^\top) \setminus \mathrm{span}(\{z_{km}\}_{m=1}^{\ell-1}), \end{aligned} \tag{10}$$

where $\mathrm{span}(\boldsymbol{x}_k^\top) \setminus \mathrm{span}(\{z_{km}\}_{m=1}^0) \coloneqq \mathrm{span}(\boldsymbol{x}_k^\top)$. Variables $\{z_{k\ell}\}_{k=1}^2$ are called the ℓ-th pair of canonical variables, and their correlation is the ℓ-th canonical correlation of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Augment $\{z_{k\ell}\}_{\ell=1}^{r_{12}}$ with any $(r_k - r_{12})$ standardized variables to be $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kr_k})^\top$ such that $\boldsymbol{z}_k^\top$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$. A detailed procedure to obtain a solution of $\{\boldsymbol{z}_k\}_{k=1}^2$ will be

presented later after Theorem 2. An important property of these augmented canonical variables is the bi-orthogonality shown in the following theorem.

Theorem 1 (Bi-orthogonality)**.** *The covariance matrix of* $\boldsymbol{z}_1$ *and* $\boldsymbol{z}_2$ *is*

$$\operatorname{cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \begin{bmatrix} \boldsymbol{\Lambda}_{12} & \mathbf{0}_{r_{12}\times(r_2-r_{12})} \\ \mathbf{0}_{(r_1-r_{12})\times r_{12}} & \mathbf{0}_{(r_1-r_{12})\times(r_2-r_{12})} \end{bmatrix},$$

where $\boldsymbol{\Lambda}_{12}$ *is a* $r_{12}\times r_{12}$ *nonsingular diagonal matrix.*

Theorem 1 implies that all correlations between $\operatorname{span}(\boldsymbol{x}_1^\top)$ and $\operatorname{span}(\boldsymbol{x}_2^\top)$ are confined between their subspaces $\operatorname{span}(\{z_{1\ell}\}_{\ell=1}^{r_{12}})$ and $\operatorname{span}(\{z_{2\ell}\}_{\ell=1}^{r_{12}})$, and moreover, $\operatorname{span}(\{z_{1\ell}, z_{2\ell}\}) \perp \operatorname{span}(\{z_{1m}, z_{2m}\})$ holds for $1 \le \ell \ne m \le r_{12}$. We hence only need to investigate the correlations within each subspace $\operatorname{span}(\{z_{1\ell}, z_{2\ell}\})$ for $1 \le \ell \le r_{12}$. The second step of our D-CCA defines the common variables $\{c_\ell\}_{\ell=1}^{r_{12}}$ by

$$c_\ell \propto \underset{w\in(\mathcal{L}_0^2,\operatorname{cov})}{\arg\max} \left\{\operatorname{corr}^2(z_{1\ell}, w) + \operatorname{corr}^2(z_{2\ell}, w)\right\} \tag{11}$$

with the constraints

$$\begin{cases} z_{k\ell} = c_\ell + d_{k\ell} \ \text{ for } \ k = 1, 2, & (12) \\ \operatorname{corr}(d_{1\ell}, d_{2\ell}) = 0, & (13) \\ \operatorname{var}(c_\ell) \text{ increases as } \rho_\ell := \operatorname{corr}(z_{1\ell}, z_{2\ell}) \text{ increases on } [0, 1]. & (14) \end{cases}$$

Constraints (12) and (13) are actually the special case of (8) and (9) for two standardized random variables. Constraint (14) indicates that c_ℓ explains more variances of $z_{1\ell}$ and $z_{2\ell}$ when their correlation ρ_ℓ increases. Although ρ_ℓ, here referring to the ℓ-th canonical correlation of $\{\boldsymbol{x}_k\}_{k=1}^2$, is always positive for $1 \le \ell \le r_{12}$, we include $\rho_\ell = 0$ to enable (11) as a general optimization problem for any two standardized variables with nonnegative correlation. The unique solution of (11) is given by

$$c_\ell = \left(1 - \sqrt{\frac{1-\rho_\ell}{1+\rho_\ell}}\right) \frac{z_{1\ell} + z_{2\ell}}{2} = \left[1 - \tan\left(\frac{\theta_\ell}{2}\right)\right] \frac{z_{1\ell} + z_{2\ell}}{2}, \tag{15}$$

where $\theta_\ell = \arccos(\rho_\ell)$ is the angle between $z_{1\ell}$ and $z_{2\ell}$ in $(\mathcal{L}_0^2, \operatorname{cov})$. More desirable than constraint (14), it easily follows from (15) that $\operatorname{var}(c_\ell)$ is a continuous and strictly monotonic increasing function for $\rho_\ell \in [0, 1]$. We defer the detailed derivation of (15) to the supplementary

material (Proposition S.2 and its proof). This solution is geometrically illustrated in Figure 1 with ℓ omitted in the subscriptions. Simply let $d_{k\ell} = z_{k\ell}$ for $r_{12}+1 \le \ell \le r_k$. The two-step optimization strategy arrives at the following decomposition of form (8): For $k = 1, 2$,

$$\boldsymbol{x}_k = \sum_{\ell=1}^{r_k} \boldsymbol{\beta}_{k\ell} z_{k\ell} = \boldsymbol{c}_k + \boldsymbol{d}_k \coloneqq \sum_{\ell=1}^{r_{12}} \boldsymbol{\beta}_{k\ell} c_\ell + \sum_{\ell=1}^{r_k} \boldsymbol{\beta}_{k\ell} d_{k\ell}, \tag{16}$$

$$\text{with} \quad \boldsymbol{\beta}_{k\ell} = \operatorname{cov}(\boldsymbol{x}_k, z_{k\ell}). \tag{17}$$

Constraints (3)-(5) or equivalently (9) are satisfied due to the bi-orthogonality in Theorem 1 and the constraints in (12) and (13).

The workflow of D-CCA can be interpreted from the perspective of blind source separation (Comon and Jutten, 2010). Jointly for $k = 1, 2$, D-CCA first uses CCA to recover the input sources $\{z_{k\ell}\}_{\ell=1}^{r_k}$ and the mixing channel $\{\boldsymbol{\beta}_{k\ell}\}_{\ell=1}^{r_k}$ that generate the output signal vector $\boldsymbol{x}_k$. Then by the constrained (11), D-CCA discovers the common components $\{c_\ell\}_{\ell=1}^{r_{12}}$ and the distinctive components $\{d_{k\ell}\}_{\ell=1}^{r_k}, k = 1, 2$ of the two sets of input sources $\{z_{k\ell}\}_{\ell=1}^{r_k}, k = 1, 2$. Finally, D-CCA separately passes $\{c_\ell\}_{\ell=1}^{r_{12}}$ and $\{d_{k\ell}\}_{\ell=1}^{r_k}$ through the mixing channel $\{\boldsymbol{\beta}_{k\ell}\}_{\ell=1}^{r_k}$ to form the common vector $\boldsymbol{c}_k$ and the distinctive vector $\boldsymbol{d}_k$ of each k-th output signal vector $\boldsymbol{x}_k$. Figure 2 illustrates such interpretation of the D-CCA decomposition structure.

The solution to the CCA problem in (10) may not be unique even when ignoring a simultaneous sign change, but all solutions yield the same $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$ as shown in the following theorem.

Theorem 2 (Uniqueness)**.** *All solutions to the problem in* (10) *for canonical variables* $\{z_{1\ell}, z_{2\ell}\}_{\ell=1}^{r_{12}}$ *give the same* $\boldsymbol{c}_k$ *and* $\boldsymbol{d}_k$ *defined in* (16)*.*

We now present a procedure to obtain the augmented canonical variables $\{\boldsymbol{z}_1, \boldsymbol{z}_2\}$. For $k = 1, 2$, let a singular value decomposition (SVD) of $\boldsymbol{\Sigma}_k$ be $\boldsymbol{\Sigma}_k = \mathbf{V}_k \boldsymbol{\Lambda}_k \mathbf{V}_k^\top$, where $\boldsymbol{\Lambda}_k = \operatorname{diag}(\sigma_1(\boldsymbol{\Sigma}_k), \ldots, \sigma_{r_k}(\boldsymbol{\Sigma}_k))$ and $\mathbf{V}_k$ is a $p_k \times r_k$ matrix with orthonormal columns. Let $\boldsymbol{z}_k^* = \boldsymbol{\Lambda}_k^{-1/2} \mathbf{V}_k^\top \boldsymbol{x}_k$, then we have $\operatorname{cov}(\boldsymbol{z}_k^*) = \mathbf{I}_{r_k \times r_k}$. Define

$$\boldsymbol{\Theta} = \operatorname{cov}(\boldsymbol{z}_1^*, \boldsymbol{z}_2^*) = \boldsymbol{\Lambda}_1^{-1/2} \mathbf{V}_1^\top \boldsymbol{\Sigma}_{12} \mathbf{V}_2 \boldsymbol{\Lambda}_2^{-1/2}.$$

The rank of $\boldsymbol{\Theta}$ is also r_{12}. Denote a full SVD of $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta} = \mathbf{U}_{\theta 1} \boldsymbol{\Lambda}_\theta \mathbf{U}_{\theta 2}^\top$, where $\mathbf{U}_{\theta 1}$ and $\mathbf{U}_{\theta 2}$ are two orthogonal matrices, and $\boldsymbol{\Lambda}_\theta$ is a $r_1 \times r_2$ rectangular diagonal matrix for which the main

diagonal is $(\sigma_1(\boldsymbol{\Theta}), \ldots, \sigma_{r_{12}}(\boldsymbol{\Theta}), \mathbf{0}_{1\times(r_{\min}-r_{12})})$. We then define

$$\boldsymbol{z}_k = \mathbf{U}_{\theta k}^\top \boldsymbol{z}_k^* = \boldsymbol{\Gamma}_k^\top \boldsymbol{x}_k \qquad \text{with} \qquad \boldsymbol{\Gamma}_k \coloneqq \mathbf{V}_k \boldsymbol{\Lambda}_k^{-1/2} \mathbf{U}_{\theta k}, \tag{18}$$

which satisfies $\mathrm{cov}(\boldsymbol{z}_k) = \mathbf{I}_{r_k \times r_k}$ and $\mathrm{corr}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{\Lambda}_\theta$. Note that $\sigma_\ell(\boldsymbol{\Theta}) = \rho_\ell$ for $\ell \le r_{12}$ are the canonical correlations between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Now look back to $\boldsymbol{c}_k = \sum_{\ell=1}^{r_{12}} \boldsymbol{\beta}_{k\ell} c_\ell$ that is defined in (16). Plugging (17) and (15) for $\boldsymbol{\beta}_{k\ell}$ and c_ℓ in the formula together with $\{\boldsymbol{z}_j\}_{j=1}^2$ given in (18), we obtain

$$\boldsymbol{c}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{[1:r_{12}]}) \mathbf{A}_C \sum_{j=1}^{2} \boldsymbol{z}_j^{[1:r_{12}]}, \tag{19}$$

where $\mathbf{A}_C = \mathrm{diag}(a_1, \ldots, a_{r_{12}})$ and $a_\ell = \frac{1}{2}\left[1 - \left(\frac{1-\sigma_\ell(\boldsymbol{\Theta})}{1+\sigma_\ell(\boldsymbol{\Theta})}\right)^{1/2}\right]$ for $\ell \le r_{12}$. Replacing random vector $\boldsymbol{z}_k = \boldsymbol{\Gamma}_k^\top \boldsymbol{x}_k$ by its sample matrix $\mathbf{Z}_k \coloneqq \boldsymbol{\Gamma}_k^\top \mathbf{X}_k$ in the rightmost of (19) yields

$$\mathbf{C}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{[1:r_{12}]}) \mathbf{A}_C \sum_{j=1}^{2} \mathbf{Z}_j^{[1:r_{12},:]}. \tag{20}$$

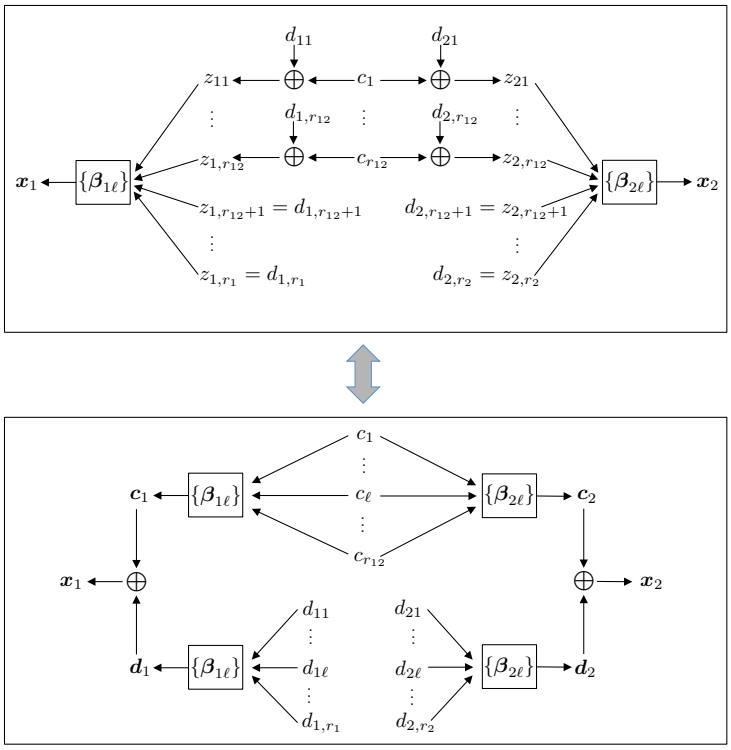


Figure 2: The decomposition structure of D-CCA.

This equation is useful to our design of estimators for $\mathbf{C}_k$ and $\mathbf{D}_k{=}\mathbf{X}_k{-}\mathbf{C}_k$ in the next subsection.

2.2 Estimation of D-CCA matrices

In this subsection, we discuss the estimation of the matrices defined by D-CCA under model (1) for two high-dimensional datasets. For simplicity, we write the proposed estimators with true ranks r_1, r_2 and r_{12}. In practice, we can replace those unknown true ranks by the estimated ranks given in Subsection 2.3 with a theoretical guarantee provided in Section 3.

Recall that $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k$ with $k = 1, 2$. Our first task is to obtain a good initial estimator, denoted by $\widetilde{\mathbf{X}}_k$, of $\mathbf{X}_k$. Under the approximate factor model given in (6) and (7), our construction of $\widetilde{\mathbf{X}}_k$ is inspired by the S-POET method (Wang and Fan, 2017) for spiked covariance matrix estimation. Let the full SVD of $\mathbf{Y}_k$ be

$$\mathbf{Y}_k = \mathbf{U}_{k1}\mathbf{\Lambda}_{y_k}\mathbf{U}_{k2}^\top, \tag{21}$$

where $\mathbf{U}_{k1}$ and $\mathbf{U}_{k2}$ are two orthogonal matrices and $\mathbf{\Lambda}_{y_k}$ is a rectangular diagonal matrix with the singular values in decreasing order on its main diagonal. The matrix $\widetilde{\mathbf{X}}_k$ is then obtained via soft-thresholding the singular values of $\mathbf{Y}_k$ by

$$\widetilde{\mathbf{X}}_k = \mathbf{U}_{k1}^{[:,1:r_k]}\,\mathrm{diag}(\widehat{\sigma}_1^S(\mathbf{Y}_k), \ldots, \widehat{\sigma}_{r_k}^S(\mathbf{Y}_k))(\mathbf{U}_{k2}^{[:,1:r_k]})^\top, \tag{22}$$

with $\widehat{\sigma}_\ell^S(\mathbf{Y}_k) = \sqrt{\max\{\sigma_\ell^2(\mathbf{Y}_k) - \tau_k p_k, 0\}}$ and $\tau_k = \sum_{\ell=r_k+1}^{p_k}\sigma_\ell^2(\mathbf{Y}_k)/(np_k - nr_k - p_k r_k)$. Let $\widetilde{r}_k = \mathrm{rank}(\widetilde{\mathbf{X}}_k)$. Under Assumption 1 that will be given later, it can be shown that $\widetilde{r}_k = r_k$ with probability tending to 1 (see the proof of Theorem 3).

We next use $\widetilde{\mathbf{X}}_k$ to develop estimators for $\mathbf{C}_k$ in (20) and $\mathbf{D}_k{=}\mathbf{X}_k{-}\mathbf{C}_k$. Define the estimators of $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_{12}$ as $\widehat{\mathbf{\Sigma}}_k = n^{-1}\widetilde{\mathbf{X}}_k\widetilde{\mathbf{X}}_k^\top$ and $\widehat{\mathbf{\Sigma}}_{12} = n^{-1}\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_2^\top$, respectively. Then, based on $\widehat{\mathbf{\Sigma}}_k$ and $\widehat{\mathbf{\Sigma}}_{12}$, we obtain estimators $\widehat{\mathbf{V}}_k, \widehat{\mathbf{\Lambda}}_k, \widehat{\mathbf{U}}_{\theta k}{=}\,\mathrm{diag}(\widehat{\mathbf{U}}_{\theta k}^{[1:\widetilde{r}_k,1:\widetilde{r}_k]}, \mathbf{I}_{(r_k-\widetilde{r}_k)\times(r_k-\widetilde{r}_k)})$ and $\widehat{\mathbf{\Lambda}}_\theta$ in the same way as their true counterparts $\mathbf{V}_k, \mathbf{\Lambda}_k, \mathbf{U}_{\theta k}$ and $\mathbf{\Lambda}_\theta$ with a $r_1{\times}r_2$ matrix $\widehat{\mathbf{\Theta}} := (\widehat{\mathbf{\Lambda}}_1^\dagger)^{1/2}\widehat{\mathbf{V}}_1^\top\widehat{\mathbf{\Sigma}}_{12}\widehat{\mathbf{V}}_2(\widehat{\mathbf{\Lambda}}_2^\dagger)^{1/2}$. Define $\widehat{\mathbf{Z}}_k^* = (\widehat{\mathbf{\Lambda}}_k^\dagger)^{1/2}\widehat{\mathbf{V}}_k^\top\widetilde{\mathbf{X}}_k$ and $\widehat{\mathbf{Z}}_k = \widehat{\mathbf{U}}_{\theta k}^\top\widehat{\mathbf{Z}}_k^*$. We have

$$n^{-1}\widehat{\mathbf{Z}}_k^*(\widehat{\mathbf{Z}}_k^*)^\top = n^{-1}\widehat{\mathbf{Z}}_k(\widehat{\mathbf{Z}}_k)^\top = \mathrm{diag}(\mathbf{I}_{\widetilde{r}_k\times\widetilde{r}_k}, \mathbf{0}_{(r_k-\widetilde{r}_k)\times(r_k-\widetilde{r}_k)}),$$

$$\widehat{\mathbf{\Theta}} = \widehat{\mathbf{U}}_{\theta 1}\widehat{\mathbf{\Lambda}}_\theta\widehat{\mathbf{U}}_{\theta 2}^\top = n^{-1}\widehat{\mathbf{Z}}_1^*(\widehat{\mathbf{Z}}_2^*)^\top \quad \text{and} \quad n^{-1}\widehat{\mathbf{Z}}_1(\widehat{\mathbf{Z}}_2)^\top = \widehat{\mathbf{\Lambda}}_\theta.$$

From Theorem 1 in Björck and Golub (1973), it follows that $n^{-1/2}\widehat{\mathbf{Z}}_1^{[\ell,:]}$ and $n^{-1/2}\widehat{\mathbf{Z}}_2^{[\ell,:]}$ for $\ell \leq r_{\min}$ are the principal vectors of the row spaces of $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$, and moreover, $\sigma_\ell(\widehat{\mathbf{\Theta}}) \leq 1$. Let $\widehat{\mathbf{A}}_C^{(r)} = \mathrm{diag}(\widehat{a}_1, \ldots, \widehat{a}_r)$ with $\widehat{a}_\ell = \frac{1}{2}\Big[1 - \big(\frac{1-\sigma_\ell(\widehat{\mathbf{\Theta}})}{1+\sigma_\ell(\widehat{\mathbf{\Theta}})}\big)^{1/2}\Big]$ for $\ell \leq \widetilde{r}_{12} := \mathrm{rank}(\widehat{\mathbf{\Theta}})$ and otherwise $\widehat{a}_\ell = 0$. Define estimators of $\mathbf{C}_k$, $\mathbf{D}_k$ and $\mathbf{X}_k$ by

$$\widehat{\mathbf{C}}_k = n^{-1}\widetilde{\mathbf{X}}_k(\widehat{\mathbf{Z}}_k^{[1:r_{12},:]})^\top \widehat{\mathbf{A}}_C^{(r_{12})} \sum_{j=1}^{2} \widehat{\mathbf{Z}}_j^{[1:r_{12},:]}, \tag{23}$$

$$\widehat{\mathbf{D}}_k = \widetilde{\mathbf{X}}_k - n^{-1}\widetilde{\mathbf{X}}_k(\widehat{\mathbf{Z}}_k^{[1:\widetilde{r}_{12},:]})^\top \widehat{\mathbf{A}}_C^{(\widetilde{r}_{12})} \sum_{j=1}^{2} \widehat{\mathbf{Z}}_j^{[1:\widetilde{r}_{12},:]} \tag{24}$$

and

$$\widehat{\mathbf{X}}_k = \widehat{\mathbf{C}}_k + \widehat{\mathbf{D}}_k. \tag{25}$$

Here, we substitute $\widehat{\mathbf{X}}_k$ for $\widetilde{\mathbf{X}}_k$ as the estimator of $\mathbf{X}_k$. The latter can be written as

$$\widetilde{\mathbf{X}}_k = \widehat{\mathbf{C}}_k^{(\widetilde{r}_{12})} + \widehat{\mathbf{D}}_k \tag{26}$$

with

$$\widehat{\mathbf{C}}_k^{(r)} := n^{-1}\widetilde{\mathbf{X}}_k(\widehat{\mathbf{Z}}_k^{[1:r,:]})^\top \widehat{\mathbf{A}}_C^{(r)} \sum_{j=1}^{2} \widehat{\mathbf{Z}}_j^{[1:r,:]}. \tag{27}$$

Note that $\widehat{\mathbf{C}}_k := \widehat{\mathbf{C}}_k^{(r_{12})}$. When $r_{12} \geq \widetilde{r}_{12}$, we have $\widehat{\mathbf{C}}_k = \widehat{\mathbf{C}}_k^{(\widetilde{r}_{12})}$. But when $r_{12} < \widetilde{r}_{12}$, $\widehat{\mathbf{C}}_k^{(\widetilde{r}_{12})}$ redundantly keeps the nonzero approximated samples of the zero common variable of $z_{1\ell}$ and $z_{2\ell}$ for $r_{12} < \ell \leq \widetilde{r}_{12}$.

Similar to the decomposition of $\{\boldsymbol{x}_k\}_{k=1}^2$ given in (16) that is built on the inner product space $(\mathcal{L}_0^2, \mathrm{cov})$, the decomposition of $\{\widetilde{\mathbf{X}}_k\}_{k=1}^2$ in (26) is constructed by an analogy of (12) and (15) on the $\mathbb{R}^n$ space with the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}/n$ for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. We thus have the appealing property $\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}_{p_1 \times p_2}$, which corresponds to the orthogonal relationship between the distinctive structures given in (4).

Throughout our estimation construction, the key idea is to develop a good estimator of $\{\mathbf{X}_k, \mathbf{Z}_k\}_{k=1}^2$. Thus, the S-POET method (Wang and Fan, 2017) may be replaced by any other good approach, but with possibly different assumptions. For example, given the cleaned signal data $\mathbf{X}_k$'s, Chen et al. (2013) and Gao et al. (2015, 2017) showed that sparse CCA algorithms can consistently estimate the canonical coefficient matrix $\mathbf{\Gamma}_k$ for $\mathbf{Z}_k = \mathbf{\Gamma}_k^\top \mathbf{X}_k$ by imposing certain sparsity on $\mathbf{\Gamma}_k$'s and that all eigenvalues of $\mathrm{cov}(\boldsymbol{x}_k)$ are bounded from above and below by positive con-

stants. These two conditions are not assumed for our proposed method. In particular, their bounded eigenvalue condition contradicts our low-rank structure of signal $\boldsymbol{x}_k$ that introduces the spiked covariance matrix $\mathrm{cov}(\boldsymbol{y}_k)$. The sparse CCA algorithms need the cleaned signal data $\mathbf{X}_k$'s available beforehand. Alternatively, they may be directly applicable to the observable data $\mathbf{Y}_k$'s by assuming zero $\mathbf{E}_k$'s, if the bounded eigenvalue condition holds for $\mathrm{cov}(\boldsymbol{y}_k)$. For the TCGA datasets in our real-data application, the scree plots given later in Figure 6 favorably suggest our spiked eigenvalue assumption. Moreover, the approximate factor model with spiked covariance structure has been widely used in various fields such as signal processing (Nadakuditi and Silverstein, 2010) and machine learning (Huang, 2017), and fits the low-rank plus noise structure considered in the six competing methods mentioned in Section 1. Our paper hence focuses on this spiked covariance model and leaves the extension to sparse CCA models for future research.

2.3 Rank selection

In practice, matrix ranks r_1, r_2 and r_{12} are usually unknown and need to be determined. There is a rich literature on determining $r_k, k \in \{1, 2\}$, which is the number of latent factors for the high-dimensional approximate factor model. Examples of consistent estimators include but are not limited to Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). Several heuristic approaches for selecting r_{12}, the number of nonzero canonical correlations for the high-dimensional CCA, have been proposed by Song et al. (2016). In this paper, we apply the edge distribution (ED) method of Onatski (2010) to determine r_k for $k = 1, 2$ by

$$\widehat{r}_k = \max\{\ell \leq T_k : \widehat{\lambda}_{k\ell} - \widehat{\lambda}_{k,\ell+1} \geq \delta\}, \tag{28}$$

where $\widehat{\lambda}_{k\ell}$ is the ℓ-th eigenvalue of $\mathbf{Y}_k\mathbf{Y}_k^\top/n$. The upper bound is chosen as $T_k = \min\left(\#\{i|\widehat{\lambda}_{ki} \geq \frac{1}{m_k}\sum_{\ell=1}^{m_k}\widehat{\lambda}_{k\ell}\}, m_k/10\right)$ with $m_k = \min(n, p_k)$, which is recommended by Ahn and Horenstein (2013), and parameter δ is calibrated as in Section IV of Onatski (2010). It is believed that $r_{12} > 0$ if two variables from different cleaned datasets have a significant nonzero correlation detected by, e.g., the normal approximation test of DiCiccio and Romano (2017). Otherwise, it is unnecessary to conduct the proposed matrix decomposition. We select the nonzero r_{12} by using the minimum

description length information-theoretic criterion (MDL-IC) proposed by Song et al. (2016):

$$\widehat{r}_{12} = \underset{r\in[1,\min(\widehat{r}_1,\widehat{r}_2)]}{\arg\min} \left\{ n\sum_{\ell=1}^{r}\log(1-s_\ell^2) + r(r_1+r_2-r)\log(n) \right\}, \tag{29}$$

where s_ℓ is the ℓ-th singular value of $(\mathbf{U}_{12}^{[:,1:\widehat{r}_1]})^\top \mathbf{U}_{22}^{[:,1:\widehat{r}_2]}$ with $\mathbf{U}_{12}$ and $\mathbf{U}_{22}$ defined in (21). The ranks r_1, r_2, and r_{12} determined by (28) and (29) perform well in our numerical studies.

3 Theoretical Properties of D-CCA Estimators

In this section, we establish asymptotic results for the high-dimensional D-CCA matrix estimators proposed in Subsection 2.2.

Assumption 1. *We assume the following conditions for model given in (6) and (7).*

(I) Let $\lambda_{k1} > \cdots > \lambda_{k,r_k} > \lambda_{k,r_k+1} \geq \cdots \geq \lambda_{k,p_k} > 0$ *be the eigenvalues of* $\mathrm{cov}(\boldsymbol{y}_k)$. *There exist positive constants* κ_1, κ_2 *and* δ_0 *such that* $\kappa_1 \leq \lambda_{k\ell} \leq \kappa_2$ *for* $\ell > r_k$ *and* $\min_{\ell\leq r_k}(\lambda_{k\ell} - \lambda_{k,\ell+1})/\lambda_{k\ell} \geq \delta_0$.

(II) Assume $p_k > \kappa_0 n$ *with a constant* $\kappa_0 > 0$. *When* $n \to \infty$, *assume* $\lambda_{k,r_k} \to \infty$, $p_k/(n\lambda_{k\ell})$ *is upper bounded for* $\ell \leq r_k$, $\lambda_{k1}/\lambda_{k,r_k}$ *is bounded from above and below, and* $\sqrt{p_k}(\log n)^{1/\gamma_{k2}} = o(\lambda_{r_k})$ *with* γ_{k2} *given in (V) below.*

(III) The columns of $\mathbf{Z}_k^{(y)} := (\mathbf{\Lambda}_k^{(y)})^{-1/2}(\mathbf{V}_k^{(y)})^\top \mathbf{Y}_k$ *are i.i.d. copies of* $\boldsymbol{z}_k^{(y)} := (\mathbf{\Lambda}_k^{(y)})^{-1/2}(\mathbf{V}_k^{(y)})^\top \boldsymbol{y}_k$, *where* $\mathbf{V}_k^{(y)}\mathbf{\Lambda}_k^{(y)}(\mathbf{V}_k^{(y)})^\top$ *is the full SVD of* $\mathrm{cov}(\boldsymbol{y}_k)$ *with* $\mathbf{\Lambda}_k^{(y)} = \mathrm{diag}(\lambda_{k1},\ldots,\lambda_{k,p_k})$. *The entries of* $\boldsymbol{z}_k^{(y)}$, $z_{k1}^{(y)},\ldots,z_{k,p_k}^{(y)}$ *are independent with* $\mathbb{E}(z_{ki}^{(y)}) = 0$, $\mathrm{var}(z_{ki}^{(y)}) = 1$, *and the sub-Gaussian norm* $\sup_{q\geq 1} q^{-1/2}(\mathbb{E}|z_{ki}^{(y)}|^q)^{1/q} \leq K$ *with a constant* $K > 0$ *for all* $i \leq p_k$.

(IV) The matrix $\mathbf{B}_k^\top \mathbf{B}_k$ *is a diagonal matrix, and* $|\mathbf{B}_k^{[i,\ell]}| \leq M\sqrt{\lambda_{k\ell}/p_k}$ *with a constant* $M > 0$ *holds for all* $i \leq p_k$ *and* $\ell \leq r_k$.

(V) Denote $\boldsymbol{e}_k = (e_{k1},\ldots,e_{k,p_k})^\top$ *and* $\boldsymbol{f}_k = (f_{k1},\ldots,f_{k,r_k})^\top$. *Assume* $\|\mathrm{cov}(\boldsymbol{e}_k)\|_\infty < s_0$ *with a constant* $s_0 > 0$. *For all* $i \leq p_k$ *and* $\ell \leq r_k$, *there exist positive constants* $\gamma_{k1}, \gamma_{k2}, b_{k1}$ *and* b_{k2} *such that for* $t > 0$, $\mathbb{P}(|e_{ki}| > t) \leq \exp(-(t/b_{k1})^{\gamma_{k1}})$ *and* $\mathbb{P}(|f_{k\ell}| > t) \leq \exp(-(t/b_{k2})^{\gamma_{k2}})$.

Assumption 1 follows assumptions 2.1-2.3 and 4.1-4.2 of Wang and Fan (2017) which guarantee desirable performance of the initial signal estimators $\widetilde{\mathbf{X}}_k$'s defined in (22). The diverging leading eigenvalues of $\mathrm{cov}(\boldsymbol{y}_k)$ assumed in conditions (I) and (II), together with the approximate sparsity constraint $\|\mathrm{cov}(\boldsymbol{e}_k)\|_\infty < s_0$ in condition (V), indicate the necessity of sufficiently strong signals for soft-thresholding. Although Wang and Fan (2017) considered $p > n$, it is not difficult to relax it to $p_k > \kappa_0 n$, as given in our condition (II). A random variable is said to be sub-Gaussian if its sub-Gaussian norm is bounded (Vershynin, 2012). Condition (III) imposes the sub-Gaussianity on all entries of $\boldsymbol{z}_k^{(y)}$ with a uniform bound. Simply letting $\boldsymbol{f}_k = \boldsymbol{z}_k^*$ can lead to a diagonal matrix $\mathbf{B}_k^\top \mathbf{B}_k$ that is required by condition (IV). In condition (V), the approximately sparse constraint is imposed on $\mathrm{cov}(\boldsymbol{e}_k)$ rather than $\mathbf{E}_k$. See Wang and Fan (2017) and also Fan et al. (2013) for more detailed discussions of the above assumption.

We consider the relative errors of the proposed matrix estimators in the spectral norm and also in the Frobenius norm. For convenience, we use $\|\cdot\|_{(\cdot)}$ as general notation for one of these two matrix norms. Define $\alpha_{\mathbf{C}_k,(\cdot)} = \|\mathbf{C}_k\|_{(\cdot)}/\|\mathbf{X}_k\|_{(\cdot)}$ and $\alpha_{\mathbf{D}_k,(\cdot)} = \|\mathbf{D}_k\|_{(\cdot)}/\|\mathbf{X}_k\|_{(\cdot)}$.

Theorem 3. *For $k = 1, 2$, assume $\widehat{\mathbf{C}}_k$, $\widehat{\mathbf{D}}_k$, $\widehat{\mathbf{X}}_k$ and $\widehat{\boldsymbol{\Theta}}$ defined in Subsection 2.2 are constructed with true r_k and r_{12}. Suppose that $r_{12} \geq 1$ and Assumption 1 hold. Define $\Delta = \delta_\theta^{1/2}$ and*

$$\delta_\theta = \min\left\{\frac{1}{\sqrt{n}} + \sum_{k=1}^{2}\sqrt{\frac{p_k \log p_k}{n\lambda_1(\boldsymbol{\Sigma}_k)}}, 1\right\}.$$

Then, we have the following relative error bounds of the matrix estimators

$$\frac{\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_{(\cdot)}}{\|\mathbf{C}_k\|_{(\cdot)}} = O_P\left(\frac{\Delta}{\alpha_{\mathbf{C}_k,(\cdot)}}\right), \quad \frac{\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\|_{(\cdot)}}{\|\mathbf{D}_k\|_{(\cdot)}} = O_P\left(\frac{\Delta}{\alpha_{\mathbf{D}_k,(\cdot)}}\right), \quad \frac{\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_{(\cdot)}}{\|\mathbf{X}_k\|_{(\cdot)}} = O_P\left(\Delta\right),$$

and the error bound of canonical correlation estimators

$$\max_{1\leq \ell \leq r_{\min}} |\sigma_\ell(\widehat{\boldsymbol{\Theta}}) - \sigma_\ell(\boldsymbol{\Theta})| = O_P(\delta_\theta).$$

Provided that matrix ranks r_1, r_2 and r_{12} are correctly selected, Theorem 3 shows the consistency of the proposed matrix estimators in the relative errors that are the norms of estimation errors divided by the norms of true matrices, with associated convergence rates. The ratios $\alpha_{\mathbf{C}_k,(\cdot)}$ and $\alpha_{\mathbf{D}_k,(\cdot)}$ in the convergence rates of $\widehat{\mathbf{C}}_k$ and $\widehat{\mathbf{D}}_k$ can be removed if the relative errors are instead scaled by the norms of the signal matrices.

Although the ED estimators of r_1 and r_2 given in (28) are consistent under some mild conditions (Onatski, 2010), the consistency of the MDL-IC estimator in (29) for r_{12} is still unclear. However, the following corollary indicates the robustness of our proposed matrix estimators given in (23) and (25) when r_{12} is misspecified but r_1 and r_2 are appropriately selected.

Corollary 1. *For $k = 1, 2$, assume $\widehat{\mathbf{C}}_k^{(r)}$, $\widehat{\mathbf{D}}_k$ and $\widehat{\mathbf{\Theta}}$ defined in Subsection 2.2 are constructed with the unknown r_k replaced by an estimator $\breve{r}_k$ satisfying $\breve{r}_k \xrightarrow{P} r_k$. Define $\widehat{\mathbf{X}}_k^{(r)} = \widehat{\mathbf{C}}_k^{(r)} + \widehat{\mathbf{D}}_k$ with $\min(r_{12}, \widetilde{r}_{12}) \le r \le r_{\min}$, and $\sigma_\ell(\widehat{\mathbf{\Theta}}) = 0$ for $\ell > \min(\breve{r}_1, \breve{r}_2)$. Suppose that $r_{12} \ge 1$ and Assumption 1 hold. Then, with Δ and δ_θ defined in Theorem 3, we have*

$$\frac{\|\widehat{\mathbf{C}}_k^{(r)} - \mathbf{C}_k\|_{(\cdot)}}{\|\mathbf{C}_k\|_{(\cdot)}} = O_P\left(\frac{\Delta}{\alpha_{\mathbf{C}_k,(\cdot)}}\right), \quad \frac{\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\|_{(\cdot)}}{\|\mathbf{D}_k\|_{(\cdot)}} = O_P\left(\frac{\Delta}{\alpha_{\mathbf{D}_k,(\cdot)}}\right),$$

$$\frac{\|\widehat{\mathbf{X}}_k^{(r)} - \mathbf{X}_k\|_{(\cdot)}}{\|\mathbf{X}_k\|_{(\cdot)}} = O_P\left(\Delta\right), \text{ and } \max_{1\le\ell\le r_{\min}} |\sigma_\ell(\widehat{\mathbf{\Theta}}) - \sigma_\ell(\mathbf{\Theta})| = O_P(\delta_\theta).$$

Corollary 1 provides an acceptable range, $[\min(r_{12}, \widetilde{r}_{12}), r_{\min}]$, for the choice of r_{12} when r_1 and r_2 are consistently estimated, which can theoretically lead to the same convergence rates (up to a constant factor) as those in Theorem 3. Note that the distinctive matrices $\widehat{\mathbf{D}}_k$'s are independent of r_{12}.

4 Simulation Studies

We consider the following three simulation setups to evaluate the finite sample performance of the proposed D-CCA estimators comparing with the six competing methods mentioned in Section 1 and also the decomposition of Hallin and Liška (2011) (denoted as GDFM).

- Setup 1: Let $\boldsymbol{x}_1 \overset{d}{=} \boldsymbol{x}_2$, with $r_1 = 3$, $r_{12} = 1$, and $\lambda_\ell(\mathbf{\Sigma}_1) = 500 - 200(\ell - 1)$ for $\ell \le 3$. Set $z_{k1}, z_{k2}, z_{k3} \overset{i.i.d.}{\sim} N(0, 1)$ for each $k = 1, 2$. Randomly generate $\mathbf{V}_1$ with orthonormal columns, which is the same for all replications. Let $\boldsymbol{x}_k = \mathbf{V}_1 \mathbf{\Lambda}_1^{1/2} \boldsymbol{z}_k$. Generate $e_{ki}, k \le 2, i \le p_k \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$ that are independent of $\{\boldsymbol{x}_k\}_{k=1}^2$. Vary dimension p_1 from 100 to 1,500, the first canonical angle $\theta_1 = \arccos(\rho_1)$ from 0° to 75° with $\rho_1 = \text{corr}(z_{11}, z_{21})$, and the noise variance σ_e^2 from 0.01 to 16.

- Setup 2: Use the same settings for $\boldsymbol{x}_1$ and $\{\boldsymbol{e}_k\}_{k=1}^2$ as in Setup 1. For $\boldsymbol{x}_2$, fix $p_2 = 300$, and set $r_2 = 5$ and $\lambda_\ell(\mathbf{\Sigma}_2) = 500 - 100(\ell - 1)$ for $\ell \le 5$. Simulate $\boldsymbol{x}_2 = \mathbf{V}_2 \mathbf{\Lambda}_2^{1/2} \boldsymbol{z}_2$ with

$z_{21}, \ldots, z_{25} \overset{i.i.d.}{\sim} N(0,1)$ and a randomly generated $\mathbf{V}_2$ that is the same for all replications. Let $r_{12} = 1$. Vary p_1, θ_1 and σ_e^2 according to Setup 1.

- Setup 3 is for visual purposes: Fix $p_1 = 3p_2 = 900, \theta_1 = 45°$, and $\sigma_e^2 = 1$. Generate two independent variables v_1 and v_2 such that $v_1 \sim \text{Unif}(\{0, \pm 1/\sqrt{2}, \pm\sqrt{2}\})$ and $v_2 \sim N(0,1)$. Let $z_{11} = [v_1 + v_2 \tan(\theta_1/2)]/\sqrt{1+\tan^2(\theta_1/2)}$ and $z_{21} = [v_1 - v_2 \tan(\theta_1/2)]/\sqrt{1+\tan^2(\theta_1/2)}$. Set $\mathbf{V}_k^{[:,1]} = \frac{1}{\sqrt{p_k}}(1,1,\ldots,1)^\top$ and randomly generate $\mathbf{V}_k^{[:,2:r_k]}$ for $k = 1, 2$. The other settings are the same as those in Setup 2.

We fixed the sample size $n = 300$ and conducted 1,000 replications for Setups 1 and 2. Setup 3 is only used for the purpose of visually comparing D-CCA with the seven other methods. Setup 3 is similar to Setup 2, but it has the common variable of the first pair of canonical variables following a discrete uniform distribution instead of a Gaussian distribution. We ran a single replication of Setup 3 for the visual comparison in Figure 5. To determine the ranks r_1, r_2, and r_{12}, we respectively used the ED method given in (28) and the MDL-IC method in (29). Additional simulations with AR(1) matrices for $\{\text{cov}(\boldsymbol{e}_k)\}_{k=1}^2$ are given in the supplementary material (Section S.2).

The results obtained by D-CCA for Setups 1 and 2 are summarized in Figures 3 and 4 and Table 1. The first rows of the two figures show the average relative errors (AREs) for $\theta_1 = 45°, \sigma_e^2 = 1$ and varying p_1; the second rows are for $p_1 = 900, \sigma_e^2 = 1$ and varying θ_1; and the third rows are for $p_1 = 900, \theta_1 = 45°$ and varying σ_e^2. Both figures reveal that the curves based on the estimated ranks almost overlap with those based on the true ranks. The ranks are selected with very high accuracy (>99.7%).

Consider Figure 3 of Setup 1 as an example. We have nearly identical plots for the two datasets that are generated from the same distribution. From the first row, where all considered cases have almost the same set of average ratios $\{\alpha_{\mathbf{C}_k,(\cdot)}, \alpha_{\mathbf{D}_k,(\cdot)}\}$, all the AREs become bigger as the dimension p_1 increases. For the second row, the increasing canonical angle θ_1 results in a change in the average ratios $\alpha_{\mathbf{C}_k,2}$ from 0.997 down to 0.18 and in $\alpha_{\mathbf{C}_k,F}$ from 0.74 down to 0.14; $\alpha_{\mathbf{D}_k,2}$ is stable around 0.78 for the first 5 values of θ_1 and then increases to 0.87 at $\theta_1 = 75°$; and $\alpha_{\mathbf{D}_k,F}$ changes from 0.67 to 0.93. Meanwhile, this leads to increasing AREs of $\widehat{\mathbf{C}}_k$ and decreasing AREs of $\widehat{\mathbf{D}}_k$, but does not affect the AREs of $\widehat{\mathbf{X}}_k$. The third row shows that all the AREs increase as the noise variance σ_e^2 becomes bigger. Note that increasing σ_e^2 is equivalent to

decreasing the eigenvalues of $\mathbf{\Sigma}_k$ by scaling σ_e^2 to 1. These results agree with the influence of p_1, α and $\lambda_1(\mathbf{\Sigma}_k)$ on the convergence rates given in Theorem 3.

For Setup 2, with similar arguments, we find a similar pattern of estimation performance for D-CCA, as shown in the second and third rows and the plots of the first dataset in the first row of Figure 4. For the first row of Figure 4, the considered cases of the second dataset have a fixed dimension p_2 and stable ratios $\{\alpha_{\mathbf{C}_2,(\cdot)}, \alpha_{\mathbf{D}_2,(\cdot)}\}$. The corresponding AREs are still acceptable and interestingly are not much impacted by the change in the dimension p_1 of the first dataset. From Table 1, we see that the estimated canonical angles and correlations perform well for Setups 1 and 2 even in the presence of strong noise levels.

The comparison of D-CCA and the seven other methods is shown in Tables 2 and 3, and Figure 5. First consider these methods other than GDFM (Hallin and Liška, 2011). Table 2 reports the results for Setups 1 and 2 when we set $p_1 = 900$, $\theta_1 = 45°$ (i.e., $\rho_1 = 0.707$), and $\sigma_e^2 = 1$. All methods except OnPLS have comparably good performance for the estimation of signal matrices. As expected, D-CCA outperforms all the six competing methods in terms of estimating the common and distinctive matrices. In particular, AJIVE and COBE are unable to discover the common matrices. Figure 5 visually shows a similar comparison based on a single replication of Setup 3. The signal, common, and distinctive matrices are recovered well by the D-CCA method. In contrast, the common matrix estimators estimated from the six state-of-the-art methods significantly differ from the ground truth. AJIVE and COBE still yield zero matrices as the estimators of the common matrices, which appears not reasonable when the first canonical correlation ρ_1 has a high value of 0.707. Table 3 shows the proportion of significant nonzero correlations among the $p_1 \times p_2$ pairs of variables between $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ that were detected by the normal approximation test (DiCiccio and Romano, 2017) using each method's estimates of $\mathbf{D}_1$ and $\mathbf{D}_2$. The procedure of Benjamini and Hochberg (1995) was applied to the multiple tests to control false discovery rate at 0.05. Results are omitted for AJIVE and COBE with $\widehat{\mathbf{C}}_k = \mathbf{0}$, and also for D-CCA and R.JIVE due to zero correlation estimates by $\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$. All the other methods have a large amount of significant nonzero correlations retained between their distinctive structures.

Now consider the GDFM method (Hallin and Liška, 2011). We set the sample temporal cross-covariances to be zero in GDFM estimation for our simulated data and TCGA datasets

that have no temporal dependence. GDFM decomposes each data matrix by $\mathbf{Y}_k = \boldsymbol{\chi}_k^* + \boldsymbol{\xi}_k^* = (\boldsymbol{\phi}_k + \boldsymbol{\psi}_k) + (\boldsymbol{\nu}_k + \boldsymbol{\xi}_k^*) = \boldsymbol{\chi}_k + \boldsymbol{\xi}_k$ with each component's name shown in Table 6. By Remark S.1 (in the supplementary material), theoretically for our simulated i.i.d. data with no correlations between signals and noises, the weakly idiosyncratic matrix $\boldsymbol{\nu}_k$ is zero, and the joint common matrix $\boldsymbol{\chi}_k^*$ and the marginal common matrix $\boldsymbol{\chi}_k$ are both equal to the signal matrix $\mathbf{X}_k$. Moreover, the strongly common matrix $\boldsymbol{\phi}_k$ is zero, when $\mathrm{span}(\boldsymbol{x}_1^\top) \cap \mathrm{span}(\boldsymbol{x}_2^\top) = \{0\}$, i.e., the first canonical correlation ρ_1 between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is smaller than 1. The above theoretical results are evidenced by our simulations. In Table 2, the relative errors of estimators $\widehat{\boldsymbol{\chi}}_k^*$ and $\widehat{\boldsymbol{\chi}}_k$ to signal $\mathbf{X}_k$ are as comparably small as those of $\widehat{\mathbf{X}}_k$ by our D-CCA and the other five well performed methods. The similarly small norm ratios of $\widehat{\boldsymbol{\nu}}_k$ to $\widehat{\boldsymbol{\chi}}_k^*$ numerically support $\boldsymbol{\nu}_k = \mathbf{0}$. The squares of these quantities are much smaller, and especially in the Frobenius norm are equivalent to matrix-variation ratios. The strongly common matrix estimate $\widehat{\boldsymbol{\phi}}_k$ is zero for the setups, with $\rho_1 = 0.707 < 1$, considered in the table. These numerical evidences are more clearly seen in Figure 5(b) under a similar setup.

Table 1: Averages (standard errors) of D-CCA estimates for the first canonical angle/correlation.

(p_1, σ_e^2)	$\theta_1 = 0°/\rho_1 = 1$	$\theta_1 = 45°/\rho_1 = 0.707$	$\theta_1 = 60°/\rho_1 = 0.5$	$\theta_1 = 75°/\rho_1 = 0.259$
		Setup 1		
$(100, 1)$	3.59°(0.21°)/0.998(0.000)	44.7°(2.38°)/0.710(0.029)	59.3°(2.88°)/0.509(0.043)	73.5°(3.06°)/0.284(0.051)
$(600, 1)$	3.61°(0.21°)/0.998(0.000)	44.7°(2.39°)/0.710(0.029)	59.4°(2.89°)/0.509(0.043)	73.5°(3.07°)/0.284(0.051)
$(900, 1)$	3.61°(0.21°)/0.998(0.000)	44.7°(2.39°)/0.710(0.029)	59.4°(2.90°)/0.509(0.043)	73.5°(3.09°)/0.283(0.052)
$(1500, 1)$	3.61°(0.21°)/0.998(0.000)	44.7°(2.39°)/0.710(0.029)	59.3°(2.89°)/0.509(0.043)	73.5°(3.08°)/0.284(0.051)
$(900, 0.01)$	0.36°(0.02°)/1.000(0.000)	44.6°(2.38°)/0.711(0.025)	59.3°(2.89°)/0.508(0.038)	73.5°(3.08°)/0.280(0.046)
$(900, 1)$	3.61°(0.21°)/0.998(0.000)	44.7°(2.39°)/0.709(0.026)	59.4°(2.90°)/0.507(0.038)	73.5°(3.09°)/0.280(0.046)
$(900, 9)$	11.0°(0.66°)/0.992(0.001)	45.6°(2.43°)/0.705(0.026)	59.9°(2.92°)/0.504(0.039)	73.7°(3.08°)/0.279(0.046)
$(900, 16)$	14.9°(0.91°)/0.966(0.004)	46.4°(2.47°)/0.688(0.028)	60.4°(2.93°)/0.492(0.040)	73.9°(3.06°)/0.273(0.047)
		Setup 2		
$(100, 1)$	3.58°(0.21°)/0.998(0.000)	44.5°(2.36°)/0.712(0.029)	59.0°(2.83°)/0.514(0.042)	72.7°(2.90°)/0.296(0.048)
$(600, 1)$	3.59°(0.21°)/0.998(0.000)	44.5°(2.36°)/0.712(0.029)	59.0°(2.83°)/0.514(0.042)	72.7°(2.89°)/0.297(0.048)
$(900, 1)$	3.60°(0.21°)/0.998(0.000)	44.5°(2.37°)/0.712(0.029)	59.0°(2.84°)/0.514(0.043)	72.7°(2.90°)/0.296(0.048)
$(1500, 1)$	3.60°(0.21°)/0.998(0.000)	44.5°(2.36°)/0.712(0.029)	59.0°(2.82°)/0.514(0.042)	72.7°(2.89°)/0.296(0.048)
$(900, 0.01)$	0.36°(0.02°)/1.000(0.000)	44.4°(2.35°)/0.714(0.029)	59.0°(2.82°)/0.515(0.042)	72.7°(2.89°)/0.297(0.048)
$(900, 1)$	3.60°(0.21°)/0.998(0.000)	44.5°(2.37°)/0.712(0.029)	59.0°(2.84°)/0.514(0.043)	72.7°(2.90°)/0.296(0.048)
$(900, 9)$	10.9°(0.64°)/0.982(0.002)	45.4°(2.41°)/0.701(0.030)	59.6°(2.87°)/0.506(0.043)	73.0°(2.93°)/0.292(0.049)
$(900, 16)$	14.6°(0.87°)/0.967(0.004)	46.3°(2.45°)/0.691(0.031)	60.0°(2.89°)/0.499(0.044)	73.2°(2.93°)/0.289(0.049)

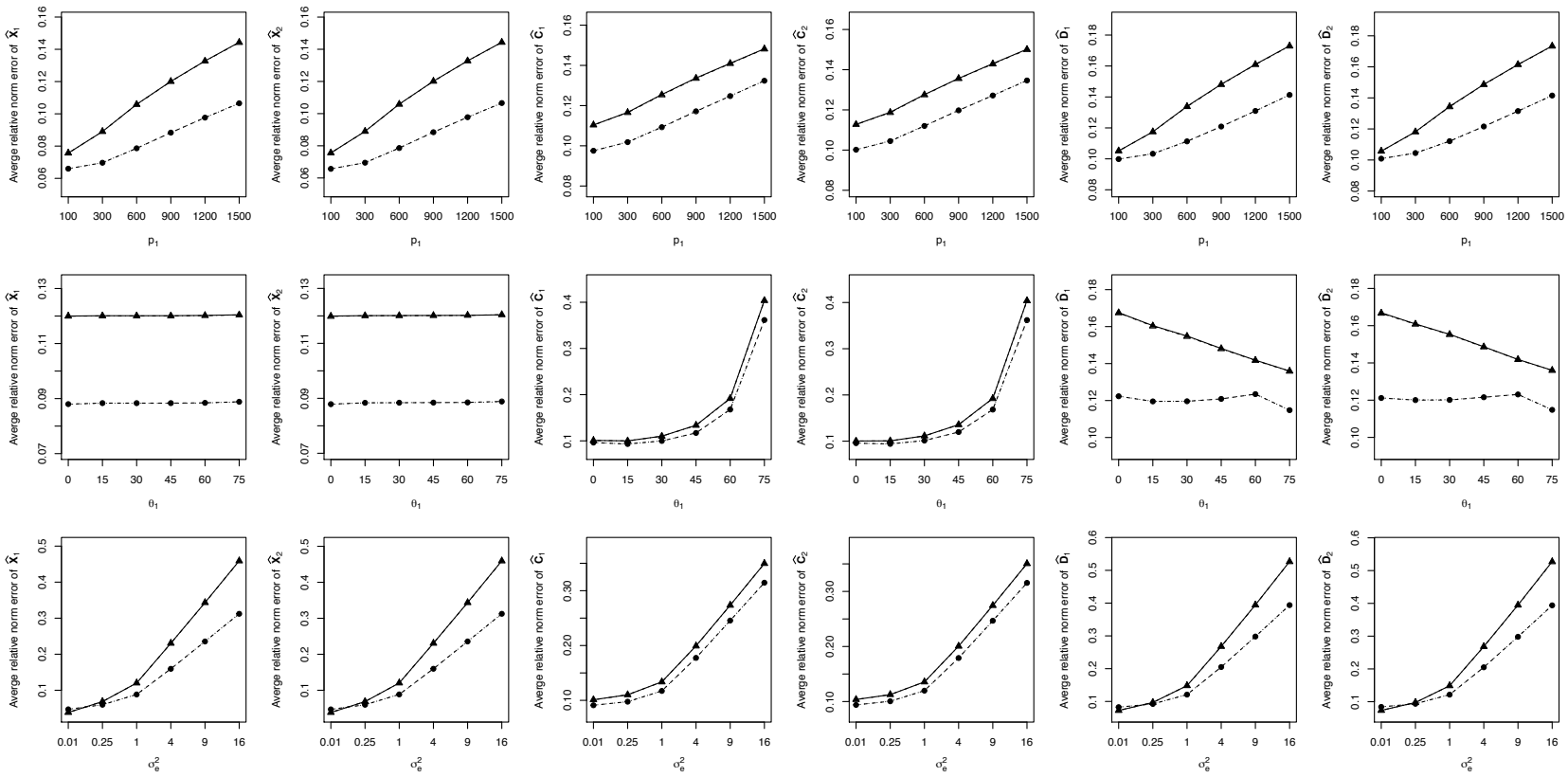


Figure 3: Average relative errors of D-CCA estimates under Setup 1 in spectral norm (○) and Frobenius norm (△) using true r_1, r_2 and r_{12}, and those in spectral norm (●) and Frobenius norm (▲) using $\widehat{r}_1, \widehat{r}_2$ and $\widehat{r}_{12}$.

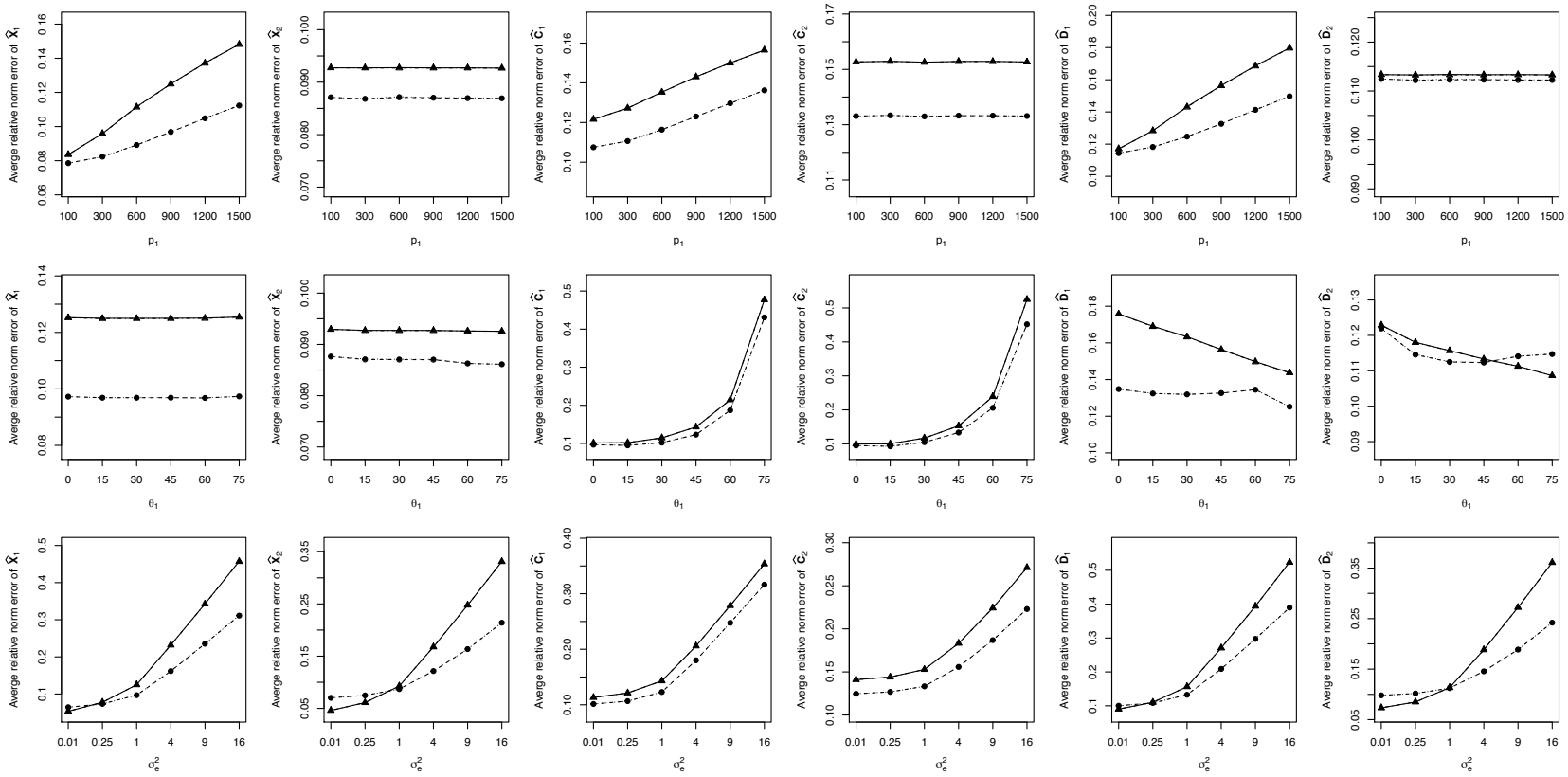


Figure 4: Average relative errors of D-CCA estimates under Setup 2 in spectral norm (○) and Frobenius norm (△) using true r_1, r_2 and r_{12}, and those in spectral norm (●) and Frobenius norm (▲) using $\widehat{r}_1, \widehat{r}_2$ and $\widehat{r}_{12}$.

Table 2: Averages (standard errors) of norm ratios when $p_1 = 900, \theta_1 = 45^\circ$ and $\sigma_e^2 = 1$.

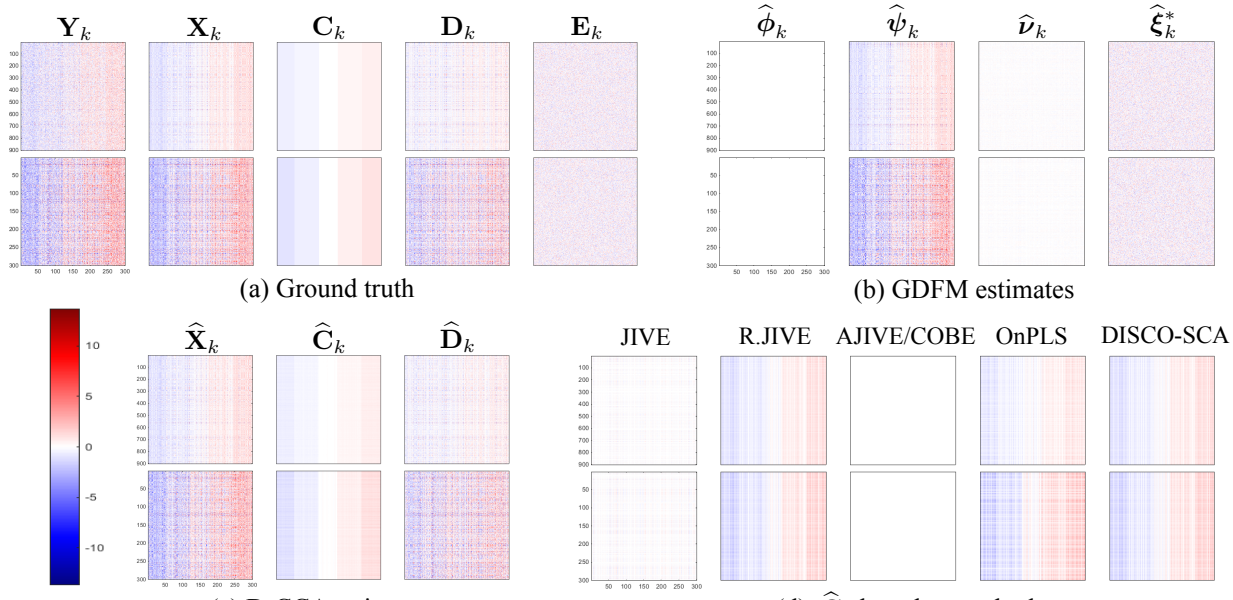
Ratio	Method	Spectral norm	Frobenius norm	Spectral norm	Frobenius norm
		Setup 1		Setup 2	
		$k=1$ / $k=2$	$k=1$ / $k=2$	$k=1$ / $k=2$	$k=1$ / $k=2$
$\frac{\|\widehat{\mathbf{X}}_k-\mathbf{X}_k\|_{(\cdot)}}{\|\mathbf{X}_k\|_{(\cdot)}}$	D-CCA	0.088(0.010)/0.088(0.010)	0.120(0.006)/0.120(0.006)	0.097(0.012)/0.087(0.017)	0.125(0.007)/0.093(0.006)
	JIVE	0.108(0.005)/0.109(0.005)	0.141(0.004)/0.141(0.004)	0.116(0.005)/0.067(0.004)	0.145(0.004)/0.090(0.002)
	R.JIVE	0.109(0.018)/0.089(0.015)	0.139(0.013)/0.140(0.009)	0.108(0.018)/0.102(0.026)	0.139(0.012)/0.105(0.011)
	AJIVE	0.080(0.004)/0.081(0.004)	0.116(0.003)/0.116(0.004)	0.081(0.004)/0.051(0.002)	0.116(0.003)/0.082(0.002)
	OnPLS	0.390(0.111)/0.399(0.112)	0.315(0.076)/0.321(0.077)	0.397(0.111)/0.550(0.116)	0.320(0.077)/0.331(0.064)
	DISCO-SCA	0.083(0.003)/0.083(0.004)	0.154(0.004)/0.154(0.005)	0.084(0.004)/0.053(0.002)	0.174(0.005)/0.093(0.002)
	COBE	0.080(0.004)/0.081(0.004)	0.116(0.003)/0.116(0.004)	0.081(0.004)/0.051(0.002)	0.116(0.003)/0.082(0.002)
$\frac{\|\widehat{\mathbf{C}}_k-\mathbf{C}_k\|_{(\cdot)}}{\|\mathbf{C}_k\|_{(\cdot)}}$	D-CCA	0.117(0.028)/0.120(0.027)	0.134(0.028)/0.136(0.027)	0.123(0.028)/0.133(0.036)	0.143(0.029)/0.153(0.038)
	JIVE	0.996(0.009)/0.996(0.008)	1.024(0.014)/1.024(0.013)	0.998(0.009)/0.990(0.015)	1.037(0.025)/1.013(0.019)
	R.JIVE	1.000(0.043)/0.576(0.032)	1.003(0.043)/0.588(0.031)	1.003(0.049)/0.576(0.041)	1.006(0.052)/0.589(0.042)
	AJIVE	1(0)/1(0)	1(0)/1(0)	1(0)/1(0)	1(0)/1(0)
	OnPLS	0.787(0.112)/0.777(0.113)	0.817(0.143)/0.805(0.142)	0.779(0.105)/0.796(0.071)	0.804(0.117)/0.815(0.098)
	DISCO-SCA	1.023(0.065)/1.023(0.066)	1.057(0.087)/1.058(0.089)	0.772(0.183)/1.052(0.112)	0.826(0.227)/1.190(0.237)
	COBE	1(0)/1(0)	1(0)/1(0)	1(0)/1(0)	1(0)/1(0)
$\frac{\|\widehat{\mathbf{D}}_k-\mathbf{D}_k\|_{(\cdot)}}{\|\mathbf{D}_k\|_{(\cdot)}}$	D-CCA	0.121(0.016)/0.122(0.016)	0.148(0.010)/0.149(0.009)	0.133(0.018)/0.112(0.019)	0.156(0.011)/0.113(0.009)
	JIVE	0.703(0.040)/0.703(0.040)	0.541(0.023)/0.541(0.023)	0.704(0.040)/0.599(0.036)	0.546(0.024)/0.371(0.016)
	R.JIVE	0.689(0.040)/0.405(0.032)	0.535(0.019)/0.337(0.020)	0.690(0.041)/0.350(0.031)	0.536(0.022)/0.238(0.017)
	AJIVE	0.706(0.040)/0.706(0.040)	0.538(0.022)/0.539(0.022)	0.705(0.040)/0.605(0.035)	0.538(0.022)/0.369(0.015)
	OnPLS	0.655(0.093)/0.654(0.095)	0.574(0.064)/0.576(0.066)	0.656(0.094)/0.658(0.113)	0.574(0.063)/0.476(0.057)
	DISCO-SCA	0.704(0.049)/0.704(0.049)	0.558(0.041)/0.559(0.041)	0.532(0.114)/0.628(0.067)	0.462(0.092)/0.432(0.078)
	COBE	0.706(0.040)/0.706(0.040)	0.538(0.022)/0.539(0.022)	0.705(0.040)/0.605(0.035)	0.538(0.022)/0.369(0.015)
$\frac{\|\widehat{\mathbf{x}}_k-\mathbf{X}_k\|_{(\cdot)}}{\|\mathbf{X}_k\|_{(\cdot)}}$	GDFM	0.080(0.004)/0.081(0.004)	0.116(0.003)/0.116(0.004)	0.081(0.004)/0.052(0.002)	0.116(0.003)/0.082(0.002)
$\frac{\|\widehat{\mathbf{x}}_k^*-\mathbf{X}_k\|_{(\cdot)}}{\|\mathbf{X}_k\|_{(\cdot)}}$	GDFM	0.083(0.003)/0.083(0.004)	0.154(0.004)/0.154(0.005)	0.084(0.004)/0.053(0.002)	0.174(0.005)/0.093(0.002)
$\frac{\|\widehat{\boldsymbol{\nu}}_k\|_{(\cdot)}}{\|\widehat{\mathbf{x}}_k^*\|_{(\cdot)}}$	GDFM	0.080(0.003)/0.080(0.004)	0.099(0.003)/0.099(0.003)	0.082(0.003)/0.047(0.002)	0.128(0.004)/0.044(0.001)

Table 3: The proportions of significant nonzero correlations between $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ for simulation setups (with $p_1 = 900$, $\theta_1 = 45^\circ$ and $\sigma_e^2 = 1$) and TCGA datasets. Averages (standard errors) are shown for Setups 1 and 2. Significant correlations are detected by the normal approximation test (DiCiccio and Romano, 2017) using $\widehat{\mathbf{D}}_1$ and $\widehat{\mathbf{D}}_2$, with false discovery rate controlled at 0.05.

Method	Setup 1	Setup 2	Setup 3	EXP90/METH90b	EXP90/METH90a
D-CCA	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$
JIVE	69.9%(2.5%)	60.8%(3.0%)	98.7%	85.0%	58.2%
R.JIVE	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$	$\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_2^\top = \mathbf{0}$
AJIVE	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$
OnPLS	56.6%(11.1%)	32.7%(8.2%)	52.5%	72.9%	68.6%
DISCO-SCA	50.3%(4.6%)	25.2%(6.7%)	25.1%	67.8%	64.2%
COBE	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$	$\widehat{\mathbf{C}}_k = \mathbf{0}$
GDFM ($\widehat{\mathbf{D}}_k$=$\widehat{\boldsymbol{\psi}}_k$)	70.3%(2.4%)	61.5%(2.7%)	98.6%	100%	100%
GDFM ($\widehat{\mathbf{D}}_k$=$\widehat{\boldsymbol{\psi}}_k$+$\widehat{\boldsymbol{\nu}}_k$)	73.8%(1.8%)	64.8%(2.3%)	97.0%	85.8%	87.0%

5 Analysis of TCGA Breast Cancer Data

In this section, we apply the proposed D-CCA method to analyze genomic datasets produced from TCGA breast cancer tumor samples. We investigate the ability to separate tumor subtypes for matrices obtained from D-CCA in comparison to those obtained from the six competing



(a) Ground truth (b) GDFM estimates

(c) D-CCA estimates (d) $\widehat{\mathbf{C}}_k$ by other methods

Figure 5: Color maps for a single replication of Setup 3.

methods as well as GDFM that are mentioned in Section 1. We consider the mRNA expression data and DNA methylation data for a common set of 660 samples. The two datasets are publicly available at https://tcga-data.nci.nih.gov/docs/publications and have been respectively preprocessed by Ciriello et al. (2015) and Koboldt et al. (2012). The 660 samples were classified by Ciriello et al. (2015) into 4 subtypes using the PAM50 model (Parker et al., 2009) based on mRNA expression data. Specifically, the samples consist of 112 basal-like, 55 HER2-enriched, 331 luminal A, and 162 luminal B tumors.

To quantify the extent of subtype separation, we adopt the standardized within-class sum of squares (SWISS; Cabanski et al., 2010)

$$\text{SWISS}(\mathbf{A}) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(A_{ij} - \bar{A}_{i,s(j)})^2}{\sum_{i=1}^{p}\sum_{j=1}^{n}(A_{ij} - \bar{A}_{i\cdot})^2}$$

for matrix $\mathbf{A} = (A_{ij})_{p\times n}$, where $\bar{A}_{i,s(j)}$ is the average of the j-th sample's subtype on the i-th row and $\bar{A}_{i\cdot}$ is the average of the i-th row's elements. The SWISS score represents the variation within the subtypes as a proportion of the total variation. A lower score indicates better subtype separation. For the mRNA expression data, we filtered out the subset consisting of the 1,195 variably expressed genes with marginal SWISS$\leq$0.9 from the original 20,533 genes, and denote this subset as EXP90. The 2,083 variably methylated probes of the DNA methylation data, orig-

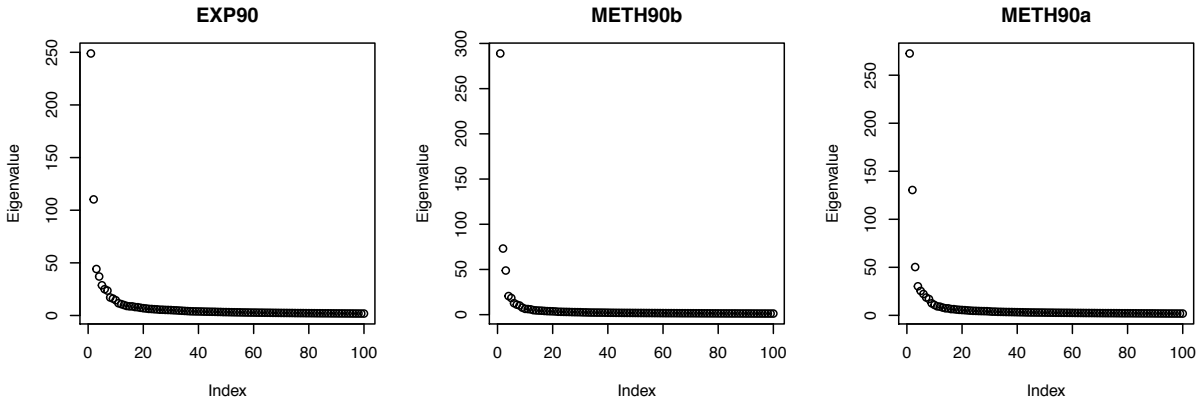


Figure 6: The scree plot of the sample covariance matrix $\frac{1}{n}\mathbf{Y}_k\mathbf{Y}_k^\top$ for each TCGA dataset.

inally with 21,986 probes, are included in the analysis. We denote the 881 probes with marginal SWISS$\leq$0.9 as METH90b and the remaining 1,202 probes as METH90a. We conducted the analysis for the pair of EXP90 and METH90b as well as the pair of EXP90 and METH90a.

The ranks and proportions of explained signal variation for the matrix estimators obtained by D-CCA and the six competing methods (except GDFM) are given in Table 4, and their SWISS scores are shown in Table 5. We see in Table 4 that D-CCA, AJIVE and COBE give much lower ranks for the estimated signal matrices $\{\widehat{\mathbf{X}}_k\}_{k=1}^2$ than the other methods. Particularly for the EXP90 dataset, the rank of $\widehat{\mathbf{X}}_1$ obtained by the remaining four methods is inconsistent for the two pairs. As shown in the scree plots of Figure 6, the ranks of signal matrices selected by D-CCA and AJIVE look reasonable because the few most leading principal components of the observed data are captured for denoising, while the signal matrix ranks for the METH90b and METH90a datasets seem to be underestimated by COBE. Using D-CCA, the estimated canonical correlations and angles of signal vectors are $(0.934, 0.431)$ and $(20.9°, 64.4°)$ between the EXP90 and METH90b datasets, and are $(0.610, 0.275)$ and $(52.4°, 74.0°)$ between the EXP90 and METH90a datasets.

From Table 5, for the pair of EXP90 and METH90b datasets, the matrix $\widehat{\mathbf{X}}_k$ obtained by all the seven methods gains an improved SWISS score compared to the noisy data matrix $\mathbf{Y}_k$. Other than AJIVE and COBE with $\widehat{\mathbf{C}}_k = \mathbf{0}$, a clear pattern of increasing SWISS scores, from $\widehat{\mathbf{C}}_k$ to $\widehat{\mathbf{X}}_k$ and then to $\widehat{\mathbf{D}}_k$, can be seen for the remaining methods except JIVE. This indicates that an enhanced ability to separate the tumor samples by subtype can be expected when integrating two datasets that can exhibit such a distinction to a moderate extent. Also note that the estimated

Table 4: Ranks (and proportions of explained signal variation, i.e., $\|\cdot\|_F^2/\|\widehat{\mathbf{X}}_k\|_F^2$) of matrix estimates for TCGA datasets.

Matrix	Method	EXP90 / METH90b	EXP90 / METH90a
$\widehat{\mathbf{X}}_k$	D-CCA	2 / 3	2 / 3
	JIVE	35 / 18	41 / 29
	R.JIVE	40 / 27	44 / 49
	AJIVE	2 / 3	2 / 3
	OnPLS	13 / 10	12 / 10
	DISCO-SCA	13 / 13	17 / 17
	COBE	2 / 1	2 / 2
$\widehat{\mathbf{C}}_k$	D-CCA	2 (0.472) / 2 (0.301)	2 (0.120) / 2 (0.062)
	JIVE	1 (0.068) / 1 (0.086)	3 (0.236) / 3 (0.167)
	R.JIVE	1 (0.212) / 1 (0.505)	3 (0.274) / 3 (0.602)
	AJIVE	0 / 0	0 / 0
	OnPLS	3 (0.516) / 3 (0.510)	2 (0.455) / 2 (0.166)
	DISCO-SCA	6 (0.732) / 6 (0.571)	8 (0.745) / 8 (0.363)
	COBE	0 / 0	0 / 0
$\widehat{\mathbf{D}}_k$	D-CCA	2 (0.223) / 3 (0.506)	2 (0.564) / 3 (0.797)
	JIVE	34 (0.932) / 17 (0.914)	38 (0.764) / 26 (0.833)
	R.JIVE	39 (0.788) / 26 (0.495)	41 (0.726) / 46 (0.398)
	AJIVE	2 (1) / 3 (1)	2 (1) / 3 (1)
	OnPLS	10 (0.484) / 7 (0.490)	10 (0.545) / 8 (0.834)
	DISCO-SCA	7 (0.268) / 7 (0.429)	9 (0.255) / 9 (0.637)
	COBE	2 (1) / 1 (1)	2 (1) / 2 (1)

Table 5: SWISS scores for TCGA breast cancer subtypes. Lower scores indicate better subtype separation.

Matrix	Method	EXP90 / METH90b	EXP90 / METH90a
$\mathbf{Y}_k$	For all	0.773 / 0.814	0.773 / 0.952
$\widehat{\mathbf{X}}_k$	D-CCA	0.313 / 0.623	0.313 / 0.925
	JIVE	0.632 / 0.698	0.643 / 0.920
	R.JIVE	0.642 / 0.689	0.647 / 0.931
	AJIVE	0.314 / 0.623	0.314 / 0.925
	OnPLS	0.523 / 0.669	0.515 / 0.905
	DISCO-SCA	0.526 / 0.663	0.553 / 0.904
	COBE	0.314 / 0.545	0.314 / 0.926
$\widehat{\mathbf{C}}_k$	D-CCA	0.240 / 0.269	0.528 / 0.606
	JIVE	0.831 / 0.831	0.639 / 0.736
	R.JIVE	0.373 / 0.373	0.564 / 0.885
	AJIVE	NA / NA	NA / NA
	OnPLS	0.398 / 0.312	0.419 / 0.494
	DISCO-SCA	0.447 / 0.400	0.470 / 0.717
	COBE	NA / NA	NA / NA
$\widehat{\mathbf{D}}_k$	D-CCA	0.623 / 0.940	0.320 / 0.979
	JIVE	0.691 / 0.741	0.830 / 0.963
	R.JIVE	0.833 / 0.997	0.874 / 0.998
	AJIVE	0.314 / 0.623	0.314 / 0.925
	OnPLS	0.878 / 0.978	0.871 / 0.989
	DISCO-SCA	0.935 / 0.992	0.944 / 0.995
	COBE	0.314 / 0.545	0.314 / 0.926

Table 6: Ranks, variation ratios (VR=$\|\cdot\|_F^2/\|\widehat{\boldsymbol{\chi}}_k^*\|_F^2$), and SWISS scores of GFDM matrix estimates for TCGA datasets.

Matrix Estimate	EXP90 / METH90b		EXP90 / METH90a	
	Rank (VR)	SWISS	Rank (VR)	SWISS
$\widehat{\boldsymbol{\chi}}_k^*$ (joint common)	4 / 4	0.373 / 0.569	4 / 4	0.378 / 0.850
$\widehat{\boldsymbol{\chi}}_k$ (marginal common)	3 (0.986) / 3 (0.986)	0.364 / 0.566	3 (0.990) / 3 (0.974)	0.372 / 0.851
$\widehat{\boldsymbol{\phi}}_k$ (strongly common)	2 (0.755) / 2 (0.626)	0.288 / 0.348	2 (0.770) / 2 (0.372)	0.302 / 0.613
$\widehat{\boldsymbol{\psi}}_k$ (weakly common)	1 (0.231) / 1 (0.360)	0.764 / 0.991	1 (0.220) / 1 (0.602)	0.811 / 0.996
$\widehat{\boldsymbol{\nu}}_k$ (weakly idiosyncratic)	1 (0.014) / 1 (0.014)	0.987 / 0.760	1 (0.010) / 1 (0.026)	0.997 / 0.812
$\widehat{\boldsymbol{\psi}}_k + \widehat{\boldsymbol{\nu}}_k$	2 (0.245) / 2 (0.374)	0.777 / 0.982	2 (0.230) / 2 (0.628)	0.819 / 0.988
$\widehat{\boldsymbol{\xi}}_k^*$ (strongly idiosyncratic)	656 (2.070) / 656 (1.196)	0.985 / 0.987	656 (2.151) / 656 (1.764)	0.977 / 0.989
$\widehat{\boldsymbol{\xi}}_k$ (marginal idiosyncratic)	657 (2.084) / 657 (1.211)	0.985 / 0.983	657 (2.160) / 657 (1.790)	0.977 / 0.987

common matrices of our D-CCA have the lowest SWISS scores. While considering the pair of EXP90 and METH90a datasets, for all the seven methods we find a big gap between the SWISS scores of the two estimated signal matrices, and that the denoised matrix of the METH90a dataset still has nearly no discriminative power with SWISS close to 1. The ability on subtype separation seems more likely to be a distinctive feature of EXP90 dataset comparing to METH90a dataset. The estimated distinctive matrix of EXP90 is thus expected to have a lower SWISS score than its estimated common matrix. However, only D-CCA meets this point, except that AJIVE and COBE yield zero common matrices. The failure of the six competing methods may be caused by their inappropriate decomposition constructions, which are mentioned in Section 1. In particular, from Table 3, we see that a lot of significant nonzero correlations exist among all gene-probe pairs based on the estimated distinctive matrices, respectively, obtained by JIVE, OnPLS and DISCO-SCA.

The GDFM method (Hallin and Liška, 2011) was also applied to the TCGA datasets. Table 6 summarizes the results of GDFM matrix estimates. As estimators of signal matrix $\mathbf{X}_k$, matrix $\widehat{\boldsymbol{\chi}}_k$ has comparable rank and SWISS score as those of our D-CCA estimator $\widehat{\mathbf{X}}_k$ given in Tables 4 and 5. Besides, $\boldsymbol{\chi}_k^* = \boldsymbol{\chi}_k$, i.e., $\boldsymbol{\nu}_k = \mathbf{0}$, is numerically suggested by the remarkably small variation ratios of $\widehat{\boldsymbol{\nu}}_k$ to $\widehat{\boldsymbol{\chi}}_k^*$ that are likely just induced by estimation errors. With very large ranks and uninformative SWISS scores, both $\widehat{\boldsymbol{\xi}}_k^*$ and $\widehat{\boldsymbol{\xi}}_k$ appear to be noises. One may let $\widehat{\mathbf{C}}_k = \widehat{\boldsymbol{\phi}}_k$, $\widehat{\mathbf{D}}_k = \widehat{\boldsymbol{\psi}}_k$ (or $\widehat{\mathbf{D}}_k = \widehat{\boldsymbol{\psi}}_k + \widehat{\boldsymbol{\nu}}_k$), and $\widehat{\mathbf{X}}_k = \boldsymbol{\chi}_k$ (or $\widehat{\mathbf{X}}_k = \boldsymbol{\chi}_k^*$) for GDFM. Inspecting Table 6 reveals that the discussion given in the preceding paragraph also holds even when we include GDFM.

6 Discussion

In this paper, we study a typical model for the joint analysis of two high-dimensional datasets. We develop a novel and promising decomposition-based CCA method, D-CCA, to appropriately define the common and distinctive matrices. In particular, the conventionally underemphasized orthogonal relationship between the distinctive matrices is now well designed on the $\mathcal{L}^2$ space of random variables. A soft-thresholding-based approach is then proposed for estimating these D-CCA-defined matrices with a theoretical guarantee and satisfactory numerical performance. The proposed D-CCA outperforms some state-of-the-art methods in both simulated and real data analyses.

There are many possible further studies beyond the current proposed D-CCA. The first is to generalize the D-CCA for three or more datasets. We may assume that at least two datasets have mutually orthogonal distinctive structures. An immediate idea starts from substituting the multiset CCA (Kettenring, 1971) for the two-set CCA in D-CCA. However, the challenge is that the iteratively obtained sets of canonical variables are not guaranteed to have the bi-orthogonality given in Theorem 1. Hence, we cannot follow the proposed D-CCA to simply break down the decomposition problem to each set of canonical variables, and need a more sophisticated design to meet the desirable constraint. Another direction is to incorporate the nonlinear relationship between the two datasets. The D-CCA only considers the linear relationship by using the traditional CCA based on Pearson's correlation. It is worth trying the kernel CCA (Fukumizu et al., 2007) or the distance correlation (Székely et al., 2007) to capture the nonlinear dependence. Inspired by the time series analysis of Hallin and Liška (2011) and Barigozzi et al. (2018), we also expect to generalize D-CCA to general dynamic factor models with comparisons to their methods. These interesting and challenging studies are under investigation and will be reported in future work.

References

Ahn, S. C. and Horenstein, A. R. (2013), "Eigenvalue ratio test for the number of factors," *Econometrica*, 81, 1203–1227.

Bai, J. (2003), "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

Bai, J. and Ng, S. (2002), "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

— (2008), "Large dimensional factor analysis," *Foundations and Trends in Econometrics*, 3, 89–163.

Barigozzi, M., Hallin, M., and Soccorsi, S. (2018), "Identification of global and local shocks in international financial markets via general dynamic factor models," *Journal of Financial Econometrics*, DOI: 10.1093/jjfinec/nby006.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Björck, A. and Golub, G. H. (1973), "Numerical methods for computing angles between linear subspaces," *Mathematics of Computation*, 27, 579–594.

Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., Wilkerson, M. D., Marron, J. S., Perou, C. M., and Hayes, D. N. (2010), "SWISS MADE: Standardized within class sum of squares to evaluate methodologies and dataset elements," *PLoS ONE*, 5, e9905.

Chamberlain, G. and Rothschild, M. (1983), "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, 51, 1281–1304.

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

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015), "Comprehensive molecular portraits of invasive lobular breast cancer," *Cell*, 163, 506–519.

Comon, P. and Jutten, C. (2010), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press.

DiCiccio, C. J. and Romano, J. P. (2017), "Robust permutation tests for correlation and regression coefficients," *Journal of the American Statistical Association*, 112, 1211–1220.

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B*, 75, 603–680.

Feng, Q., Jiang, M., Hannig, J., and Marron, J. (2018), "Angle-based joint and individual variation explained," *Journal of Multivariate Analysis*, 166, 241–265.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and statistics*, 82, 540–554.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2017), "Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis," *Journal of Econometrics*, 199, 74–92.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007), "Statistical consistency of kernel canonical correlation analysis," *Journal of Machine Learning Research*, 8, 361–383.

Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015), "Minimax estimation in sparse canonical correlation analysis," *The Annals of Statistics*, 43, 2168–2197.

Gao, C., Ma, Z., and Zhou, H. H. (2017), "Sparse CCA: Adaptive estimation and computational barriers," *The Annals of Statistics*, 45, 2074–2101.

Hallin, M. and Liška, R. (2011), "Dynamic factors in the presence of blocks," *Journal of Econometrics*, 163, 29–41.

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

Huang, H. (2017), "Asymptotic behavior of support vector machine for spiked population model," *Journal of Machine Learning Research*, 18, 1–21.

Kettenring, J. R. (1971), "Canonical analysis of several sets of variables," *Biometrika*, 58, 433–451.

Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., et al. (2012), "Comprehensive molecular portraits of human breast tumours," *Nature*, 490, 61–70.

Kuligowski, J., Pérez-Guaita, D., Sánchez-Illana, Á., León-González, Z., de la Guardia, M., Vento, M., Lock, E. F., and Quintás, G. (2015), "Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE)," *Analyst*, 140, 4521–4529.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013), "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types," *Annals of Applied Statistics*, 7, 523–542.

Löfstedt, T. and Trygg, J. (2011), "OnPLS–a novel multiblock method for the modelling of predictive and orthogonal variation," *Journal of Chemometrics*, 25, 441–455.

Nadakuditi, R. R. and Silverstein, J. W. (2010), "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples," *IEEE Journal of Selected Topics in Signal Processing*, 4, 468–480.

O'Connell, M. J. and Lock, E. F. (2016), "R.JIVE for exploration of multi-source molecular data," *Bioinformatics*, 32, 2877–2879.

Onatski, A. (2010), "Determining the number of factors from empirical distribution of eigenvalues," *The Review of Economics and Statistics*, 92, 1004–1016.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009), "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, 27, 1160–1167.

Ross, S. A. (1976), "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13, 341–360.

Schouteden, M., Van Deun, K., Wilderjans, T. F., and Van Mechelen, I. (2014), "Performing DISCO-SCA to search for distinctive and common information in linked data," *Behavior Research Methods*, 46, 576–587.

Smilde, A. K., Mage, I., Naes, T., Hankemeier, T., Lips, M. A., Kiers, H. A. L., Acar, E., and Bro, R. (2017), "Common and distinct components in data fusion," *Journal of Chemometrics*, 31, e2900.

Song, Y., Schreier, P. J., Ramírez, D., and Hasija, T. (2016), "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Processing*, 128, 449–458.

Stock, J. H. and Watson, M. W. (2002), "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 97, 1167–1179.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 35, 2769–2794.

Trygg, J. (2002), "O2-PLS for qualitative and quantitative analysis in multivariate calibration," *Journal of Chemometrics*, 16, 283–293.

van der Kloet, F., Sebastian-Leon, P., Conesa, A., Smilde, A., and Westerhuis, J. (2016), "Separating common from distinctive variation," *BMC Bioinformatics*, 17, S195.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., and Ugurbil, K. (2013), "The WU-Minn human connectome project: an overview," *NeuroImage*, 80, 62–79.

Vershynin, R. (2012), "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing*, Cambridge University Press, Cambridge, pp. 210–268.

Wang, W. and Fan, J. (2017), "Asymptotics of empirical eigenstructure for high dimensional spiked covariance," *The Annals of Statistics*, 45, 1342–1374.

Yu, Q., Risk, B. B., Zhang, K., and Marron, J. (2017), "Jive integration of imaging and behavioral data," *NeuroImage*, 152, 38–49.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016), "Group component analysis for multiblock data: common and individual feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, 27, 2426–2439.

Supplementary Material to "D-CCA: A Decomposition-based Canonical Correlation Analysis for High-Dimensional Datasets"

Hai Shu, Xiao Wang, and Hongtu Zhu

Abstract

Two important propositions and all technical proofs are given in Section S.1. Additional simulations are included in Section S.2.

S.1 Propositions and Technical Proofs

Proposition S.1. *Let* $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top)^\top$, $\boldsymbol{e} = (\boldsymbol{e}_1^\top, \boldsymbol{e}_2^\top)^\top$, *and* $r_x = \mathrm{rank}(\mathrm{cov}(\boldsymbol{x}))$. *Assume that ranks* r_1 *and* r_2 *are constants. When* $p \coloneqq \min(p_1, p_2) \to \infty$, *if* $\min_{k\in\{1,2\}} \lambda_{r_k}(\mathrm{cov}(\boldsymbol{x}_k))$ *diverges and* $\max_{k\in\{1,2\}} \lambda_1(\mathrm{cov}(\boldsymbol{e}_k))$ *is bounded, then* $\lambda_{r_x}(\mathrm{cov}(\boldsymbol{x}))$ *diverges and* $\lambda_1(\mathrm{cov}(\boldsymbol{e}))$ *is bounded.*

Remark S.1. *Hallin and Liška (2011) proposed a decomposition method under a general dynamic factor model that includes our approximate factor model given in* (6) *and* (7) *as a special case. Their decomposition method divides each of two observed vector processes into four components that are called strongly common, weakly common, weakly idiosyncratic, and strongly idiosyncratic, respectively. Consider applying their method to our approximate factor model that has i.i.d. samples. By their Assumption A3 and Proposition 1(a) and (b) as well as Weyl's inequality, ranks* $\{r_k\}_{k=1}^2$ *are constant, and* $\min_{k\in\{1,2\}} \lambda_{r_k}(\mathrm{cov}(\boldsymbol{x}_k))$ *diverges but* $\max_{k\in\{1,2\}} \lambda_1(\mathrm{cov}(\boldsymbol{e}_k))$ *is bounded when* $p \to \infty$. *Then with the additional condition* $\mathrm{cov}(\boldsymbol{x}, \boldsymbol{e}) = \boldsymbol{0}$, *it follows from our Proposition S.1 and their Proposition 1(c) that for each* $\boldsymbol{y}_k$, $\boldsymbol{x}_k$ *is the sum of strongly common and weakly common components,* $\boldsymbol{e}_k$ *is the strongly idiosyncratic component, and no weakly idiosyncratic component exists. Furthermore, if* $\mathrm{span}(\boldsymbol{x}_1^\top) \cap \mathrm{span}(\boldsymbol{x}_2^\top) = \{0\}$, *i.e., the first signal canonical correlation* $\rho_1 < 1$, *then there is no strongly common component, and* $\boldsymbol{x}_k$ *is entirely the weakly common component of* $\boldsymbol{y}_k$.

Proof of Proposition S.1. Recall that $\boldsymbol{\Sigma}_k = \mathrm{cov}(\boldsymbol{x}_k) = \mathbf{V}_k \boldsymbol{\Lambda}_k \mathbf{V}_k^\top$ and $\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Using (S.4) that will be shown later, we have

$$\begin{aligned}\mathrm{cov}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_1^{1/2} & \\ & \boldsymbol{\Lambda}_2^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{\theta 1} & \\ & \mathbf{U}_{\theta 2} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{r_1 \times r_1} & \boldsymbol{\Lambda}_\theta \\ \boldsymbol{\Lambda}_\theta^\top & \mathbf{I}_{r_2 \times r_2} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} \mathbf{U}_{\theta 1}^\top & \\ & \mathbf{U}_{\theta 2}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_1^{1/2} & \\ & \boldsymbol{\Lambda}_2^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^\top & \\ & \mathbf{V}_2^\top \end{bmatrix} \\ &=: \mathbf{V}_0 \boldsymbol{\Lambda}_0^{1/2} \mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top \boldsymbol{\Lambda}_0^{1/2} \mathbf{V}_0^\top. \end{aligned} \tag{S.1}$$

According to Theorem 3.3.16(d) in Horn and Johnson (1994), we have

$$\begin{aligned}\sigma_{r_x}(\mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top) &= \sigma_{r_x}(\boldsymbol{\Lambda}_0^{-1/2} \boldsymbol{\Lambda}_0^{1/2} \mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top \boldsymbol{\Lambda}_0^{1/2} \boldsymbol{\Lambda}_0^{-1/2}) \\ &\le \sigma_1(\boldsymbol{\Lambda}_0^{-1/2}) \sigma_{r_x}(\boldsymbol{\Lambda}_0^{1/2} \mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top \boldsymbol{\Lambda}_0^{1/2}) \sigma_1(\boldsymbol{\Lambda}_0^{-1/2}).\end{aligned}$$

Since r_k and $\sigma_\ell(\boldsymbol{\Theta}) = \rho_\ell$ are constant for $k \le 2$ and $\ell \le r_{12}$, $\sigma_{r_x}(\mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top) = \sigma_{r_x}(\boldsymbol{\Phi}_0)$ is a positive constant. Thus, when $p \to \infty$, we have

$$\begin{aligned}\lambda_{r_x}(\mathrm{cov}(\boldsymbol{x})) &= \sigma_{r_x}(\mathrm{cov}(\boldsymbol{x})) = \sigma_{r_x}(\boldsymbol{\Lambda}_0^{1/2} \mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top \boldsymbol{\Lambda}_0^{1/2}) \\ &\ge \sigma_{r_x}(\mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top) / \sigma_1^2(\boldsymbol{\Lambda}_0^{-1/2}) = \sigma_{r_x}(\mathbf{U}_0 \boldsymbol{\Phi}_0 \mathbf{U}_0^\top) \min_{k \in \{1,2\}} \lambda_{r_k}(\mathrm{cov}(\boldsymbol{x}_k)) \to \infty.\end{aligned}$$

For the noise covariance matrices, we denote their compact SVDs by $\mathrm{cov}(\boldsymbol{e}_k) = \mathbf{V}_{e,k} \boldsymbol{\Lambda}_{e,k} \mathbf{V}_{e,k}^\top$ for $k = 1, 2$. Similar to (S.1), we have

$$\begin{aligned}\mathrm{cov}(\boldsymbol{e}) &= \begin{bmatrix} \mathbf{V}_{e,1} & \\ & \mathbf{V}_{e,2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{e,1}^{1/2} & \\ & \boldsymbol{\Lambda}_{e,2}^{1/2} \end{bmatrix} \mathbf{U}_e \begin{bmatrix} \mathbf{I}_{r_{e,1} \times r_{e,1}} & \boldsymbol{\Lambda}_{e12} \\ \boldsymbol{\Lambda}_{e12}^\top & \mathbf{I}_{r_{e,2} \times r_{e,2}} \end{bmatrix} \mathbf{U}_e^\top \begin{bmatrix} \boldsymbol{\Lambda}_{e,1}^{1/2} & \\ & \boldsymbol{\Lambda}_{e,2}^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{e,1}^\top & \\ & \mathbf{V}_{e,2}^\top \end{bmatrix} \\ &=: \mathbf{V}_e \boldsymbol{\Lambda}_e^{1/2} \mathbf{U}_e \boldsymbol{\Phi}_e \mathbf{U}_e^\top \boldsymbol{\Lambda}_e^{1/2} \mathbf{V}_e^\top,\end{aligned}$$

where $\mathbf{U}_e$ is an orthogonal matrix, $r_{e,k} = \mathrm{rank}(\mathrm{cov}(\boldsymbol{e}_k))$ for $k = 1, 2$, and $\boldsymbol{\Lambda}_{e12}$ is a $r_{e,1} \times r_{e,2}$ rectangular diagonal matrix with the canonical correlations between $\boldsymbol{e}_k$'s on its main diagonal. Finally, as $p \to \infty$, we have

$$\lambda_1(\mathrm{cov}(\boldsymbol{e})) = \|\mathrm{cov}(\boldsymbol{e})\|_2 \le \|\boldsymbol{\Phi}_e\|_1 \cdot \max_{k \in \{1,2\}} \lambda_1(\mathrm{cov}(\boldsymbol{e}_k)) \le 2 \cdot \max_{k \in \{1,2\}} \lambda_1(\mathrm{cov}(\boldsymbol{e}_k)) < \infty.$$

We finish the proof of Proposition S.1. □

Proposition S.2. *Equation* (15) *is the unique solution to the problem in* (11) *subject to* (12)-(14).

Proof. Let $\theta(\cdot,\cdot)$ denote the angle between two elements in space $(\mathcal{L}_0^2, \mathrm{cov})$. Then, $\cos\theta(\cdot,\cdot) = \mathrm{corr}(\cdot,\cdot)$. Hereafter, we use these two operators exchangeably. Note that $\sum_{k=1}^2 \cos^2\theta(z_{k\ell}, z_{1\ell}) \geq 1$. If $w \perp \mathrm{span}(\{z_{1\ell}, z_{2\ell}\})$, then $\sum_{k=1}^2 \cos^2\theta(z_{k\ell}, w) = 0$, and thus such a w is not an optimal solution to the right-hand side of (11). When $w \not\perp \mathrm{span}(\{z_{1\ell}, z_{2\ell}\})$, since $\cos\theta(z_{k\ell}, w) = \cos\theta(z_{k\ell}, w_0)\cos\theta(w_0, w)$, where w_0 denotes the projection of w onto $\mathrm{span}(\{z_{1\ell}, z_{2\ell}\})$, we only need to consider $w \in \mathrm{span}(\{z_{1\ell}, z_{2\ell}\})$. Let $w = az_{1\ell} + bz_{2\ell}$ with $\mathrm{var}(w) = a^2 + b^2 + 2ab\rho_\ell = 1$. Then, we have

$$\begin{aligned}\sum_{k=1}^2 \mathrm{corr}^2(z_{k\ell}, w) &= (a + b\rho_\ell)^2 + (a\rho_\ell + b)^2 \\ &= a^2 + b^2 + 4ab\rho_\ell + \rho_\ell^2(a^2 + b^2) = 1 + 2ab\rho_\ell + \rho_\ell^2(1 - 2ab\rho_\ell) \\ &= 1 + \rho_\ell^2 + 2ab\rho_\ell(1 - \rho_\ell^2). \end{aligned} \tag{S.2}$$

We first consider $\rho_\ell \in (0,1)$. Equation (S.2) is maximized only when $ab \geq 0$. Without loss of generality, we assume a and b are nonnegative. Since $2ab \leq a^2 + b^2 = 1 - 2ab\rho_\ell$, the maximizer of ab satisfies $a = b = (2 + 2\rho_\ell)^{-1/2}$. Thus, $c_\ell \propto z_{1\ell} + z_{2\ell}$. Let $c_\ell = \alpha(z_{1\ell} + z_{2\ell})$. From (13), we have

$$\begin{aligned} 0 = \mathrm{corr}(d_{1\ell}, d_{2\ell}) &= \mathrm{corr}\left((1-\alpha)z_{1\ell} - \alpha z_{2\ell}, (1-\alpha)z_{2\ell} - \alpha z_{1\ell}\right) \\ &= 2(\rho_\ell + 1)\alpha^2 - 2(\rho_\ell + 1)\alpha + \rho_\ell. \end{aligned}$$

Hence, we obtain

$$\alpha = \frac{1}{2}\left(1 \pm \sqrt{\frac{1-\rho_\ell}{1+\rho_\ell}}\right).$$

It follows that

$$\mathrm{var}(c_\ell) = \frac{1}{2}\left(1 \pm \sqrt{1 - \frac{2}{1 + 1/\rho_\ell}}\right)^2 (1 + \rho_\ell). \tag{S.3}$$

To satisfy (14), c_ℓ must be the one in (15) when $\rho_\ell \in (0,1)$.

Now we consider the solution of c_ℓ when $\rho_\ell \in \{0,1\}$. By (S.3), for c_ℓ that is defined in (15) when $\rho_\ell \in (0,1)$, we have $\lim_{\rho_\ell \to 0^+} \mathrm{var}(c_\ell) = 0$ and $\lim_{\rho_\ell \to 1^-} \mathrm{var}(c_\ell) = 1$. Then by (14), when $\rho_\ell = 0$, then $\mathrm{var}(c_\ell) = 0$, and thus $c_\ell = 0$ which satisfies (15) as well as (11) and (13). We also obtain $\mathrm{var}(c_\ell) \geq 1$ when $\rho_\ell = 1$. Now consider $\rho_\ell = 1$. By (S.2), we have $\max_{w \in \mathcal{L}_0^2} \sum_{k=1}^2 \mathrm{corr}^2(z_{k\ell}, w) = 2$,

and thus $c_\ell \propto z_{1\ell} = z_{2\ell}$. If $\mathrm{var}(c_\ell) > 1$ or $c_\ell = -z_{1\ell} = -z_{2\ell}$, then $d_{1\ell} = d_{2\ell} \neq 0$ and $\mathrm{corr}(d_{1\ell}, d_{2\ell}) = 1$. Thus, to satisfy (13), $c_\ell = z_{1\ell} = z_{2\ell}$, which is equivalent to (15).

From the above, we have that c_ℓ must be the one in (15) when $\rho_\ell \in [0, 1]$. □

Lemma S.1. *When $n \to \infty$, if $a_n = O_P(b_n)$ holds on a given event $\mathcal{A}_n$ that has $\mathbb{P}(\mathcal{A}_n) \to 1$, then we have $a_n = O_P(b_n)$.*

Proof. By the given assumptions, for any $\varepsilon > 0$, there exist constants M_ε and N_ε such that $\mathbb{P}\left(|a_n| \leq M_\varepsilon b_n \middle| \mathcal{A}_n\right) > 1 - \varepsilon$ and $\mathbb{P}(\mathcal{A}_n) > 1 - \varepsilon$ for all $n \geq N_\varepsilon$. Then, $\mathbb{P}(|a_n| \leq M_\varepsilon b_n) \geq \mathbb{P}\left(|a_n| \leq M_\varepsilon b_n \middle| \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n) > (1-\varepsilon)^2 > 1 - 2\varepsilon$. Hence, we obtain $a_n = O_P(b_n)$. □

Proof of Theorem 1. First, we show $\mathrm{rank}(\boldsymbol{\Theta}) = r_{12}$. Since

$$\boldsymbol{x}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^*)\boldsymbol{z}_k^* = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{\Lambda}_k^{-1/2}\mathbf{V}_k^\top \boldsymbol{x}_k)\boldsymbol{z}_k^* = \mathbf{V}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{z}_k^*,$$

we obtain

$$\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\mathbf{V}_1 \boldsymbol{\Lambda}_1^{1/2} \boldsymbol{z}_1^*, \mathbf{V}_2 \boldsymbol{\Lambda}_2^{1/2} \boldsymbol{z}_2^*) = \mathbf{V}_1 \boldsymbol{\Lambda}_1^{1/2} \boldsymbol{\Theta} \boldsymbol{\Lambda}_2^{1/2} \mathbf{V}_2^\top. \tag{S.4}$$

Hence by $\mathrm{rank}(\mathbf{M}_1\mathbf{M}_2) \leq \min(\mathrm{rank}(\mathbf{M}_1), \mathrm{rank}(\mathbf{M}_2))$ for real matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we have $\mathrm{rank}(\boldsymbol{\Theta}) \geq r_{12}$. Again using the above inequality of the rank of matrix product, by $\boldsymbol{\Theta} = \boldsymbol{\Lambda}_1^{-1/2} \mathbf{V}_1^\top \boldsymbol{\Sigma}_{12} \mathbf{V}_2 \boldsymbol{\Lambda}_2^{-1/2}$, we have $\mathrm{rank}(\boldsymbol{\Theta}) \leq r_{12}$. Thus, $\mathrm{rank}(\boldsymbol{\Theta}) = r_{12}$.

Now temporarily replace the constraint $\ell \leq r_{12}$ by $\ell \leq r_{\min}$ for (10) . Let $\{\widetilde{z}_{1\ell}, \widetilde{z}_{2\ell}\}_{\ell=1}^{r_{\min}}$ be an arbitrary solution of (10). We will later see that $\mathrm{corr}(\widetilde{z}_{1\ell}, \widetilde{z}_{2\ell}) = 0$ for all $\ell > r_{12}$. Augment $(\widetilde{z}_{k1}, \ldots, \widetilde{z}_{k,r_{\min}})^\top$ with any $(r_k - r_{\min})$ standardized variables to be $\widetilde{\boldsymbol{z}}_k = (\widetilde{z}_{k1}, \ldots, \widetilde{z}_{k,r_k})^\top$ such that $\widetilde{\boldsymbol{z}}_k^\top$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$. Denote $\widetilde{\boldsymbol{\Theta}} = \mathrm{cov}(\widetilde{\boldsymbol{z}}_1, \widetilde{\boldsymbol{z}}_2)$. When $\ell = 1$ in (10), z_{11} must be proportional to the projection of z_{21} onto $\mathrm{span}(\boldsymbol{x}_1^\top)$. Hence, $z_{21} \perp \mathrm{span}(\boldsymbol{x}_1^\top) \backslash \mathrm{span}(z_{11})$, and $\widetilde{\boldsymbol{\Theta}}^{[2:r_1,1]}$ is a zero vector. Similarly, $\widetilde{\boldsymbol{\Theta}}^{[1,2:r_2]}$ is a zero vector. Using the same argument for $\ell = 2, \ldots, r_{\min}$ yields that the only nonzero entries of $\widetilde{\boldsymbol{\Theta}}$ are located on the diagonal of $\widetilde{\boldsymbol{\Theta}}^{[1:r_{\min},1:r_{\min}]}$. Note that there exists an orthogonal matrix $\mathbf{Q}_k$ such that $\widetilde{\boldsymbol{z}}_k = \mathbf{Q}_k \boldsymbol{z}_k^*$. Then, $\widetilde{\boldsymbol{\Theta}} = \mathbf{Q}_1 \boldsymbol{\Theta} \mathbf{Q}_2^\top$ has rank r_{12}. Hence, the only nonzero entries of $\widetilde{\boldsymbol{\Theta}}$ are the first r_{12} elements of its main diagonal. We thus only need $\ell \leq r_{12}$ in (10).

The proof is complete. □

Proof of Theorem 2. We only need to show the uniqueness of $\boldsymbol{c}_1$.

Let $\{\widetilde{\boldsymbol{z}}_k\}_{k=1,2}$ be another set of augmented standardized canonical variables. Then, there exists an orthogonal matrix $\mathbf{Q}_k$ such that $\widetilde{\boldsymbol{z}}_k = \mathbf{Q}_k \boldsymbol{z}_k$ with $\boldsymbol{z}_k = \boldsymbol{\Gamma}_k^\top \boldsymbol{x}_k$ defined in (18). By Theorem 1 and the fact that $\mathrm{cov}(\widetilde{\boldsymbol{z}}_1, \widetilde{\boldsymbol{z}}_2) = \mathbf{Q}_1 \boldsymbol{\Lambda}_\theta \mathbf{Q}_2^\top$ has the same singular values of $\boldsymbol{\Lambda}_\theta$, we have $\mathrm{cov}(\widetilde{\boldsymbol{z}}_1, \widetilde{\boldsymbol{z}}_2) = \mathbf{Q}_1 \boldsymbol{\Lambda}_\theta \mathbf{Q}_2^\top = \boldsymbol{\Lambda}_\theta$. Let m be the number of distinct nonzero singular values of $\boldsymbol{\Lambda}_\theta$. Then for $k = 1, 2$, we have $\mathbf{Q}_k = \mathrm{diag}(\mathbf{M}_{k1}, \ldots, \mathbf{M}_{km}, \mathbf{M}_{k,m+1})$, where $\mathbf{M}_{k\ell}, \ell \leq m$ is an orthogonal matrix with column dimension equal to the repetition number of the ℓ-th largest distinct nonzero singular value of $\boldsymbol{\Lambda}_\theta$, and $\mathbf{M}_{k,m+1}$ might be an empty matrix. By $\boldsymbol{\Lambda}_\theta \mathbf{Q}_2^\top = \mathbf{Q}_1^\top \boldsymbol{\Lambda}_\theta$, we obtain $\mathbf{M}_{1\ell} = \mathbf{M}_{2\ell}$ for all $\ell \leq m$.

By the expression of $\boldsymbol{c}_1$ in (19), we only need to show

$$\mathrm{cov}(\boldsymbol{x}_1, \widetilde{\boldsymbol{z}}_1^{[1:r_{12}]}) \mathbf{A}_C \sum_{k=1}^{2} \widetilde{\boldsymbol{z}}_k^{[1:r_{12}]} = \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{z}_1^{[1:r_{12}]}) \mathbf{A}_C \sum_{k=1}^{2} \boldsymbol{z}_k^{[1:r_{12}]}.$$

This is true because

$$\begin{aligned}
&\mathrm{cov}(\boldsymbol{x}_1, \widetilde{\boldsymbol{z}}_1^{[1:r_{12}]}) \mathbf{A}_C \sum_{k=1}^{2} \widetilde{\boldsymbol{z}}_k^{[1:r_{12}]} \\
&= \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{z}_1) (\mathbf{Q}_1^{[1:r_{12},:]})^\top \mathbf{A}_C \sum_{k=1}^{2} \mathbf{Q}_k^{[1:r_{12},:]} \boldsymbol{z}_k \\
&= \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{z}_1^{[1:r_{12}]}) \left(\mathrm{diag}(\mathbf{M}_{11}, \ldots, \mathbf{M}_{1m})\right)^\top \mathbf{A}_C \sum_{k=1}^{2} \mathrm{diag}(\mathbf{M}_{11}, \ldots, \mathbf{M}_{1m}) \boldsymbol{z}_k^{[1:r_{12}]} \\
&= \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{z}_1^{[1:r_{12}]}) \mathbf{A}_C \sum_{k=1}^{2} \boldsymbol{z}_k^{[1:r_{12}]}.
\end{aligned}$$

□

Proof of Theorem 3. Under Assumption 1, by the proof of Theorem 4.1 in Wang and Fan (2017) (see the bound for their Δ_{L1}), we have

$$\delta_{\boldsymbol{\Sigma}_k} := \|\widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_2 = O_P(\lambda_{k1}/\sqrt{n}). \tag{S.5}$$

From Weyl's inequality [see Theorem 3.3.16(c) in Horn and Johnson (1994)],

$$|\lambda_{k\ell} - \lambda_\ell(\boldsymbol{\Sigma}_k)| \leq \|\mathrm{cov}(\boldsymbol{e}_k)\|_2 \leq s_0 \quad \text{for} \quad 1 \leq \ell \leq r_k.$$

This implies

$$\lambda_{k\ell}/\lambda_\ell(\mathbf{\Sigma}_k) \to 1 \quad \text{for} \quad 1 \le \ell \le r_k. \tag{S.6}$$

Together with the assumption that $\lambda_{k1}/\lambda_{k,r_k}$ is bounded from above and below, we have

$$\lambda_\ell(\mathbf{\Sigma}_k) \asymp \lambda_m(\mathbf{\Sigma}_k) \quad \text{for} \quad 1 \le \ell, m \le r_k. \tag{S.7}$$

By Weyl's inequality, (S.5) and (S.6),

$$\left| \|\widetilde{\mathbf{X}}_k\|_2^2/n - \lambda_1(\mathbf{\Sigma}_k) \right| = \left| \lambda_1(\widehat{\mathbf{\Sigma}}_k) - \lambda_1(\mathbf{\Sigma}_k) \right| \le \delta_{\mathbf{\Sigma}_k} = O_P(\lambda_1(\mathbf{\Sigma}_k)/\sqrt{n}). \tag{S.8}$$

Thus,

$$\frac{\|\widetilde{\mathbf{X}}_k\|_2}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} = 1 + o_P(1). \tag{S.9}$$

Under Assumption 1, by the proof of Theorem C.1 in Wang and Fan (2017) (see the bound for their $\max_{i \le p} T^{-1} \sum_{t=1}^T |\widehat{u}_{it} - u_{it}|^2$), we have

$$\left\| \mathbf{X}_k - \mathbf{U}_{k1}^{[:,1:r_k]} \operatorname{diag}(\sigma_1(\mathbf{Y}_k), \ldots, \sigma_{r_k}(\mathbf{Y}_k)) (\mathbf{U}_{k2}^{[:,1:r_k]})^\top \right\|_F = O_P(\sqrt{p_k \log p_k}). \tag{S.10}$$

Also under Assumption 1, by the proof of Theorem 3.1 and the argument in the third paragraph on page 1355 in Wang and Fan (2017), for $1 \le \ell \le r_k$,

$$\left| \frac{\sigma_\ell^2(\mathbf{Y}_k)}{n\lambda_{k\ell}} - \frac{[\widehat{\sigma}_\ell^S(\mathbf{Y}_k)]^2}{n\lambda_{k\ell}} \right| = O_P\left(\frac{\tau_k p_k}{n\lambda_{k\ell}} \right) = O_P\left(\frac{p_k}{n\lambda_{k\ell}} + \frac{1}{n} \right) = O_P(1)$$

and

$$\frac{[\widehat{\sigma}_\ell^S(\mathbf{Y}_k)]^2}{n\lambda_{k\ell}} - 1 = O_P\left(\frac{1}{\sqrt{n}} \right).$$

Hence,

$$\left| \frac{\sigma_\ell(\mathbf{Y}_k)}{\sqrt{n\lambda_{k\ell}}} - \frac{\widehat{\sigma}_\ell^S(\mathbf{Y}_k)}{\sqrt{n\lambda_{k\ell}}} \right| = O_P\left(\left| \frac{\sigma_\ell^2(\mathbf{Y}_k)}{n\lambda_{k\ell}} - \frac{[\widehat{\sigma}_\ell^S(\mathbf{Y}_k)]^2}{n\lambda_{k\ell}} \right| \right) = O_P\left(\frac{p_k}{n\lambda_{k\ell}} + \frac{1}{n} \right). \tag{S.11}$$

By (S.11) and (S.10), we have

$$\begin{aligned}
\left\|\widetilde{\mathbf{X}}_k - \mathbf{X}_k\right\|_2 &\le \left\|\widetilde{\mathbf{X}}_k - \mathbf{X}_k\right\|_F \\
&\le \left\|\mathbf{U}_{k1}^{[:,1:r_k]} \operatorname{diag}\left(\widehat{\sigma}_1^S(\mathbf{Y}_k) - \sigma_1(\mathbf{Y}_k), \ldots, \widehat{\sigma}_{r_k}^S(\mathbf{Y}_k) - \sigma_{r_k}(\mathbf{Y}_k)\right) (\mathbf{U}_{k2}^{[:,1:r_k]})^\top\right\|_F \\
&\quad + \left\|\mathbf{X}_k - \mathbf{U}_{k1}^{[:,1:r_k]} \operatorname{diag}(\sigma_1(\mathbf{Y}_k), \ldots, \sigma_{r_k}(\mathbf{Y}_k))(\mathbf{U}_{k2}^{[:,1:r_k]})^\top\right\|_F \\
&\le \sqrt{r_k} \max_{1\le \ell \le r_k} \left|\widehat{\sigma}_\ell^S(\mathbf{Y}_k) - \sigma_\ell(\mathbf{Y}_k)\right| + \left\|\mathbf{X}_k - \sum_{\ell=1}^{r_k} \sigma_\ell(\mathbf{Y}_k)\mathbf{U}_{k1}^{[:,\ell]}(\mathbf{U}_{k2}^{[:,\ell]})^\top\right\|_F \\
&= O_P\left(\frac{p_k}{\sqrt{n\lambda_{k1}}} + \sqrt{\frac{\lambda_{k1}}{n}} + \sqrt{p_k \log p_k}\right) \\
&= O_P\left(\sqrt{\frac{\lambda_1(\mathbf{\Sigma}_k)}{n}} + \sqrt{p_k \log p_k}\right). \qquad \text{(S.12)}
\end{aligned}$$

It is easy to show $\mathbb{E}(f_{k\ell}^4)$ is upper bounded for all $1 \le \ell \le r_k$. Thus, $\operatorname{var}(f_{k\ell} f_{km})$ is upper bounded for all $1 \le \ell, m \le r_k$. Then from the central limit theorem, $\|\mathbf{F}_k\mathbf{F}_k^\top/n - \mathbf{I}_{r_k\times r_k}\|_{\max} = O_P(1/\sqrt{n})$. Hence,

$$\frac{1}{n}\|\mathbf{F}_k\|_F^2 = \text{trace}\left(\frac{1}{n}\mathbf{F}_k\mathbf{F}_k^\top\right) = r_k + O_P(1/\sqrt{n}).$$

Then by Lemma 1 in Lam and Fan (2009), the fact $\sigma_\ell(\mathbf{B}_k) = \lambda_\ell^{1/2}(\mathbf{\Sigma}_k)$ for $1 \le \ell \le r_k$, and (S.7), there exists a constant $\kappa_3 \in (0, 1]$ such that

$$\begin{aligned}
\kappa_3\sqrt{r_k} + o_P(1) &\le \frac{\sqrt{\lambda_{r_k}(\mathbf{\Sigma}_k)}\|\mathbf{F}_k\|_F}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} \\
&\le \frac{\|\mathbf{X}_k\|_F}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} \\
&= \frac{\|\mathbf{B}_k\mathbf{F}_k\|_F}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} \le \frac{\sqrt{\lambda_1(\mathbf{\Sigma}_k)}\|\mathbf{F}_k\|_F}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} = \sqrt{r_k} + o_P(1). \qquad \text{(S.13)}
\end{aligned}$$

By $\|\mathbf{X}_k\|_2 \le \|\mathbf{X}_k\|_F \le \sqrt{r_k}\|\mathbf{X}_k\|_2$, we have

$$\kappa_3 + o_P(1) \le \frac{\|\mathbf{X}_k\|_2}{\sqrt{n\lambda_1(\mathbf{\Sigma}_k)}} \le \sqrt{r_k} + o_P(1). \qquad \text{(S.14)}$$

From (S.12), (S.13), (S.9) and $\|\widetilde{\mathbf{X}}_k\|_F \le \sqrt{r_k}\|\widetilde{\mathbf{X}}_k\|_2$, we obtain

$$\begin{aligned}\delta_{\mathbf{X}_k,2} &:= \left\|\widetilde{\mathbf{X}}_k - \mathbf{X}_k\right\|_2 \le \delta_{\mathbf{X}_k,F} := \left\|\widetilde{\mathbf{X}}_k - \mathbf{X}_k\right\|_F \\ &= O_P\bigg(\min\Big\{\sqrt{\frac{\lambda_1(\mathbf{\Sigma}_k)}{n}} + \sqrt{p_k \log p_k}, \sqrt{n\lambda_1(\mathbf{\Sigma}_k)}\Big\}\bigg). \end{aligned} \tag{S.15}$$

From Weyl's inequality and (S.8), for all $1 \le \ell \le p_k$,

$$|\lambda_\ell(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell(\mathbf{\Sigma}_k)| \le \|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k\|_2 = O_P(\lambda_1(\mathbf{\Sigma}_k)/\sqrt{n}). \tag{S.16}$$

Then by (S.7),

$$\lambda_{r_k}(\widehat{\mathbf{\Sigma}}_k) \ge \lambda_{r_k}(\mathbf{\Sigma}_k) - |\lambda_{r_k}(\widehat{\mathbf{\Sigma}}_k) - \lambda_{r_k}(\mathbf{\Sigma}_k)| \ge (1 - o_P(1))\lambda_{r_k}(\mathbf{\Sigma}_k). \tag{S.17}$$

It follows that $\widetilde{r}_k = r_k$ with probability tending to 1 as $n \to \infty$. Due to Lemma S.1, we simply assume $\widetilde{r}_k = r_k$ in the rest of the proof.

By the mean value theorem and (S.16), uniformly for $\ell = 1, \ldots, r_k$, we have

$$|\lambda_\ell^{1/2}(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell^{1/2}(\mathbf{\Sigma}_k)| \le \frac{1}{2}[(1 - o_P(1))\lambda_{r_k}(\mathbf{\Sigma}_k)]^{-1/2}|\lambda_\ell(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell(\mathbf{\Sigma}_k)| = O_P(\lambda_1^{1/2}(\mathbf{\Sigma}_k)n^{-1/2}), \tag{S.18}$$

$$|\lambda_\ell^{-1/2}(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell^{-1/2}(\mathbf{\Sigma}_k)| \le \frac{1}{2}[(1 - o_P(1))\lambda_{r_k}(\mathbf{\Sigma}_k)]^{-3/2}|\lambda_\ell(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell(\mathbf{\Sigma}_k)| = O_P(\lambda_1^{-1/2}(\mathbf{\Sigma}_k)n^{-1/2}), \tag{S.19}$$

$$|\lambda_\ell^{-1}(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell^{-1}(\mathbf{\Sigma}_k)| \le [(1 - o_P(1))\lambda_{r_k}(\mathbf{\Sigma}_k)]^{-2}|\lambda_\ell(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell(\mathbf{\Sigma}_k)| = O_P(\lambda_1^{-1}(\mathbf{\Sigma}_k)n^{-1/2}).$$

By the uniqueness given in Theorem 2, we let $\mathbf{V}_k$ satisfy $(\widehat{\mathbf{V}}_k^{[:,\ell]})^\top \mathbf{V}_k^{[:,\ell]} \ge 0$ for all $k = 1, 2$ and $\ell = 1, \ldots, r_k$. By Corollary 1 in Yu et al. (2015), (S.6), (S.7) and $\min_{\ell \le r_k}(\lambda_{k\ell} - \lambda_{k,\ell+1})/\lambda_{k\ell} \ge \delta_0$, we have

$$\begin{aligned}\|\widehat{\mathbf{V}}_k - \mathbf{V}_k\|_F &= O\left(\|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k\|_2 \Big/ \min_{\ell \le r_k}\{\lambda_\ell(\mathbf{\Sigma}_k) - \lambda_{\ell+1}(\mathbf{\Sigma}_k)\}\right) \\ &= O_P(1/\sqrt{n}). \end{aligned} \tag{S.20}$$

Note that

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{B}}-\mathbf{A}\mathbf{B}\|_2 \begin{cases} = \|\widehat{\mathbf{A}}\widehat{\mathbf{B}}-\widehat{\mathbf{A}}\mathbf{B}+\widehat{\mathbf{A}}\mathbf{B}-\mathbf{A}\mathbf{B}\|_2 \le \|\widehat{\mathbf{A}}\|_2\|\widehat{\mathbf{B}}-\mathbf{B}\|_2+\|\mathbf{B}\|_2\|\widehat{\mathbf{A}}-\mathbf{A}\|_2, \\ = \|\widehat{\mathbf{B}}^\top\widehat{\mathbf{A}}^\top-\mathbf{B}^\top\mathbf{A}^\top\|_2 \le \|\widehat{\mathbf{B}}\|_2\|\widehat{\mathbf{A}}-\mathbf{A}\|_2+\|\mathbf{A}\|_2\|\widehat{\mathbf{B}}-\mathbf{B}\|_2. \end{cases} \tag{S.21}$$

Now we consider the error bounds for the columns of $\widehat{\mathbf{U}}_{\theta k}$. We first consider $\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\|_2$. By (S.21), (S.20), (S.19) and (S.7), we have

$$\begin{aligned} \delta_1^{(k)} &:= \|\widehat{\mathbf{\Lambda}}_k^{-1/2}\widehat{\mathbf{V}}_k^\top-\mathbf{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\|_2 \\ &\le \lambda_{r_k}^{-1/2}(\mathbf{\Sigma}_k)\|\mathbf{V}_k-\widehat{\mathbf{V}}_k\|_F+\|\widehat{\mathbf{V}}_k^\top\|_2\max_{1\le\ell\le r_k}|\lambda_\ell^{-1/2}(\widehat{\mathbf{\Sigma}}_k)-\lambda_\ell^{-1/2}(\mathbf{\Sigma}_k)| \\ &= O_P(\lambda_1^{-1/2}(\mathbf{\Sigma}_k)n^{-1/2}). \end{aligned} \tag{S.22}$$

By (S.21), (S.17), (S.22), (S.14), (S.15) and (S.7),

$$\begin{aligned} \delta_{Z_k} &:= \|\widehat{\mathbf{\Lambda}}_k^{-1/2}\widehat{\mathbf{V}}_k^\top\widetilde{\mathbf{X}}_k-\mathbf{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\mathbf{X}_k\|_2 \\ &\le \delta_1^{(k)}\|\mathbf{X}_k\|_2+\delta_{\mathbf{X}_k,2}\|\widehat{\mathbf{\Lambda}}_k^{-1/2}\|_2 \\ &= O_P(\delta_1^{(k)}\|\mathbf{X}_k\|_2+\delta_{\mathbf{X}_k,2}\lambda_{r_k}^{-1/2}(\mathbf{\Sigma}_k)) \\ &= O_P\left(\min\left\{1+\sqrt{p_k\lambda_1^{-1}(\mathbf{\Sigma}_k)\log p_k},\sqrt{n}\right\}\right). \end{aligned}$$

Define $\mathbf{Z}_k^*=\mathbf{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\mathbf{X}_k$. Then by (S.21), (S.17), (S.9), (S.14) and (S.7), we have

$$\begin{aligned} \left\|\frac{1}{n}\widehat{\mathbf{Z}}_1^*(\widehat{\mathbf{Z}}_2^*)^\top-\frac{1}{n}\mathbf{Z}_1^*(\mathbf{Z}_2^*)^\top\right\|_2 &\le \frac{1}{n}\left[\delta_{Z_1}\lambda_{r_2}^{-1/2}(\mathbf{\Sigma}_2)\|\mathbf{X}_2\|_2+\delta_{Z_2}\lambda_{r_1}^{-1/2}(\widehat{\mathbf{\Sigma}}_1)\|\widetilde{\mathbf{X}}_1\|_2\right] \\ &= O_P\left(\min\left\{\frac{1}{\sqrt{n}}+\sum_{k=1}^2\sqrt{\frac{p_k\log p_k}{n\lambda_1(\mathbf{\Sigma}_k)}},1\right\}\right). \end{aligned}$$

Since $\boldsymbol{z}_k^*$ and $\boldsymbol{f}_k$ are both orthonormal bases of $\mathrm{span}(\boldsymbol{x}_k^\top)$, $\boldsymbol{z}_k^*=\mathbf{Q}_{zf_k}\boldsymbol{f}_k$ with a $r_k\times r_k$ orthogonal matrix $\mathbf{Q}_{zfk}$. Since $\mathbb{E}(f_{k\ell}^4)$ is upper bounded for all $\ell\le r_k$ and $k\le 2$, $\mathrm{var}(f_{1\ell}f_{2m})$ is upper bounded for all $\ell<r_1$ and $m\le r_2$. Then, we can use the central limit theorem to obtain

$$\begin{aligned} &\left\|\frac{1}{n}\mathbf{Z}_1^*(\mathbf{Z}_2^*)^\top-\mathbf{\Theta}\right\|_2=\left\|\mathbf{Q}_{zf_1}\left(\frac{1}{n}\mathbf{F}_1\mathbf{F}_2^\top-\mathrm{cov}(\boldsymbol{f}_1,\boldsymbol{f}_2)\right)\mathbf{Q}_{zf_2}^\top\right\|_2 \\ &\le \|\mathbf{Q}_{zf_1}\|_2\left\|\frac{1}{n}\mathbf{F}_1\mathbf{F}_2^\top-\mathrm{cov}(\boldsymbol{f}_1,\boldsymbol{f}_2)\right\|_2\|\mathbf{Q}_{zf_2}^\top\|_2=\left\|\frac{1}{n}\mathbf{F}_1\mathbf{F}_2^\top-\mathbb{E}(\boldsymbol{f}_1\boldsymbol{f}_2^\top)\right\|_2=O_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Therefore,

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_2 &\le \left\|\frac{1}{n}\widehat{\mathbf{Z}}_1^*(\widehat{\mathbf{Z}}_2^*)^\top-\frac{1}{n}\mathbf{Z}_1^*(\mathbf{Z}_2^*)^\top\right\|_2+\left\|\frac{1}{n}\mathbf{Z}_1^*(\mathbf{Z}_2^*)^\top-\boldsymbol{\Theta}\right\|_2\\ &\lesssim_P \min\left\{\frac{1}{\sqrt{n}}+\sum_{k=1}^{2}\sqrt{\frac{p_k\log p_k}{n\lambda_1(\boldsymbol{\Sigma}_k)}},1\right\}\\ &= \delta_\theta.\end{aligned}$$

Here and also in the following text, for simplicity, we write $A\lesssim_P B$ if and only if $A=O_P(B)$. From Weyl's inequality, we have the bound for canonical correlation estimators

$$\max_{1\le\ell\le r_{\min}}|\sigma_\ell(\widehat{\boldsymbol{\Theta}})-\sigma_\ell(\boldsymbol{\Theta})|\le\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_2\lesssim_P\delta_\theta. \tag{S.23}$$

Using (S.21), we obtain

$$\begin{aligned}&\max\{\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Theta}}^\top-\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\|_2,\|\widehat{\boldsymbol{\Theta}}^\top\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\|_2\}\\ &\le(\|\widehat{\boldsymbol{\Theta}}\|_2+\|\boldsymbol{\Theta}\|_2)\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_2\\ &\lesssim_P(2\|\boldsymbol{\Theta}\|_2+\delta_\theta)\delta_\theta\lesssim_P(\sigma_1(\boldsymbol{\Theta})+\delta_\theta)\delta_\theta\\ &\lesssim_P\delta_\theta.\end{aligned}$$

Let $\{\widetilde{\mathbf{U}}_{\theta k}\}_{k=1,2}$ be one pair of orthogonal matrices such that $\boldsymbol{\Theta}=\widetilde{\mathbf{U}}_{\theta1}\boldsymbol{\Lambda}_\theta\widetilde{\mathbf{U}}_{\theta2}^\top$. Define $\sigma_{\theta,1}>\cdots>\sigma_{\theta,r_\theta}$ to be the distinct nonzero singular values of $\boldsymbol{\Theta}$, and $\sigma_{\theta,r_\theta+1}=0$. By Lemma 1 in Lam and Fan (2009) and Theorem 2 in Yu et al. (2015), there exists a matrix $\mathbf{Q}_k=\mathrm{diag}(\mathbf{Q}_{k1},\ldots,\mathbf{Q}_{kr_\theta})$, where $\mathbf{Q}_{k\ell}$ is an orthogonal matrix with column dimension equal to the repetition number of $\sigma_{\theta,\ell}$, such that

$$\begin{aligned}\|\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}-\widetilde{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}\mathbf{Q}_k\|_F&\le\|\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}\mathbf{Q}_k^\top-\widetilde{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}\|_F\|\mathbf{Q}_k\|_2\\ &\lesssim_P\min\left\{\delta_\theta\Big/\min_{1\le\ell\le r_\theta}\{\sigma_{\theta,\ell}^2-\sigma_{\theta,\ell+1}^2\},1\right\}\\ &\lesssim_P\delta_\theta.\end{aligned}$$

Note that $\widetilde{\mathbf{U}}_{\theta1}^{[:,1:r_{12}]}\mathbf{Q}_1\boldsymbol{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\mathbf{Q}_1^\top(\widetilde{\mathbf{U}}_{\theta2}^{[:,1:r_{12}]})^\top=\widetilde{\mathbf{U}}_{\theta1}^{[:,1:r_{12}]}\boldsymbol{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\widetilde{\mathbf{U}}_{\theta2}^{[:,1:r_{12}]})^\top=\boldsymbol{\Theta}$. By the uniqueness given in Theorem 2, we let $\mathbf{U}_{\theta k}=(\widetilde{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}\mathbf{Q}_1,\widetilde{\mathbf{U}}_{\theta k}^{[:,(r_{12}+1):r_k]})$. Define $\mathbf{U}_{\theta2}^\star=$

$(\widetilde{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]}\mathbf{Q}_2, \widetilde{\mathbf{U}}_{\theta 2}^{[:,(r_{12}+1):r_2]})$. We have

$$\|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_F \lesssim_P \delta_\theta \tag{S.24}$$

and

$$\|\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]}\|_F \lesssim_P \delta_\theta. \tag{S.25}$$

Then by (S.21) and (S.23),

$$\begin{aligned}
&\left\|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]}\widehat{\mathbf{\Lambda}}_\theta^{[1:r_{12},1:r_{12}]}(\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]})^\top - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\right\|_2 \\
&\le \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]}\widehat{\mathbf{\Lambda}}_\theta^{[1:r_{12},1:r_{12}]} - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\|_2\|(\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]})^\top\|_2 \\
&\quad + \|\mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\|_2\|(\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]})^\top - (\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_2 \\
&\le \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_2\|\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\|_2 + \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]}\|_2\|\widehat{\mathbf{\Lambda}}_\theta^{[1:r_{12},1:r_{12}]} - \mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\|_2 \\
&\quad + \|\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}\|_2\|(\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]})^\top - (\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_2 \\
&\lesssim_P \sigma_1(\mathbf{\Theta})\delta_\theta + \delta_\theta \\
&\lesssim_P \delta_\theta.
\end{aligned}$$

By the above inequality, the inequality $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \lesssim_P \delta_\theta$, and the triangular inequality of matrix norms, we have

$$\|\mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_2 \lesssim_P \delta_\theta.$$

It follows that

$$\begin{aligned}
&\|\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_F \\
&\le \sqrt{r_{12}}\|\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_2 \\
&\le \sqrt{r_{12}}\|(\mathbf{U}_{\theta 1}^{[:,1:r_{12}]})^\top\|_2\|\mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_2 \\
&\lesssim_P \delta_\theta.
\end{aligned} \tag{S.26}$$

Define $\mathbf{\Gamma}_2^\star = \mathbf{V}_2\mathbf{\Lambda}_2^{-1/2}\mathbf{U}_{\theta 2}^\star$. Note that

$$\begin{aligned}
\mathbf{C}_1 = \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_1^{[:,1:r_{12}]})^\top\mathbf{X}_1 + \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\mathbf{X}_2 \\
+ \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{[:,1:r_{12}]} - \mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\mathbf{X}_2.
\end{aligned} \tag{S.27}$$

Let $\widehat{\mathbf{\Gamma}}_k = \widehat{\mathbf{V}}_k \widehat{\mathbf{\Lambda}}_k^{-1/2} \widehat{\mathbf{U}}_{\theta k}$ for $k = 1, 2$. By (S.21), (S.22) and (S.24), we have

$$\begin{aligned}
&\|\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} - \mathbf{\Gamma}_1^{[:,1:r_{12}]}\|_2 \\
&= \|\widehat{\mathbf{V}}_1 \widehat{\mathbf{\Lambda}}_1^{-1/2} \widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{V}_1 \mathbf{\Lambda}_1^{-1/2} \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_2 \\
&\le \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]}\|_2 \|\widehat{\mathbf{V}}_1 \widehat{\mathbf{\Lambda}}_1^{-1/2} - \mathbf{V}_1 \mathbf{\Lambda}_1^{-1/2}\|_2 + \|\mathbf{V}_1 \mathbf{\Lambda}_1^{-1/2}\|_2 \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_F \\
&\lesssim_P \lambda_1^{-1/2}(\mathbf{\Sigma}_1) n^{-1/2} + \lambda_1^{-1/2}(\mathbf{\Sigma}_1) \delta_\theta \\
&=: \delta_\gamma^{(1)}
\end{aligned} \tag{S.28}$$

and similarly,

$$\begin{aligned}
&\|\widehat{\mathbf{\Gamma}}_2^{[:,1:r_{12}]} - \mathbf{\Gamma}_2^{\star[:,1:r_{12}]}\|_2 \\
&= \|\widehat{\mathbf{V}}_2 \widehat{\mathbf{\Lambda}}_2^{-1/2} \widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{V}_2 \mathbf{\Lambda}_2^{-1/2} \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]}\|_2 \\
&\lesssim_P \lambda_1^{-1/2}(\mathbf{\Sigma}_2) n^{-1/2} + \lambda_1^{-1/2}(\mathbf{\Sigma}_2) \delta_\theta \\
&=: \delta_\gamma^{(2)}.
\end{aligned} \tag{S.29}$$

By (S.21), (S.20) and (S.18),

$$\begin{aligned}
&\|\widehat{\mathbf{V}}_1 \widehat{\mathbf{\Lambda}}_1^{1/2} - \mathbf{V}_1 \mathbf{\Lambda}_1^{1/2}\|_2 \\
&\le \lambda_1^{1/2}(\mathbf{\Sigma}_1) \|\mathbf{V}_1 - \widehat{\mathbf{V}}_1\|_F + \|\widehat{\mathbf{V}}_1^\top\|_2 \max_{1 \le \ell \le r_1} |\lambda_\ell^{1/2}(\widehat{\mathbf{\Sigma}}_1) - \lambda_\ell^{1/2}(\mathbf{\Sigma}_1)| \\
&= O_P(\lambda_1^{1/2}(\mathbf{\Sigma}_1) n^{-1/2}).
\end{aligned} \tag{S.30}$$

Then by (S.24),

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1 \widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} - \mathbf{\Sigma}_1 \mathbf{\Gamma}_1^{[:,1:r_{12}]}\|_2 \\
&= \|\widehat{\mathbf{V}}_1 \widehat{\mathbf{\Lambda}}_1^{1/2} \widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{V}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_2 \\
&\le \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]}\|_2 \|\widehat{\mathbf{V}}_1 \widehat{\mathbf{\Lambda}}_1^{1/2} - \mathbf{V}_1 \mathbf{\Lambda}_1^{1/2}\|_2 + \|\mathbf{V}_1 \mathbf{\Lambda}_1^{1/2}\|_2 \|\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_2 \\
&\lesssim_P \lambda_1^{1/2}(\mathbf{\Sigma}_1) n^{-1/2} + \lambda_1^{1/2}(\mathbf{\Sigma}_1) \delta_\theta \\
&=: \delta_2.
\end{aligned} \tag{S.31}$$

Now consider the error bound for $\widehat{\mathbf{A}}_C^{(r)}$. Let $f(x) = \frac{1-x}{1+x}$. We notice that the derivative of $f^{\frac{1}{2}}(x)$ is unbound near $x = 1$. Thus, rather than using the mean value theorem directly for

$|f^{\frac{1}{2}}(\sigma_\ell(\widehat{\mathbf{\Theta}})) - f^{\frac{1}{2}}(\sigma_\ell(\mathbf{\Theta}))|$, we use the following technique:

$$\begin{aligned}
\left|f^{\frac{1}{2}}(\sigma_\ell(\widehat{\mathbf{\Theta}})) - f^{\frac{1}{2}}(\sigma_\ell(\mathbf{\Theta}))\right|^2 &\le \left|f^{\frac{1}{2}}(\sigma_\ell(\widehat{\mathbf{\Theta}})) - f^{\frac{1}{2}}(\sigma_\ell(\mathbf{\Theta}))\right| \left|f^{\frac{1}{2}}(\sigma_\ell(\widehat{\mathbf{\Theta}})) + f^{\frac{1}{2}}(\sigma_\ell(\mathbf{\Theta}))\right| \\
&= \left|f(\sigma_\ell(\widehat{\mathbf{\Theta}})) - f(\sigma_\ell(\mathbf{\Theta}))\right| \\
&\le \sup_{0\le x\le 1} |f'(x)| \left|\sigma_\ell(\widehat{\mathbf{\Theta}}) - \sigma_\ell(\mathbf{\Theta})\right| \\
&\le \sup_{0\le x\le 1} \frac{2}{(x+1)^2} \left|\sigma_\ell(\widehat{\mathbf{\Theta}}) - \sigma_\ell(\mathbf{\Theta})\right| \\
&\lesssim_P \delta_\theta,
\end{aligned}$$

where the last inequality holds uniformly for all $\ell = 1, \ldots, r_{\min}$ due to (S.23). Hence,

$$\max_{1\le\ell\le r_{\min}} |\widehat{a}_\ell - a_\ell| \lesssim_P \delta_\theta^{1/2}. \tag{S.32}$$

From (S.21), (S.31), and (S.32),

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1 \widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} \widehat{\mathbf{A}}_C^{(r_{12})} - \mathbf{\Sigma}_1 \mathbf{\Gamma}_1^{[:,1:r_{12}]} \mathbf{A}_C\|_2 \\
&\le \|\widehat{\mathbf{A}}_C^{(r_{12})}\|_2 \|\widehat{\mathbf{\Sigma}}_1 \widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} - \mathbf{\Sigma}_1 \mathbf{\Gamma}_1^{[:,1:r_{12}]}\|_2 + \|\mathbf{V}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\|_2 \|\widehat{\mathbf{A}}_C^{(r_{12})} - \mathbf{A}_C\|_2 \\
&\lesssim_P \delta_2 + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\delta_\theta^{1/2} \\
&=: \delta_{\sigma\gamma a}.
\end{aligned} \tag{S.33}$$

Then by (S.21), (S.28), (S.17) and (S.7),

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1 \widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} \widehat{\mathbf{A}}_C^{(r_{12})} (\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]})^\top - \mathbf{\Sigma}_1 \mathbf{\Gamma}_1^{[:,1:r_{12}]} \mathbf{A}_C (\mathbf{\Gamma}_1^{[:,1:r_{12}]})^\top\|_2 \\
&\le \|(\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]})^\top \widehat{\mathbf{\Lambda}}_1^{-1/2} \widehat{\mathbf{V}}_1^\top\|_2 \|\widehat{\mathbf{\Sigma}}_1 \widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]} \widehat{\mathbf{A}}_C^{(r_{12})} - \mathbf{\Sigma}_1 \mathbf{\Gamma}_1^{[:,1:r_{12}]} \mathbf{A}_C\|_2 \\
&\qquad + \|\mathbf{V}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{U}_{\theta 1}^{[:,1:r_{12}]} \mathbf{A}_C\|_2 \|(\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]})^\top - (\mathbf{\Gamma}_1^{[:,1:r_{12}]})^\top\|_2 \\
&\lesssim_P \lambda_{r_1}^{-1/2}(\widehat{\mathbf{\Sigma}}_1)\delta_{\sigma\gamma a} + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\delta_\gamma^{(1)} \\
&\lesssim_P \lambda_{r_1}^{-1/2}(\mathbf{\Sigma}_1)\delta_{\sigma\gamma a} + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\delta_\gamma^{(1)} \\
&\lesssim_P \delta_\theta + \delta_\theta^{1/2} \\
&\lesssim_P \delta_\theta^{1/2} \\
&=: \delta_{1,1}.
\end{aligned} \tag{S.34}$$

Similarly, from (S.29),

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]}\widehat{\mathbf{A}}_C^{(r_{12})}(\widehat{\mathbf{\Gamma}}_2^{[:,1:r_{12}]})^\top - \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\|_2 \\
&\lesssim_P \lambda_{r_2}^{-1/2}(\mathbf{\Sigma}_2)\delta_{\sigma\gamma a} + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\delta_\gamma^{(2)} \\
&\lesssim_P \delta_{1,1}\lambda_1^{1/2}(\mathbf{\Sigma}_1)\lambda_1^{-1/2}(\mathbf{\Sigma}_2) \\
&=: \delta_{1,2}. \qquad \text{(S.35)}
\end{aligned}$$

By (S.21) and its variant under the Frobenius norm following from Lemma 1 in Lam and Fan (2009),

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]}\widehat{\mathbf{A}}_C^{(r_{12})}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]})^\top\widetilde{\mathbf{X}}_1 - \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_1^{[:,1:r_{12}]})^\top\mathbf{X}_1\|_{(\cdot)} \\
&\le \|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]}\widehat{\mathbf{A}}_C^{(r_{12})}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]})^\top - \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_1^{[:,1:r_{12}]})^\top\|_2\|\mathbf{X}_1\|_{(\cdot)} \\
&\quad + \|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]}\widehat{\mathbf{A}}_C^{(r_{12})}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]})^\top\|_2\|\widetilde{\mathbf{X}}_1 - \mathbf{X}_1\|_{(\cdot)} \\
&\lesssim_P \delta_{1,1}\|\mathbf{X}_1\|_{(\cdot)} + \frac{1}{2}\lambda_1^{1/2}(\widehat{\mathbf{\Sigma}}_1)\lambda_{r_1}^{-1/2}(\widehat{\mathbf{\Sigma}}_1)\delta_{\mathbf{X}_1,(\cdot)} \\
&\lesssim_P \delta_{1,1}\|\mathbf{X}_1\|_{(\cdot)} + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\lambda_{r_1}^{-1/2}(\mathbf{\Sigma}_1)\delta_{\mathbf{X}_1,(\cdot)} \\
&=: \delta_{C,(\cdot)}^{(1)} \qquad \text{(S.36)}
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r_{12}]}\widehat{\mathbf{A}}_C^{(r_{12})}(\widehat{\mathbf{\Gamma}}_2^{[:,1:r_{12}]})^\top\widetilde{\mathbf{X}}_2 - \mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\mathbf{X}_2\|_{(\cdot)} \\
&\lesssim_P \delta_{1,2}\|\mathbf{X}_2\|_{(\cdot)} + \lambda_1^{1/2}(\mathbf{\Sigma}_1)\lambda_{r_2}^{-1/2}(\mathbf{\Sigma}_2)\delta_{\mathbf{X}_2,(\cdot)} \\
&=: \delta_{C,(\cdot)}^{(2)}. \qquad \text{(S.37)}
\end{aligned}$$

By the fact that $1 - \sqrt{\frac{1-x}{1+x}} \le 1 - \frac{1-x}{1+x} \le 2x$ for $x \in [0,1]$ and inequality (S.26), we have

$$\|\mathbf{A}_C(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_F \le \|\mathbf{\Lambda}_\theta^{[1:r_{12},1:r_{12}]}(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]} - \mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\|_F \lesssim_P \delta_\theta.$$

It follows that

$$\begin{aligned}
&\|\mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{[:,1:r_{12}]}-\mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\mathbf{X}_2\|_{(\cdot)}\\
&\le\|\mathbf{V}_1\mathbf{\Lambda}_1^{1/2}\mathbf{U}_{\theta 1}^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{U}_{\theta 2}^{[:,1:r_{12}]}-\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]})^\top\mathbf{\Lambda}_2^{-1/2}\mathbf{V}_2^\top\|_2\|\mathbf{X}_2\|_{(\cdot)}\\
&\lesssim_P\lambda_1^{1/2}(\mathbf{\Sigma}_1)\lambda_{r_2}^{-1/2}(\mathbf{\Sigma}_2)\|\mathbf{X}_2\|_{(\cdot)}\delta_\theta\\
&\lesssim_P\delta_{1,2}\|\mathbf{X}_2\|_{(\cdot)}.
\end{aligned}\tag{S.38}$$

By the definition of $\widehat{\mathbf{C}}_1$, (S.27), (S.36), (S.37) and (S.38), we obtain

$$\|\widehat{\mathbf{C}}_1-\mathbf{C}_1\|_{(\cdot)}\lesssim_P\delta_{C,(\cdot)}^{(1)}+\delta_{C,(\cdot)}^{(2)}.\tag{S.39}$$

Together with (S.15), (S.13) and (S.14), we obtain the claimed bound for $\|\widehat{\mathbf{C}}_1-\mathbf{C}_1\|_{(\cdot)}/\|\mathbf{C}_1\|_{(\cdot)}$.

Now consider the relative error bound for $\widehat{\mathbf{D}}_1$. Write $\widehat{\mathbf{C}}_1^{(r)}$ equivalently by

$$\widehat{\mathbf{C}}_1^{(r)}=\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)}\sum_{k=1}^{2}(\widehat{\mathbf{\Gamma}}_k^{[:,1:r]})^\top\widetilde{\mathbf{X}}_k.$$

Note that $\widetilde{\mathbf{X}}_1=\widehat{\mathbf{C}}_1^{(\widetilde{r}_{12})}+\widehat{\mathbf{D}}_1$. We have

$$\|\widehat{\mathbf{D}}_1-\mathbf{D}_1\|_{(\cdot)}\le\|\widehat{\mathbf{C}}_1^{(\widetilde{r}_{12})}-\mathbf{C}_1\|_{(\cdot)}+\|\widetilde{\mathbf{X}}_1-\mathbf{X}_1\|_{(\cdot)}.\tag{S.40}$$

When $\widetilde{r}_{12}\le r_{12}$, $\widehat{\mathbf{C}}_1=\widehat{\mathbf{C}}_1^{(\widetilde{r}_{12})}$ and thus $\|\widehat{\mathbf{D}}_1-\mathbf{D}_1\|_{(\cdot)}\le\|\widehat{\mathbf{C}}_1-\mathbf{C}_1\|_{(\cdot)}+\|\widetilde{\mathbf{X}}_1-\mathbf{X}_1\|_{(\cdot)}$, immediately leading to the bound for $\|\widehat{\mathbf{D}}_1-\mathbf{D}_1\|_{(\cdot)}/\|\mathbf{D}_1\|_{(\cdot)}$. Now consider the case when $\widetilde{r}_{12}>r_{12}$. Let $r\in(r_{12},r_{\min}]$. We first look at $\|\widehat{\mathbf{C}}_1^{(r)}-\mathbf{C}_1\|_{(\cdot)}$. Define $\widetilde{\mathbf{\Gamma}}_k^{(r)}=\mathbf{V}_k\mathbf{\Lambda}_k^{-1/2}(\mathbf{U}_{\theta k}^{[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta k}^{[:,(r_{12}+1):r]})$ and $\widetilde{\mathbf{\Gamma}}_2^{\star(r)}=\mathbf{V}_2\mathbf{\Lambda}_2^{-1/2}(\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta 2}^{[:,(r_{12}+1):r]})$. We have

$$\begin{aligned}
\mathbf{C}_1=&\mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_1^{(r)})^\top\mathbf{X}_1+\mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_2^{\star(r)})^\top\mathbf{X}_2\\
&+\mathbf{\Sigma}_1\mathbf{\Gamma}_1^{[:,1:r_{12}]}\mathbf{A}_C(\mathbf{\Gamma}_2^{[:,1:r_{12}]}-\mathbf{\Gamma}_2^{\star[:,1:r_{12}]})^\top\mathbf{X}_2
\end{aligned}\tag{S.41}$$

with $\mathbf{A}_C^{(r)}\coloneqq\mathrm{diag}(a_1,\ldots,a_r)$ and $a_\ell=0$ for $\ell>r_{12}$. By (S.24) and (S.25),

$$\begin{aligned}
\max\Big\{&\|(\widehat{\mathbf{U}}_{\theta 1}^{[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta 1}^{[:,(r_{12}+1):r]})-(\mathbf{U}_{\theta 1}^{[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta 1}^{[:,(r_{12}+1):r]})\|_F,\\
&\|(\widehat{\mathbf{U}}_{\theta 2}^{[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta 2}^{[:,(r_{12}+1):r]})-(\mathbf{U}_{\theta 2}^{\star[:,1:r_{12}]},\widehat{\mathbf{U}}_{\theta 2}^{[:,(r_{12}+1):r]})\|_F\Big\}\lesssim_P\delta_\theta.
\end{aligned}\tag{S.42}$$

Then following the proof lines for (S.28), (S.29), (S.31) and (S.33)-(S.35), we can obtain

$$\|\widehat{\mathbf{\Gamma}}_1^{[:,1:r]} - \widetilde{\mathbf{\Gamma}}_1^{(r)}\|_2 \lesssim_P \delta_\gamma^{(1)},$$

$$\|\widehat{\mathbf{\Gamma}}_2^{[:,1:r]} - \widetilde{\mathbf{\Gamma}}_2^{\star(r)}\|_2 \lesssim_P \delta_\gamma^{(2)},$$

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]} - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\|_2 \lesssim_P \delta_2,$$

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)} - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}\|_2 \lesssim_P \delta_{a\gamma a},$$

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r]})^\top - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_1^{(r)})^\top\|_2 \lesssim_P \delta_{1,1}$$

and

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r]})^\top - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_2^{\star(r)})^\top\|_2 \lesssim_P \delta_{1,2}.$$

Following the derivation of (S.36) and (S.37), we can obtain

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)}(\widehat{\mathbf{\Gamma}}_1^{[:,1:r]})^\top\widetilde{\mathbf{X}}_1 - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_1^{(r)})^\top\mathbf{X}_1\|_{(\cdot)} \lesssim_P \delta_{C,(\cdot)}^{(1)}$$

and

$$\|\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{\Gamma}}_1^{[:,1:r]}\widehat{\mathbf{A}}_C^{(r)}(\widehat{\mathbf{\Gamma}}_2^{[:,1:r]})^\top\widetilde{\mathbf{X}}_2 - \mathbf{\Sigma}_1\widetilde{\mathbf{\Gamma}}_1^{(r)}\mathbf{A}_C^{(r)}(\widetilde{\mathbf{\Gamma}}_2^{\star(r)})^\top\mathbf{X}_2\|_{(\cdot)} \lesssim_P \delta_{C,(\cdot)}^{(2)}.$$

By the above two inequalities, (S.38), the definition of $\widehat{\mathbf{C}}_1^{(r)}$, and (S.41), we obtain

$$\|\widehat{\mathbf{C}}_1^{(r)} - \mathbf{C}_1\|_{(\cdot)} \lesssim_P \delta_{C,(\cdot)}^{(1)} + \delta_{C,(\cdot)}^{(2)}, \tag{S.43}$$

which has the same bound for $\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_{(\cdot)}$ given in (S.39). Then using (S.40) gives

$$\|\widehat{\mathbf{D}}_1 - \mathbf{D}_1\|_{(\cdot)} \lesssim_P \delta_{C,(\cdot)}^{(1)} + \delta_{C,(\cdot)}^{(2)} \tag{S.44}$$

and the claimed bound for $\|\widehat{\mathbf{D}}_1 - \mathbf{D}_1\|_{(\cdot)}/\|\mathbf{D}_1\|_{(\cdot)}$ in the theorem.

The relative error bound for $\widehat{\mathbf{X}}_1$ immediately follows from

$$\|\widehat{\mathbf{X}}_1 - \mathbf{X}_1\|_{(\cdot)} \le \|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_{(\cdot)} + \|\widehat{\mathbf{D}}_1 - \mathbf{D}_1\|_{(\cdot)} \lesssim_P \delta_{C,(\cdot)}^{(1)} + \delta_{C,(\cdot)}^{(2)}.$$

Similarly, we can obtain the bounds for estimated matrices of the second dataset. □

Proof of Corollary 1. For $k = 1, 2$, since $\breve{r}_k \xrightarrow{P} r_k$ and $\breve{r}_k$ is an integer, we have $\mathbb{P}(\breve{r}_k = r_k) \to 1$ as $n \to \infty$. Due to Lemma S.1, in this proof we simply assume $\breve{r}_k = r_k$. Hence, we only need to prove the relative error bounds for $\widehat{\mathbf{C}}_k^{(r)}$ and $\widehat{\mathbf{X}}_k^{(r)}$, and refer the other two bounds to Theorem 3.

When $\widetilde{r}_{12} \le r_{12}$, we have $\widetilde{r}_{12} = \min(r_{12}, \widetilde{r}_{12}) \le r \le r_{\min}$ and thus $\widehat{\mathbf{C}}_k^{(r)} = \widehat{\mathbf{C}}_k^{(\widetilde{r}_{12})} = \widehat{\mathbf{C}}_k$. Then, the result stated in the corollary has been given in Theorem 3. On the other hand, when $\widetilde{r}_{12} > r_{12}$, we have $r_{12} = \min(r_{12}, \widetilde{r}_{12}) \le r \le r_{\min}$. Then by (S.43), we can immediately obtain the claimed result in the corollary. □

S.2 Additional Simulations

We consider Setups 1* and 2* which have the same settings as those in Setups 1 and 2, respectively, except for the noise covariance matrices $\mathrm{cov}(\boldsymbol{e}_k) = (0.7^{|i-j|}\sigma_e^2)_{1\le i,j\le p_k}$, $k = 1, 2$. Note that $\lambda_1(\mathrm{cov}(\boldsymbol{e}_k)) \in (5.62\sigma_e^2, 5.67\sigma_e^2)$ for $100 \le p_k \le 1500$. Especially when $\sigma_e^2 = 16$, $\lambda_1(\mathrm{cov}(\boldsymbol{e}_k)) \approx 90$ is quite close to 100 that is the minimum nonzero eigenvalue of $\boldsymbol{\Sigma}_k$, resulting in challenging cases for estimation (see conditions (I), (II) and (V) in Assumption 1, and also the result in (S.6)). The finite sample performance of our D-CCA estimates shown in Table S.1 and Figures S.1 and S.2 is similar to that in Table 1 and Figures 3 and 4.

Table S.1: Averages (standard errors) of D-CCA estimates for the first canonical angle/correlation.

(p_1, σ_e^2)	$\theta_1 = 0°/\rho_1 = 1$	$\theta_1 = 45°/\rho_1 = 0.707$	$\theta_1 = 60°/\rho_1 = 0.5$	$\theta_1 = 75°/\rho_1 = 0.259$
		Setup 1*		
$(100, 1)$	$4.15°(0.24°)/0.997(0.000)$	$44.7°(2.39°)/0.710(0.029)$	$59.4°(2.89°)/0.509(0.043)$	$73.5°(3.08°)/0.283(0.051)$
$(600, 1)$	$3.65°(0.22°)/0.998(0.000)$	$44.7°(2.39°)/0.710(0.029)$	$59.4°(2.89°)/0.509(0.043)$	$73.5°(3.08°)/0.283(0.051)$
$(900, 1)$	$3.65°(0.22°)/0.998(0.000)$	$44.7°(2.39°)/0.710(0.029)$	$59.4°(2.89°)/0.509(0.043)$	$73.5°(3.07°)/0.283(0.051)$
$(1500, 1)$	$3.64°(0.22°)/0.998(0.000)$	$44.7°(2.38°)/0.710(0.029)$	$59.4°(2.89°)/0.509(0.043)$	$73.5°(3.08°)/0.283(0.051)$
$(900, 0.01)$	$0.36°(0.02°)/1.000(0.000)$	$44.6°(2.38°)/0.712(0.029)$	$59.3°(2.89°)/0.510(0.043)$	$73.5°(3.08°)/0.284(0.051)$
$(900, 1)$	$3.65°(0.22°)/0.998(0.000)$	$44.7°(2.39°)/0.710(0.029)$	$59.4°(2.89°)/0.509(0.043)$	$73.5°(3.07°)/0.283(0.051)$
$(900, 9)$	$12.1°(0.81°)/0.978(0.003)$	$45.9°(2.46°)/0.696(0.031)$	$60.1°(2.92°)/0.499(0.044)$	$73.9°(3.05°)/0.277(0.051)$
$(900, 16)$	$17.6°(1.28°)/0.953(0.007)$	$47.4°(2.57°)/0.676(0.033)$	$61.1°(2.98°)/0.482(0.046)$	$74.9°(3.16°)/0.260(0.053)$
		Setup 2*		
$(100, 1)$	$3.97°(0.24°)/0.998(0.000)$	$44.5°(2.36°)/0.712(0.029)$	$59.0°(2.82°)/0.514(0.042)$	$72.7°(2.88°)/0.296(0.048)$
$(600, 1)$	$3.72°(0.23°)/0.998(0.000)$	$44.5°(2.36°)/0.712(0.029)$	$59.0°(2.82°)/0.514(0.042)$	$72.7°(2.88°)/0.296(0.048)$
$(900, 1)$	$3.72°(0.22°)/0.998(0.000)$	$44.5°(2.36°)/0.712(0.029)$	$59.0°(2.83°)/0.514(0.042)$	$72.7°(2.88°)/0.297(0.048)$
$(1500, 1)$	$3.72°(0.23°)/0.998(0.000)$	$44.5°(2.37°)/0.712(0.029)$	$59.0°(2.84°)/0.514(0.043)$	$72.7°(2.89°)/0.296(0.048)$
$(900, 0.01)$	$0.37°(0.02°)/1.000(0.000)$	$44.4°(2.35°)/0.714(0.029)$	$59.0°(2.82°)/0.515(0.042)$	$72.7°(2.89°)/0.297(0.048)$
$(900, 1)$	$3.72°(0.22°)/0.998(0.000)$	$44.5°(2.36°)/0.712(0.029)$	$59.0°(2.83°)/0.514(0.042)$	$72.7°(2.88°)/0.297(0.048)$
$(900, 9)$	$12.0°(0.79°)/0.978(0.003)$	$45.6°(2.42°)/0.698(0.030)$	$59.7°(2.86°)/0.505(0.043)$	$73.0°(2.89°)/0.292(0.048)$
$(900, 16)$	$17.3°(2.73°)/0.954(0.030)$	$47.1°(2.53°)/0.680(0.032)$	$60.7°(2.90°)/0.488(0.044)$	$74.0°(2.97°)/0.275(0.050)$

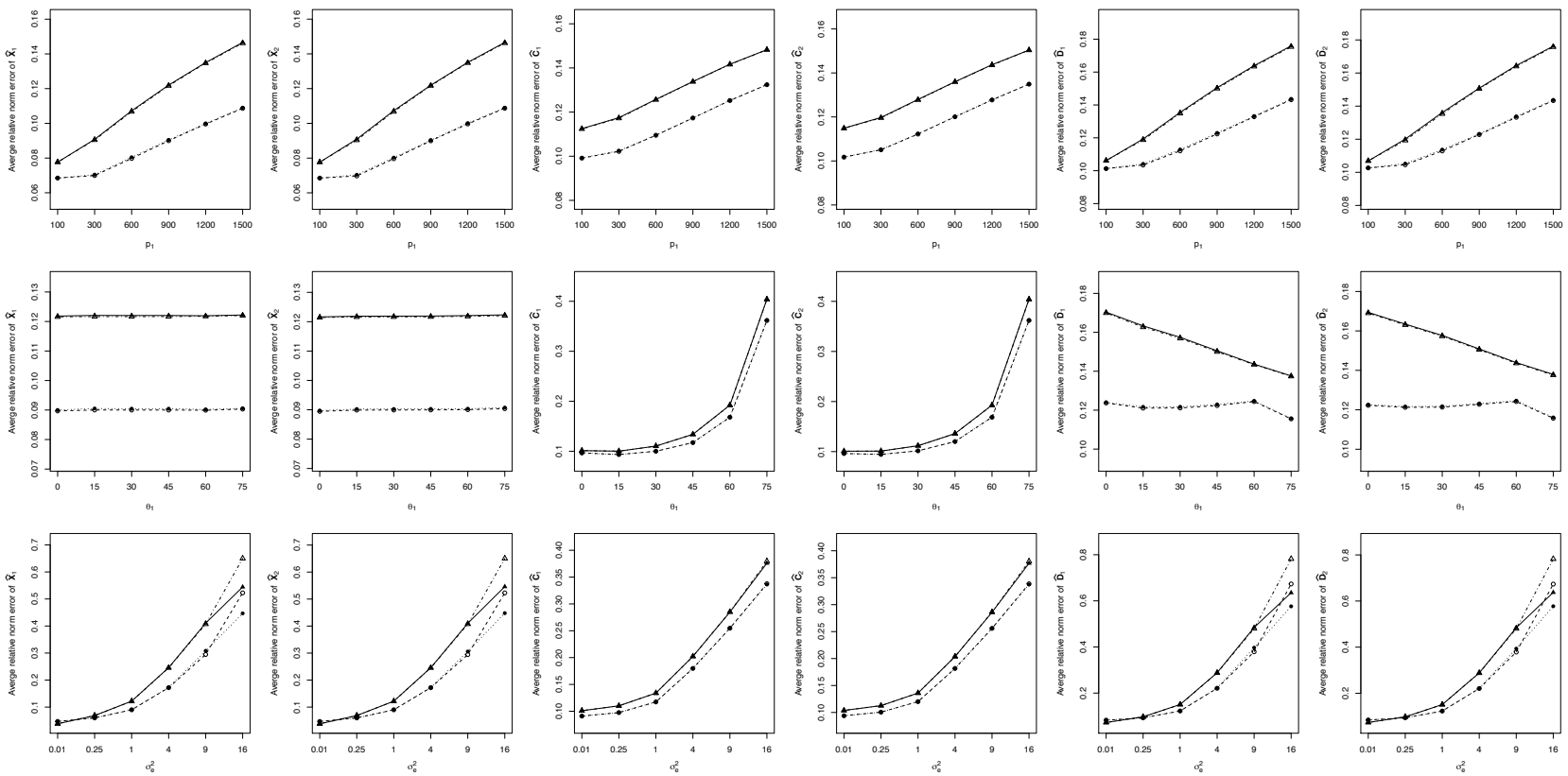


Figure S.1: Average relative errors of D-CCA estimates under Setup 1* in spectral norm (○) and Frobenius norm (△) using true r_1, r_2 and r_{12}, and those in spectral norm (●) and Frobenius norm (▲) using $\widehat{r}_1, \widehat{r}_2$ and $\widehat{r}_{12}$.

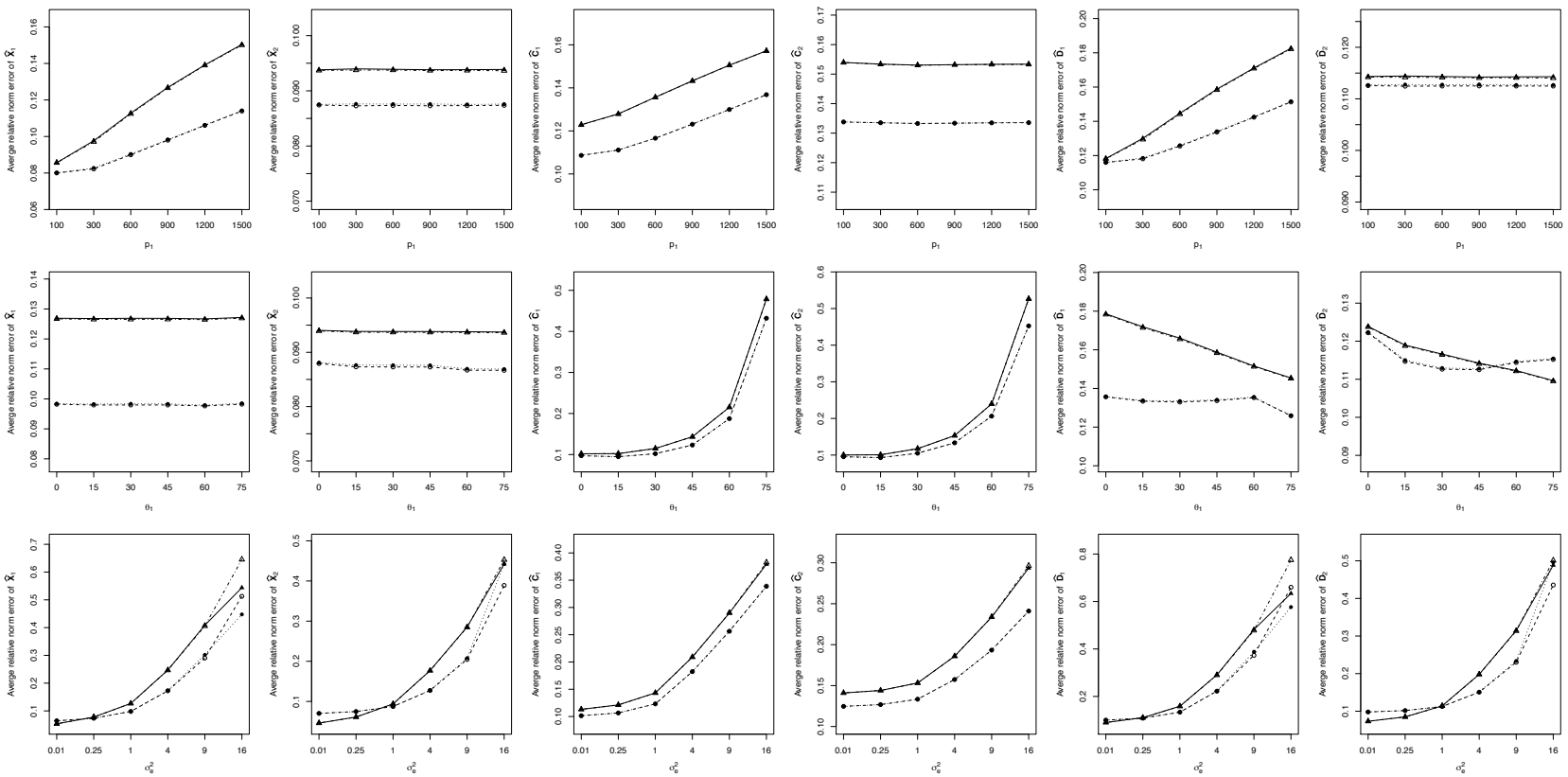


Figure S.2: Average relative errors of D-CCA estimates under Setup 2* in spectral norm (○) and Frobenius norm (△) using true r_1, r_2 and r_{12}, and those in spectral norm (●) and Frobenius norm (▲) using $\widehat{r}_1, \widehat{r}_2$ and $\widehat{r}_{12}$.

References

Hallin, M. and Liška, R. (2011), "Dynamic factors in the presence of blocks," *Journal of Econometrics*, 163, 29–41.

Horn, R. A. and Johnson, C. R. (1994), *Topics in Matrix Analysis*, Cambridge University Press, Cambridge.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation," *The Annals of Statistics*, 37, 4254–4278.

Wang, W. and Fan, J. (2017), "Asymptotics of empirical eigenstructure for high dimensional spiked covariance," *The Annals of Statistics*, 45, 1342–1374.

Yu, Y., Wang, T., and Samworth, R. J. (2015), "A useful variant of the Davis–Kahan theorem for statisticians," *Biometrika*, 102, 315–323.