D-GCCA: Decomposition-based Generalized Canonical Correlation Analysis for Multiple High-dimensional Datasets

Hai Shu hs120@nyu.edu
Department of Biostatistics
New York University
New York, NY 10003, USA

Zhe Qu zqu2@tulane.edu
Department of Mathematics
Tulane University
New Orleans, LA 70118, USA

Hongtu Zhu htzhu@email.unc.edu
Department of Biostatistics
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
and
AI Labs, Didi Chuxing, Beijing, China

Editor:

arXiv:2001.02856v1 [stat.ML] 9 Jan 2020

Abstract

Modern biomedical studies often collect multiple types of high-dimensional data on a common set of objects. A popular model for the joint analysis of multi-type datasets decomposes each data matrix into a low-rank common-variation matrix generated by latent factors shared across all datasets, a low-rank distinctive-variation matrix corresponding to each dataset, and an additive noise matrix. We propose decomposition-based generalized canonical correlation analysis (D-GCCA), a novel decomposition method that appropriately defines those matrices on the $\mathcal{L}^2$ space of random variables, whereas most existing methods are developed on its approximation, the Euclidean dot product space. Moreover to well calibrate common latent factors, we impose a desirable orthogonality constraint on distinctive latent factors. Existing methods inadequately consider such orthogonality and can thus suffer from substantial loss of undetected common variation. Our D-GCCA takes one step further than GCCA by separating common and distinctive variations among canonical variables, and enjoys an appealing interpretation from the perspective of principal component analysis. Consistent estimators of our common-variation and distinctive-variation matrices are established with good finite-sample numerical performance, and have closed-form expressions leading to efficient computation especially for large-scale datasets. The superiority of D-GCCA over state-of-the-art methods is also corroborated in simulations and real-world data examples.

Keywords: Canonical variable, common structure, distinctive structure, data integration, high-dimensional data.

1. Introduction

Data integration is widely used in biomedical studies to extract data from disparate sources on a common set of objects into meaningful and valuable information. Such studies include The Cancer Genome Atlas (TCGA; Hoadley et al., 2018) with multi-platform genomic data for tumor samples, and Human Connectome Project (HCP; Van Essen et al., 2013) with multi-modal brain images of healthy adults, among many others (Crawford et al., 2016; Jensen et al., 2017). The use of multiple data types can allow us to enhance understanding the etiology of many complex diseases, such as cancers (Ciriello et al., 2015; Campbell et al., 2018) and neurodegenerative diseases (Weiner et al., 2013; Saeed et al., 2017). Researchers hence have became highly interested in studying the shared information and individual features across multi-type datasets through separating their common and distinctive variation structures (van der Kloet et al., 2016; Smilde et al., 2017; Li et al., 2018).

Let $\mathbf{Y}_k \in \mathbb{R}^{p_k \times n}$ be the k-th row-mean centered dataset obtained on a common set of n objects for $k = 1, \ldots, K$, where p_k is the number of variables for the k-th dataset. One popular approach for disentangling their common and distinctive variation structures is to decompose each data matrix into

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k \qquad \text{for} \quad k = 1, \ldots, K, \tag{1}$$

where $\{\mathbf{X}_k\}_{k=1}^K$ are low-rank signal matrices with $\{\mathbf{E}_k\}_{k=1}^K$ being additive noise matrices, $\{\mathbf{C}_k\}_{k=1}^K$ are low-rank common-variation matrices that represent the signal data coming from the common mechanism shared across all datasets, and $\{\mathbf{D}_k\}_{k=1}^K$ are low-rank distinctive-variation matrices each from the distinctive mechanism of each single dataset that is not shared by all. Both common and distinctive mechanisms, also known as latent factors, denote the underlying causes of variation in the data (Schouteden et al., 2014).

Various orthogonality constraints for defining common-variation and distinctive-variation matrices in model (1) have been proposed by six state-of-the-art decomposition methods, including orthogonal n-block partial least squares (OnPLS; Löfstedt and Trygg, 2011), distinctive and common components with simultaneous component analysis (DISCO-SCA; Schouteden et al., 2014), common orthogonal basis extraction (COBE Zhou et al., 2016), joint and individual variation explained (JIVE; Lock et al., 2013) and its variants R.JIVE (O'Connell and Lock, 2016) and AJIVE (Feng et al., 2018). These methods can be applied to multiple datasets, $K \geq 2$, but suffer from two major issues: (i) all their decompositions are built on the inappropriate Euclidean dot product space $(\mathbb{R}^n, \cdot)$, which simply approximates the $\mathcal{L}^2$ space of random variables; (ii) they inadequately consider orthogonality constraints among distinctive-variation matrices $\{\mathbf{D}_k\}_{k=1}^K$, and thus these matrices may retain some important common variation. To address these issues, a nice decomposition, called decomposition-based canonical correlation analysis (D-CCA), is recently proposed in Shu et al. (2019) based on the canonical correlation analysis (CCA; Hotelling, 1936), but unfortunately, it is limited to two datasets, $K = 2$.

The aim of this paper is to address issues (i) and (ii) for multiple datasets, $K \geq 2$. We assume that the columns of each matrix in (1) are n independent copies of the corresponding random vector in

$$\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \boldsymbol{c}_k + \boldsymbol{d}_k + \boldsymbol{e}_k \in \mathbb{R}^{p_k}, \tag{2}$$

with entries of $\boldsymbol{c}_k$, $\boldsymbol{d}_k$, and $\boldsymbol{e}_k$ belonging to $\mathcal{L}_0^2$, where $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$ are, respectively, generated by common and distinctive latent factors. Here, $\mathcal{L}_0^2$ is the vector space composed of all real-valued random variables with zero mean and finite variance. We denote $(\mathcal{L}_0^2, \mathrm{cov})$ to be the inner product space of $\mathcal{L}_0^2$ that is endowed with the covariance operator as the inner product.

The traditional dot product of two data sample matrices is equivalent to the sample covariance matrix between their corresponding random vectors, for example, $\mathbf{C}_1\mathbf{D}_1^\top/n \approx \mathrm{cov}(\boldsymbol{c}_1, \boldsymbol{d}_1)$. The matrices $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ of the above six multi-set decomposition methods are defined under the orthogonality of $(\mathbb{R}^n, \cdot)$, and thus can only be viewed as approximations of the data samples for $\{\boldsymbol{c}_k, \boldsymbol{d}_k\}_{k=1}^K$ defined similarly on $(\mathcal{L}_0^2, \mathrm{cov})$. Moreover, unlike their sample-based definitions on $(\mathbb{R}^n, \cdot)$, our population-based definition from $(\mathcal{L}_0^2, \mathrm{cov})$ naturally enables the investigation of estimation consistency for recovering unobserved $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ from observable $\{\mathbf{Y}_k\}_{k=1}^K$.

Even translated into $(\mathcal{L}_0^2, \mathrm{cov})$, the six competing methods focus on the orthogonality (i.e., uncorrelatedness) between $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$, but inadequately consider orthogonality constraints among $\{\boldsymbol{d}_k\}_{k=1}^K$. Specifically, OnPLS, COBE, JIVE, and AJIVE do not impose any orthogonality on $\{\boldsymbol{d}_k\}_{k=1}^K$; R.JIVE enforces such orthogonality at the price of relegating its unexplainable portion of signal $\boldsymbol{x}_k$ into noise $\boldsymbol{e}_k$; DISCO-SCA often only approximates but not exactly achieves its target orthogonality for $\{\boldsymbol{d}_k\}_{k=1}^K$ (van der Kloet et al., 2016). When $K = 2$, the orthogonality between $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ desirably assures no common latent factors retained between them. For $K > 2$, with the same aim to well capture the common mechanism, a similar desirable orthogonality constraint on $\{\boldsymbol{d}_k\}_{k=1}^K$ is that at least one pair among them are uncorrelated. However, it has been unclear how to build a decomposition for all $K \geq 2$ that can ensure both the above desirable orthogonality among $\{\boldsymbol{d}_k\}_{k=1}^K$ and the interpretability of associated $\{\boldsymbol{c}_k\}_{k=1}^K$. After all, the former alone is insufficient to guarantee the latter.

We propose a novel method, called decomposition-based generalized canonical correlation analysis (D-GCCA), to handle cases with $K \geq 2$. Our method is equivalent to D-CCA when $K = 2$. The key idea is to divide the decomposition problem (2) into multiple subproblems via Carroll's generalized canonical correlation analysis (GCCA; Carroll, 1968). We slightly relax the aforementioned desirable orthogonality of $\{\boldsymbol{d}_k\}_{k=1}^K$ by enforcing it for each subproblem. This in turn leads to a geometrically interpretable definition of $\{\boldsymbol{c}_k\}_{k=1}^K$ on space $(\mathcal{L}_0^2, \mathrm{cov})$ by connecting the Carroll's GCCA with principal component analysis (PCA; Hotelling, 1933). In particular, our defined common latent factors of $\{\boldsymbol{x}_k\}_{k=1}^K$ represent the same contribution made by the principal basis of the entire signal space $\sum_{k=1}^K \mathrm{span}(\boldsymbol{x}_k^\top)$ in generating each of the K signal subspaces $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. Here, $\mathrm{span}(\boldsymbol{v}^\top)$ denotes the subspace spanned by entries of $\boldsymbol{v}$ for any random vector $\boldsymbol{v}$ in $(\mathcal{L}_0^2, \mathrm{cov})$.

Recovering high-dimensional matrices $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ poses practical difficulties because only matrices $\{\mathbf{Y}_k\}_{k=1}^K$ are observable and they are often high-rank. If the high-dimensional, high-rank $\mathbf{Y}_k$ is treated as the signal $\mathbf{X}_k$, its associated high-rank covariance matrix $\mathrm{cov}(\boldsymbol{x}_k)$ can be inconsistently estimated by the traditional sample covariance matrix due to the curse of "intrinsic" high dimensionality (Yin et al., 1988; Vershynin, 2012). Low-rank $\mathbf{X}_k$ or equivalently low-rank $\mathrm{cov}(\boldsymbol{x}_k)$ is often assumed to facilitate the construction of consistent estimates (Shu et al., 2019). Fortunately, big data matrices are often approximately low-rank in many real-world applications (Udell and Townsend, 2019), and their low-rank

approximations render feasible or more efficient computation, while retaining the major portion of information (Kishore Kumar and Schneider, 2017). We consider the low-rank plus noise structure given in (1) and (2) under the widely used high-dimensional spiked covariance model (Fan et al., 2013; Wang and Fan, 2017; Shu et al., 2019). Subsequently, we propose soft-thresholding based estimators for $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$. Convergence properties of our estimators are established with reasonably good finite-sample performance shown by simulations. The proposed matrix estimators have closed-form expressions and thus are more computationally efficient than most existing methods that use time-expensive iterative optimization algorithms. For example, to decompose three 91,282×1080 data matrices in our HCP application, our approach can complete in 18 seconds on a single computing node, while some state-of-the-art methods cannot converge within 5 hours.

The contributions of this paper are summarized below:

- We propose a novel decomposition method, called D-GCCA, for tackling $K \geq 2$ datasets under model (1), based on $(\mathcal{L}_0^2, \mathrm{cov})$ instead of $(\mathbb{R}^n, \cdot)$. Our distinctive-variation matrices are especially imposed with an orthogonality constraint to avoid substantial loss of undetected common variation. The proposed common-variation matrices exhibit a geometric interpretation from the perspective of PCA. Our D-GCCA reduces to D-CCA when $K = 2$.

- We establish consistent estimators for our defined common-variation and distinctive-variation matrices under high-dimensional settings with convergence rates in both the Frobenius norm and the spectral norm. The proposed estimators have closed-form expressions and thus are computationally efficient.

- We compare our D-GCCA with the six competing methods on both simulated and real-world data to show the superiority of proposed method for separating the common and distinctive variations across multiple datasets.

- As a byproduct, we reformulate Carroll's GCCA on $(\mathcal{L}_0^2, \mathrm{cov})$ from the traditional $(\mathbb{R}^n, \cdot)$ and provide some useful properties, which may facilitate the use of GCCA in statistical data integration.

The rest of this paper is organized as follows. We introduce our random-variable version of Carroll's GCCA and propose our D-GCCA method in Section 2. We provide our estimation approach of the D-GCCA defined matrices and its asymptotic properties in Section 3. Section 4 evaluates the finite-sample performance of proposed estimators via simulations. We also compare D-GCCA with the six competing methods through simulated data in Section 4 and through two real-world data examples from TCGA and HCP in Section 5. Concluding remarks are drawn in Section 6. All theoretical proofs are provided in Appendix A, and additional simulation results are presented in Appendix B.

We now introduce some notation. For a real matrix $\mathbf{M} = (M_{ij})_{1\leq i\leq p, 1\leq j\leq n}$, the ℓ-th largest singular value is denoted by $\sigma_\ell(\mathbf{M})$, the ℓ-th largest eigenvalue when $p = n$ is $\lambda_\ell(\mathbf{M})$, the spectral norm is $\|\mathbf{M}\|_2 = \{\lambda_1(\mathbf{M}^\top\mathbf{M})\}^{1/2}$, the Frobenius norm is $\|\mathbf{M}\|_F = (\sum_{i=1}^p \sum_{j=1}^n M_{ij}^2)^{1/2}$, the matrix $\mathcal{L}^\infty$ norm is $\|\mathbf{M}\|_\infty = \max_{1\leq i\leq p} \sum_{j=1}^n |M_{ij}|$, the max norm is $\|\mathbf{M}\|_{\max} = \max_{1\leq i\leq p, 1\leq j\leq n} |M_{ij}|$, and the Moore-Penrose pseudoinverse is $\mathbf{M}^\dagger$. 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 $\mathbf{M}$, respectively. We write the j-th entry of a vector $\boldsymbol{v}$ by $\boldsymbol{v}^{[j]}$, and $\boldsymbol{v}^{[s:t]} = (\boldsymbol{v}^{[s]}, \boldsymbol{v}^{[s+1]}, \ldots, \boldsymbol{v}^{[t]})^\top$. Define $(v_i)_{i\in\mathcal{I}}$ by $(v_{i_1}, \ldots, v_{i_q})$ with $\mathcal{I} = \{i_1, \ldots, i_q\}$ and $i_1 < \ldots < i_q$. The angle between any $x, y \in (\mathcal{L}_0^2, \mathrm{cov})$ is denoted by $\theta(x, y)$, and the norm of x is $\|x\| = \sqrt{\mathrm{var}(x)}$. We use $\cos\{\theta(x, y)\}$ and $\mathrm{corr}(x, y)$ exchangeably, and define $\mathrm{corr}(x, 0) = 0$. The symbol $\perp$ used between two subspaces and/or random variables in $(\mathcal{L}_0^2, \mathrm{cov})$ means their orthogonality, that is, uncorrelatedness. Define $r_0 = 0$, $r_k = \mathrm{rank}\{\mathrm{cov}(\boldsymbol{x}_k)\}$ and $r_f = \mathrm{rank}\{\mathrm{cov}((\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_K^\top)^\top)\}$. It holds that $r_k = \dim\{\mathrm{span}(\boldsymbol{x}_k^\top)\}$ and $r_f = \dim\{\mathrm{span}((\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_K^\top))\}$. Throughout the paper, the asymptotic arguments are by default under $n \to \infty$.

2. Methodology

We first develop the random-variable version of Carroll's GCCA in $(\mathcal{L}_0^2, \mathrm{cov})$ and then use this framework to derive the D-GCCA decomposition.

2.1 Generalized canonical correlation analysis

We translate Carroll's GCCA into the space $(\mathcal{L}_0^2, \mathrm{cov})$. Carroll's GCCA was originally proposed and is often studied in $(\mathbb{R}^n, \cdot)$ using data samples (Carroll, 1968; van de Velden, 2011; Draper et al., 2014). Kettenring (1971) briefly mentioned that the random-variable version of Carroll's GCCA is a mixture of his maximum variance and minimum variance methods. We provide the solution to the optimization problem of Carroll's GCCA in $(\mathcal{L}_0^2, \mathrm{cov})$ as well as some important properties.

For subspaces $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$, the Carroll's GCCA in $(\mathcal{L}_0^2, \mathrm{cov})$ sequentially finds the closest elements among the K subspaces. The method has r_f recursive stages. The ℓ-th stage finds the closest elements, denoted as $z_1^{(\ell)}, \ldots, z_K^{(\ell)}$, among the K subspaces, which are called the ℓ-th set of canonical variables, along with an auxiliary variable $w^{(\ell)}$ as follows:

$$\begin{aligned} \{z_1^{(\ell)}, \ldots, z_K^{(\ell)}, w^{(\ell)}\} &= \underset{\{z_1,\ldots,z_K,w\}}{\arg\max} \sum_{k=1}^K \cos^2\{\theta(z_k, w)\} \\ \text{subject to} \quad &\begin{cases} z_k \in \mathrm{span}(\boldsymbol{x}_k^\top),\ \|z_k\| = 1, \\ w \perp w^{(j)},\ w \in \mathcal{L}_0^2,\ \|w\| = 1,\ j < \ell. \end{cases} \end{aligned} \tag{3}$$

Let $\boldsymbol{f}_k^\top$ be an arbitrary orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$, $\boldsymbol{f} = (\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_K^\top)^\top$, and $\{\boldsymbol{\eta}^{(\ell)}\}_{1\le\ell\le r_f}$ be any r_f orthonormal eigenvectors of $\mathrm{cov}(\boldsymbol{f})$, where $\boldsymbol{\eta}^{(\ell)} = [(\boldsymbol{\eta}_1^{(\ell)})^\top, \ldots, (\boldsymbol{\eta}_K^{(\ell)})^\top]^\top$ corresponds to $\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))$ with $\boldsymbol{\eta}_k^{(\ell)} \in \mathbb{R}^{r_k}$. Note that $r_f = \mathrm{rank}\{\mathrm{cov}(\boldsymbol{f})\}$. The following theorem presents the solution to (3) as well as some useful properties for our decomposition method.

Theorem 1 *The following results hold.*

(i) *For $\ell \le r_f$ and $k \le K$, the solution of (3) is given by*

$$z_k^{(\ell)} = \begin{cases} \textit{any standardized variable in } \mathrm{span}(\boldsymbol{x}_k^\top), & \textit{if } \boldsymbol{\eta}_k^{(\ell)} = \boldsymbol{0}, \\ \pm(\boldsymbol{\eta}_k^{(\ell)}/\|\boldsymbol{\eta}_k^{(\ell)}\|_F)^\top \boldsymbol{f}_k, & \textit{if } \boldsymbol{\eta}_k^{(\ell)} \neq \boldsymbol{0}, \end{cases} \tag{4}$$

$$w^{(\ell)} = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{-1/2} (\boldsymbol{\eta}^{(\ell)})^\top \boldsymbol{f}. \tag{5}$$

Moreover, we have

$$\cos\{\theta(z_k^{(\ell)}, w^{(\ell)})\} = \pm[\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{1/2}\|\boldsymbol{\eta}_k^{(\ell)}\|_F,$$
$$\sum_{k=1}^{K}\cos^2\{\theta(z_k^{(\ell)}, w^{(\ell)})\} = \lambda_\ell(\mathrm{cov}(\boldsymbol{f})), \tag{6}$$
$$\sum_{k=1}^{K}\mathrm{span}(\boldsymbol{x}_k^\top) = \mathrm{span}(\{w^{(\ell)}\}_{\ell=1}^{r_f}). \tag{7}$$

(ii) *For $\ell \le r_f$, re-define $z_k^{(\ell)}$ in (4) to be*

$$z_k^{(\ell)} = \begin{cases} 0, & \text{if } \boldsymbol{\eta}_k^{(\ell)} = \mathbf{0}, \text{ i.e., } w^{(\ell)} \perp \mathrm{span}(\boldsymbol{x}_k^\top), \\ (\boldsymbol{\eta}_k^{(\ell)}/\|\boldsymbol{\eta}_k^{(\ell)}\|_F)^\top \boldsymbol{f}_k, & \text{otherwise.} \end{cases} \tag{8}$$

Then, we have $\theta(z_k^{(\ell)}, w^{(\ell)}) \in [0, \pi/2]$ and $\mathrm{span}(\{z_k^{(\ell)}\}_{\ell=1}^{r_f}) = \mathrm{span}(\boldsymbol{x}_k^\top)$.

(iii) *For $z_k^{(\ell)}$ in either (4) or (8), if $\lambda_\ell(\mathrm{cov}(\boldsymbol{f})) \le 1$ and $\mathrm{span}(\{z_k^{(m)}\}_{m=1}^{\ell-1}) \ne \mathrm{span}(\boldsymbol{x}_k^\top)$ for some ℓ and k, then there exists a $w^{(\ell)} \in \mathrm{span}(\boldsymbol{x}_k^\top)$ such that $w^{(\ell)} \perp \sum_{1\le j\ne k\le K}\mathrm{span}(\boldsymbol{x}_j^\top)$.*

In the following text, if without further clarification, we refer $z_k^{(\ell)}$ to the one defined in (8) so that $\theta(z_k^{(\ell)}, w^{(\ell)}) \in [0, \pi/2]$.

2.2 Definition of common-variation and distinctive-variation matrices

In the model given by (1) and (2), the columns of each common-variation matrix $\mathbf{C}_k$ or distinctive-variation matrix $\mathbf{D}_k$ are assumed to be n independent copies of its corresponding random vector $\boldsymbol{c}_k$ or $\boldsymbol{d}_k$. We thus consider the following decomposition with noise excluded:

$$\boldsymbol{x}_k = \boldsymbol{c}_k + \boldsymbol{d}_k \quad \text{for} \quad k = 1, \ldots, K. \tag{9}$$

The estimation of $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ from noisy data $\{\mathbf{Y}_k\}_{k=1}^K$ will be given in Section 3.

Like the divide-and-conquer strategy of D-CCA, we first break down decomposition problem (9) into multiple subproblems. Each ℓ-th subproblem is solved by finding a common variable $c^{(\ell)}$ and K distinctive variables $\{d_k^{(\ell)}\}_{k=1}^K$ for the ℓ-th set of canonical variables $\{z_k^{(\ell)}\}_{k=1}^K$ such that

$$z_k^{(\ell)} = c^{(\ell)} + d_k^{(\ell)} \quad \text{for} \quad k = 1, \ldots, K. \tag{10}$$

The auxiliary variable $w^{(\ell)}$ in (3) naturally serves as the direction variable of our common variable $c^{(\ell)}$ of $\{z_k^{(\ell)}\}_{k=1}^K$. We define $c^{(\ell)}$ by

$$c^{(\ell)} = \alpha^{(\ell)} w^{(\ell)}, \tag{11}$$

where $\alpha^{(\ell)}$ satisfies the constraints:

(C.1) $|\alpha^{(\ell)}|$ is the smallest value such that at least one pair among $\{d_k^{(\ell)}\}_{k=1}^K$ are orthogonal;

(C.2) $\alpha^{(\ell)} < 0$ if (C.1) has two solutions with respect to $\alpha^{(\ell)}$.

The rationale of setting constraints (C.1) and (C.2) are given as follows. The structure of at least one orthogonal pair among $\{d_k^{(\ell)}\}_{k=1}^K$ is the relaxed analogy of the desirable orthogonality among $\{\boldsymbol{d}_k\}_{k=1}^K$ mentioned in Section 1 that is used on each ℓ-th subproblem. Let $\alpha_1^{(\ell)}$ and $\alpha_2^{(\ell)}$ be two candidate values of $\alpha^{(\ell)}$, each of which leads to the required orthogonality among $\{d_k^{(\ell)}\}_{k=1}^K$. If $|\alpha_1^{(\ell)}| < |\alpha_2^{(\ell)}|$, then the extra variance $(|\alpha_2^{(\ell)}|^2 - |\alpha_1^{(\ell)}|^2)$ for the variable $c^{(\ell)}$ of $\alpha_2^{(\ell)}$ can be alternatively explained by the variables $\{d_k^{(\ell)}\}_{k=1}^K$ of $\alpha_1^{(\ell)}$. If $\alpha_1^{(\ell)} < 0 < \alpha_2^{(\ell)}$ and $|\alpha_1^{(\ell)}| = |\alpha_2^{(\ell)}|$, then the $d_k^{(\ell)}$ corresponding to $\alpha_1^{(\ell)}$, for $k = 1, \ldots, K$, has a larger variance than that to $\alpha_2^{(\ell)}$.

The existence and computing formula of $\alpha^{(\ell)}$ is shown in the theorem below.

Theorem 2 *For $\ell \le r_f$, $w^{(\ell)}$ in (5), and $\{z_k^{(\ell)}\}_{k=1}^K$ in (8), we have that $\alpha^{(\ell)}$ in (11) exists and satisfies*

$$\alpha^{(\ell)} \in \arg\min_{\alpha_{jk}^{(\ell)}} \left\{ \left|\alpha_{jk}^{(\ell)}\right| : \alpha_{jk}^{(\ell)} = \frac{1}{2}\left[\cos\{\theta(w^{(\ell)}, z_j^{(\ell)})\} + \cos\{\theta(w^{(\ell)}, z_k^{(\ell)})\} - (\Delta_{jk}^{(\ell)})^{1/2}\right] \right.$$
$$\left. \text{for} \quad \Delta_{jk}^{(\ell)} \ge 0 \quad \text{and} \quad 1 \le j < k \le K \right\}$$

with $\Delta_{jk}^{(\ell)} = [\cos\{\theta(w^{(\ell)}, z_j^{(\ell)})\} + \cos\{\theta(w^{(\ell)}, z_k^{(\ell)})\}]^2 - 4\cos\{\theta(z_j^{(\ell)}, z_k^{(\ell)})\}$.

We interpret the decomposition given in (10) and (11) via analyzing the relationship between the entire signal space $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$ and its subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. First, from the perspective of PCA, we consider how the K signal subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$ contribute to form the whole signal space $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$. We use an arbitrary orthonormal basis $\boldsymbol{f}_k^\top$ of $\text{span}(\boldsymbol{x}_k^\top)$ to represent its contribution to $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$, because $\boldsymbol{f}_k^\top$ fully characterizes $\text{span}(\boldsymbol{x}_k^\top)$ due to $\text{span}(\boldsymbol{x}_k^\top) = \{\boldsymbol{f}_k^\top \boldsymbol{b} : \forall \boldsymbol{b} \in \mathbb{R}^{r_k}\}$, and its entries, all of which are standardized variables, provide a fair comparison among subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. By (5) and (7), we see that $\left\{[\lambda_\ell(\text{cov}(\boldsymbol{f}))]^{1/2} w^{(\ell)}\right\}_{\ell=1}^{r_f}$ are the first r_f principal components of $\boldsymbol{f}^\top = (\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_K^\top)$, which fully capture the variance of $\boldsymbol{f}$, that is, the accumulated contribution to $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$ from all subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. They also constitute an orthogonal basis of $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$ that is the closest to these subspaces in the sense of (3). This leads to the following definition.

Definition 1 *Standardized variables $\{w^{(\ell)}\}_{\ell=1}^{r_f}$ given in (5) are defined to be the principal basis of $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$ with respect to $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$.*

Next, from the perspective of the principal basis $\{w^{(\ell)}\}_{\ell=1}^{r_f}$, we conversely deduce how the entire signal space $\sum_{k=1}^K \text{span}(\boldsymbol{x}_k^\top)$ generates its subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. With $0/0 = 0$, $z_k^{(\ell)}$ is the standardized version of the projection of $w^{(\ell)}$ onto $\text{span}(\boldsymbol{x}_k^\top)$. Theorem 1 (ii) shows that the standardized projections $\{z_k^{(\ell)}\}_{\ell=1}^{r_f}$ of $\{w^{(\ell)}\}_{\ell=1}^{r_f}$ span the subspace $\text{span}(\boldsymbol{x}_k^\top)$ for each $k \le K$. Hence, the decomposition in (10) and (11) essentially measures the same contribution of the principal-basis component $w^{(\ell)}$ in generating each of the K signal subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$.

Remark 1 *Let* $L = \max\{\ell \in \{1, \ldots, r_f\} : \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) > 1\}$. *We only need to consider the first* L *principal-basis components* $\{w^{(\ell)}\}_{\ell=1}^{L}$ *due to the following reasons. For* $\ell > L$, *by Theorem 1 (iii), either there exists a* $w^{(\ell)} \in \mathrm{span}(\boldsymbol{x}_k^\top)$ *for some* k *that is orthogonal to all the other signal subspaces* $\{\mathrm{span}(\boldsymbol{x}_j^\top)\}_{j\neq k}$, *or otherwise* $\{z_k^{(m)}\}_{m=1}^{\ell-1}$ *has spanned the subspace* $\mathrm{span}(\boldsymbol{x}_k^\top)$ *for all* k. *The first scenario results in* $c^{(\ell)} = 0$, *and the second one indicates that the contribution of* $w^{(\ell)}$ *to each signal subspace has already been accomplished by the preceding components* $\{w^{(m)}\}_{m=1}^{\ell-1}$.

We now combine the decompositions for all $\ell = 1, \ldots, L$ in (10) to form the original decomposition (9). Define the index set of nonzero $c^{(\ell)}$s by $\mathcal{I}_0 = \{\ell \in \{1, \ldots, L\} : c^{(\ell)} \neq 0, \text{ i.e., } \alpha^{(\ell)} \neq 0\}$. We set $\boldsymbol{c}_k = \boldsymbol{0}_{p_k\times 1}$ and $\mathbf{C}_k = \boldsymbol{0}_{p_k\times n}$ for all k when $\mathcal{I}_0 = \emptyset$, and only consider $\mathcal{I}_0 \neq \emptyset$ in the following. Let $\boldsymbol{z}_k^{\mathcal{I}_0} = (z_k^{(\ell)})_{\ell\in\mathcal{I}_0}^\top$. The portion of $\boldsymbol{x}_k$ generated from latent factors $\boldsymbol{z}_k^{\mathcal{I}_0}$ is equivalent to the projection of $\boldsymbol{x}_k$ onto $\mathrm{span}\{(\boldsymbol{z}_k^{\mathcal{I}_0})^\top\}$, and can be written by

$$\mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger \boldsymbol{z}_k^{\mathcal{I}_0} = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger (c^{(\ell)} + d_k^{(\ell)})_{\ell\in\mathcal{I}_0}^\top. \tag{12}$$

Here, $\mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger$ is a deterministic coefficient matrix. We define the common-variation vector $\boldsymbol{c}_k$ of $\boldsymbol{x}_k$ by

$$\boldsymbol{c}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger \boldsymbol{c}^{\mathcal{I}_0}, \tag{13}$$

which is the portion of (12) comes from the common latent factors $(\boldsymbol{c}^{\mathcal{I}_0})^\top := (c^{(\ell)})_{\ell\in\mathcal{I}_0}$.

Definition 2 *We define the common-variation vector* $\boldsymbol{c}_k$ *of* $\boldsymbol{x}_k$ *by* (13) *and the distinctive-variation vector* $\boldsymbol{d}_k = \boldsymbol{x}_k - \boldsymbol{c}_k$. *The common-variation matrix* $\mathbf{C}_k$ *and distinctive-variation matrix* $\mathbf{D}_k$ *are the corresponding sample matrices of* $\boldsymbol{c}_k$ *and* $\boldsymbol{d}_k$, *respectively.*

Remark 2 *Our common-variation vectors* $\{\boldsymbol{c}_k\}_{k=1}^K$ *in* (13) *are all generated by the same latent factors* $(\boldsymbol{c}^{\mathcal{I}_0})^\top = (c^{(\ell)})_{\ell\in\mathcal{I}_0}$. *As explained in the paragraph before Remark 1, for* $\ell \leq L$, $c^{(\ell)}$ *is the contribution of the principal-basis component* $w^{(\ell)}$ *made uniformly to generating all signal subspaces* $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^K$. *Vector* $\boldsymbol{c}^{\mathcal{I}_0}$ *contains these* $c^{(\ell)}$*s that are nonzero. This generative nature of* $(\boldsymbol{c}^{\mathcal{I}_0})^\top$ *indicates that even some part of the common mechanism is possibly retained among* $\{\boldsymbol{d}_k\}_{k=1}^K$ *due to relaxing the latter's desirable orthogonality into each subproblem* (10), *it is less important than* $(\boldsymbol{c}^{\mathcal{I}_0})^\top$ *and may be further explored by recursively applying our proposed decomposition. When* $K = 2$, *by Theorem 3 below, our D-GCCA decomposition is equivalent to D-CCA, and thus ensures* $\mathrm{span}(\boldsymbol{d}_1^\top) \perp \mathrm{span}(\boldsymbol{d}_2^\top)$.

Theorem 3 *When* $K = 2$, $\{\boldsymbol{c}_k\}_{k=1}^K$ *in* (13) *are the same as those of D-CCA in* (16) *of Shu et al. (2019).*

We further investigate the uniqueness of $\{\boldsymbol{c}_k\}_{k=1}^K$.

Theorem 4 *For* $L \geq 1$, *assume that* $\lambda_1(\mathrm{cov}(\boldsymbol{f})), \ldots, \lambda_L(\mathrm{cov}(\boldsymbol{f}))$ *are distinct, then* $\{\boldsymbol{c}_k\}_{k=1}^K$ *are uniquely defined by* (13) *no matter the choice of* $\boldsymbol{f}$ *and* $\{\boldsymbol{\eta}^{(\ell)}\}_{1\leq\ell\leq L}$.

The largest L eigenvalues of $\mathrm{cov}(\boldsymbol{f})$ are invariant to the choice of $\boldsymbol{f}$. For a given $\boldsymbol{f}$, the distinctness of these L eigenvalues ensures the identifiability of $\{\boldsymbol{\eta}^{(\ell)}\}_{1\le\ell\le L}$ up to a sign change and thus simplifies the analysis. Analogous assumptions are often made in the literature, such as Zhou and He (2008), Birnbaum et al. (2013), and Wang and Fan (2017). If the joint distribution of the n ($\ge L$) samples of $\boldsymbol{f}$ is absolutely continuous or elliptically contoured, then the largest L eigenvalues of its sample covariance matrix are distinct with probability one (Okamoto, 1973; Gupta and Varga, 1991). Hence, our distinct eigenvalues assumption is plausible in practice.

3. Estimation

3.1 Matrix estimators

We derive the estimators of common-variation and distinctive-variation matrices $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ by starting with the estimation of signal matrices $\{\mathbf{X}_k = \mathbf{C}_k + \mathbf{D}_k\}_{k=1}^K$ from the observable data $\{\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k\}_{k=1}^K$.

Suppose that the low-rank plus noise structure in (1) and (2) follows the factor model:

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{B}_k\mathbf{F}_k + \mathbf{E}_k, \qquad \boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \mathbf{B}_k\boldsymbol{f}_k + \boldsymbol{e}_k, \tag{14}$$

where $\mathbf{B}_k \in \mathbb{R}^{p_k\times r_k}$ is a real deterministic matrix, the columns of $\mathbf{F}_k$ and $\mathbf{E}_k$ are, respectively, the n independent copies of $\boldsymbol{f}_k$ and $\boldsymbol{e}_k$, $\boldsymbol{f}_k^\top$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$ with $\mathrm{cov}(\boldsymbol{f}_k, \boldsymbol{e}_k) = \mathbf{0}_{r_k\times p_k}$, $\mathrm{span}(\boldsymbol{x}_k^\top)$ is a fixed space that is independent of $\{p_k\}_{k=1}^K$ and n, and $\mathbf{F} := (\mathbf{F}_1^\top, \ldots, \mathbf{F}_K^\top)^\top$ has independent columns. We assume that $\mathrm{cov}(\boldsymbol{y}_k)$ is a spiked covariance matrix for which the largest r_k eigenvalues are significantly larger than the rest, namely, signals are distinguishably stronger than noises. The r_k spiked eigenvalues are majorly contributed by signal $\boldsymbol{x}_k$, whereas the rest small eigenvalues are induced by noise $\boldsymbol{e}_k$. The spiked covariance model has been widely used in various fields, such as signal processing (Nadakuditi and Silverstein, 2010), machine learning (Huang, 2017), and economics (Chamberlain and Rothschild, 1983).

For simplicity, we define the estimators of $\{\mathbf{X}_k, \mathbf{C}_k, \mathbf{D}_k\}_{k=1}^K$ using the true $\{r_k\}_{k=1}^K$, $\mathcal{I}_0$, $r_k^* = \mathrm{rank}(\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0}))$, as well as $\mathcal{I}_\Delta^{(\ell)} = \{(j,k) : \Delta_{jk}^{(\ell)} \ge 0, 1 \le j < k \le K\}$ and $\mathrm{sign}(\alpha^{(\ell)})$ for all $\ell \in \mathcal{I}_0$. The practical selection of these nuisance parameters is discussed in Section 3.3.

We use the following soft-thresholding estimator of $\mathbf{X}_k$ proposed in Shu et al. (2019). This estimator is originally inspired by the method of Wang and Fan (2017) for spiked covariance matrix estimation:

$$\widehat{\mathbf{X}}_k = \mathbf{U}_{k1}\,\mathrm{diag}\{\widehat{\sigma}_1^S(\mathbf{Y}_k), \ldots, \widehat{\sigma}_{r_k}^S(\mathbf{Y}_k)\}\mathbf{U}_{k2}^\top,$$

where $\widehat{\sigma}_\ell^S(\mathbf{Y}_k) = [\max\{\sigma_\ell^2(\mathbf{Y}_k) - \tau_k p_k, 0\}]^{1/2}$, $\tau_k = \sum_{\ell=r_k+1}^{p_k}\sigma_\ell^2(\mathbf{Y}_k)/(np_k - nr_k - p_kr_k)$, and $\mathbf{U}_{k1}\,\mathrm{diag}(\sigma_1(\mathbf{Y}_k), \ldots, \sigma_{r_k}(\mathbf{Y}_k))\mathbf{U}_{k2}^\top$ forms the top-r_k singular value decomposition (SVD) of $\mathbf{Y}_k$.

We next use $\widehat{\mathbf{X}}_k$ to develop estimators for $\mathbf{C}_k$ and $\mathbf{D}_k = \mathbf{X}_k - \mathbf{C}_k$. Define an estimator of $\mathrm{cov}(\boldsymbol{x}_k)$ by $\widehat{\mathrm{cov}}(\boldsymbol{x}_k) = n^{-1}\widehat{\mathbf{X}}_k\widehat{\mathbf{X}}_k^\top$ for which a SVD is denoted as $\widehat{\mathrm{cov}}(\boldsymbol{x}_k) = \widehat{\mathbf{V}}_{xk}\widehat{\boldsymbol{\Lambda}}_{xk}\widehat{\mathbf{V}}_{xk}^\top$, where $\widehat{\boldsymbol{\Lambda}}_{xk} = \mathrm{diag}([\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{x}_k))]_{1\le\ell\le r_k})$ and $\widehat{\mathbf{V}}_{xk}$ has r_k orthonormal columns. We can obtain $\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) = [\widehat{\sigma}_\ell^S(\mathbf{Y}_k)]^2/n$ and $\widehat{\mathbf{V}}_{xk} = \mathbf{U}_{k1}$. Define the estimators of $\mathbf{F}_k$ and

$\mathbf{F}$ by $\widehat{\mathbf{F}}_k = (\widehat{\mathbf{\Lambda}}_{xk}^{1/2})^\dagger \widehat{\mathbf{V}}_{xk}^\top \widehat{\mathbf{X}}_k$ and $\widehat{\mathbf{F}} = (\widehat{\mathbf{F}}_1^\top, \ldots, \widehat{\mathbf{F}}_K^\top)^\top$, respectively. We estimate $\mathrm{cov}(\boldsymbol{f})$ by $\widehat{\mathrm{cov}}(\boldsymbol{f}) = n^{-1}\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\top$. Let $\widehat{\boldsymbol{\eta}}^{(\ell)} = [(\widehat{\boldsymbol{\eta}}_1^{(\ell)})^\top, \ldots, (\widehat{\boldsymbol{\eta}}_K^{(\ell)})^\top]^\top$, with $\widehat{\boldsymbol{\eta}}_k^{(\ell)} \in \mathbb{R}^{r_k}$, be a normalized eigenvector of $\widehat{\mathrm{cov}}(\boldsymbol{f})$ corresponding to $\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f}))$. We also let different $\widehat{\boldsymbol{\eta}}^{(\ell)}$s be orthogonal. Our estimated sample vector of variable $w^{(\ell)}$ is defined by $(\widehat{\boldsymbol{w}}^{(\ell)})^\top = [\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f}))]^{-1/2}(\widehat{\boldsymbol{\eta}}^{(\ell)})^\top\widehat{\mathbf{F}}$ if $\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f})) \neq 0$ and otherwise $\widehat{\boldsymbol{w}}^{(\ell)} = \mathbf{0}_{n\times 1}$, and that of variable $z_k^{(\ell)}$ is $(\widehat{\boldsymbol{z}}_k^{(\ell)})^\top = (\widehat{\boldsymbol{\eta}}_k^{(\ell)}/\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F)^\top\widehat{\mathbf{F}}_k$ if $\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F \neq 0$ and otherwise $\widehat{\boldsymbol{z}}_k^{(\ell)} = \mathbf{0}_{n\times 1}$. We initially estimate $\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})$ by $\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0}) = \widehat{\mathbf{H}}_k\widehat{\mathbf{H}}_k^\top$, where $\widehat{\mathbf{H}}_k = (\widehat{\boldsymbol{\eta}}_k^{(\ell)}/\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F)_{\ell\in\mathcal{I}_0}^\top$ with $\mathbf{0}/0 = \mathbf{0}$. Let $\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0}) = \widehat{\mathbf{V}}_{zk}\widehat{\mathbf{\Lambda}}_{zk}\widehat{\mathbf{V}}_{zk}^\top$ be its compact SVD, where $\widehat{\mathbf{\Lambda}}_{zk}$ has nonincreasing diagonal elements. With $\check{r}_k = \min[r_k^*, \mathrm{rank}\{\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})\}]$, our estimator of $\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})$ is defined as the top-$\check{r}_k$ SVD of $\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})$, that is,

$$\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0}) = \widehat{\mathbf{V}}_{zk}^{[:,1:\check{r}_k]}\widehat{\mathbf{\Lambda}}_{zk}^{[1:\check{r}_k,1:\check{r}_k]}(\widehat{\mathbf{V}}_{zk}^{[:,1:\check{r}_k]})^\top. \tag{15}$$

Replacing $\cos\{\theta(w^{(\ell)}, z_k^{(\ell)})\}$ and $\cos\{\theta(z_j^{(\ell)}, z_k^{(\ell)})\}$ by $\widehat{\cos}\{\theta(w^{(\ell)}, z_k^{(\ell)})\} = n^{-1}(\widehat{\boldsymbol{w}}^{(\ell)})^\top\widehat{\boldsymbol{z}}_k^{(\ell)}$ and $\widehat{\cos}\{\theta(z_j^{(\ell)}, z_k^{(\ell)})\} = n^{-1}(\widehat{\boldsymbol{z}}_j^{(\ell)})^\top\widehat{\boldsymbol{z}}_k^{(\ell)}$ in $\Delta_{jk}^{(\ell)}$ yields its initial estimator $\widetilde{\Delta}_{jk}^{(\ell)}$. For $(j,k) \in \mathcal{I}_\Delta^{(\ell)}$, define

$$\widehat{\alpha}_{jk}^{(\ell)} = \frac{1}{2}\left[\widehat{\cos}\{\theta(w^{(\ell)}, z_j^{(\ell)})\} + \widehat{\cos}\{\theta(w^{(\ell)}, z_k^{(\ell)})\} - (\widehat{\Delta}_{jk}^{(\ell)})^{1/2}\right]$$

with $\widehat{\Delta}_{jk}^{(\ell)} = \max(\widetilde{\Delta}_{jk}^{(\ell)}, 0)$. For $\ell \in \mathcal{I}_0$, we define

$$\widehat{\alpha}^{(\ell)} = \underset{\widehat{\alpha}_{jk}^{(\ell)}}{\arg\min}\left\{|\widehat{\alpha}_{jk}^{(\ell)}| : \widehat{\alpha}_{jk}^{(\ell)}\,\mathrm{sign}(\alpha^{(\ell)}) > 0, (j,k) \in \mathcal{I}_\Delta^{(\ell)}\right\}.$$

Using (13), we estimate the common-variation matrix $\mathbf{C}_k$ by

$$\widehat{\mathbf{C}}_k = \widehat{\mathrm{cov}}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger\widehat{\mathbf{C}}^{\mathcal{I}_0},$$

where $\widehat{\mathrm{cov}}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0}) = n^{-1}\widehat{\mathbf{X}}_k(\widehat{\boldsymbol{z}}_k^{(\ell)})_{\ell\in\mathcal{I}_0} = \widehat{\mathbf{V}}_{xk}\widehat{\mathbf{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top$, $\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})$ is given in (15), $\widehat{\mathbf{C}}^{\mathcal{I}_0} = \widehat{\mathbf{A}}\widehat{\mathbf{N}}\widehat{\mathbf{F}}$, $\widehat{\mathbf{A}} = \mathrm{diag}\{(\widehat{\alpha}^{(\ell)}[\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f}))]^{-1/2})_{\ell\in\mathcal{I}_0}\}$ with $0/0 = 0$, and $\widehat{\mathbf{N}} = (\widehat{\boldsymbol{\eta}}^{(\ell)})_{\ell\in\mathcal{I}_0}^\top$. Finally, our estimator of $\mathbf{D}_k$ is given by $\widehat{\mathbf{D}}_k = \widehat{\mathbf{X}}_k - \widehat{\mathbf{C}}_k$.

The major time cost of proposed matrix estimators comes from the SVD of each $\mathbf{Y}_k$ with complexity $O(\min\{np_k^2, n^2p_k\})$.

3.2 Asymptotic properties

We introduce an assumption used in Wang and Fan (2017) and Shu et al. (2019).

Assumption 1 *We assume the following conditions for model* (14).

(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 that* $p_k > \kappa_0 n$ *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* $p_k^{1/2}(\log n)^{1/\gamma_{k2}} = o(\lambda_{r_k})$ *with* γ_{k2} *given in (v).*

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

(iv) *Matrix* $\mathbf{B}_k^\top \mathbf{B}_k$ *is a diagonal matrix. For all* $i \leq p_k$ *and* $\ell \leq r_k$, $|\mathbf{B}_k^{[i,\ell]}| \leq \kappa_B(\lambda_{k\ell}/p_k)^{1/2}$ *with a constant* $\kappa_B > 0$.

(v) *Denote* $\boldsymbol{e}_k = (e_{k1}, \ldots, e_{k,p_k})^\top$ *and* $\boldsymbol{f}_k = (f_{k1}, \ldots, f_{k,r_k})^\top$. *Let* $\|\mathrm{cov}(\boldsymbol{e}_k)\|_\infty < s_0$ *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* $\mathbb{P}(|f_{k\ell}| > t) \leq \exp\{-(t/b_{k2})^{\gamma_{k2}}\}$.

We have the following asymptotic properties of proposed estimators.

Theorem 5 *Suppose that Assumption 1 holds and true* $\{r_k\}_{k=1}^K$ *are given. Then for each* $k \leq K$, *we have*

$$\frac{\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_\star^2}{\|\mathbf{X}_k\|_\star^2} = O_P\Big(\min\Big\{\frac{1}{n^2} + \frac{\log p_k}{n\,\mathrm{SNR}_k}, 1\Big\}\Big),$$

where $\|\cdot\|_\star$ *denotes either the Frobenius norm or the spectral norm, and* $\mathrm{SNR}_k = \frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_k)\}}{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{e}_k)\}}$ *is the signal-to-noise ratio of* $\boldsymbol{y}_k$. *Additionally assume that* K *is a constant,* $\mathcal{I}_0 \neq \emptyset$, $\{\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))\}_{\ell=1}^L$ *are distinct, and true* $\big\{\mathcal{I}_0, \{r_k^*\}_{k=1}^K, \{\mathcal{I}_\Delta^{(\ell)}, \mathrm{sign}(\alpha^{(\ell)})\}_{\ell\in\mathcal{I}_0}\big\}$ *are given. If* $\delta_\eta = \frac{1}{\sqrt{n}} + \sum_{k=1}^K \sqrt{\frac{\log p_k}{n\,\mathrm{SNR}_k}} = o(1)$, *then*

$$\max\left\{\frac{\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_\star^2}{\|\mathbf{X}_k\|_\star^2}, \frac{\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\|_\star^2}{\|\mathbf{X}_k\|_\star^2}\right\} = O_P(\delta_\eta), \tag{16}$$

$$\left|\frac{\|\widehat{\mathbf{C}}_k\|_F^2}{\|\widehat{\mathbf{X}}_k\|_F^2} - \frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c}_k)\}}{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_k)\}}\right| = O_P(\delta_\eta^{1/2}).$$

When $K = 2$, the error bounds of $\widehat{\mathbf{C}}_k$ and $\widehat{\mathbf{D}}_k$ in (16) are equivalent to those in Theorem 3 of the D-CCA paper (Shu et al., 2019). The quantity $\mathrm{PVE}(\boldsymbol{c}_k) := \mathrm{tr}\{\mathrm{cov}(\boldsymbol{c}_k)\}/\,\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_k)\}$ is the proportion of $\boldsymbol{x}_k$'s variance explained by $\boldsymbol{c}_k$, which reflects the influence of $\boldsymbol{c}_k$ on $\boldsymbol{x}_k$. Following Smilde et al. (2017), $[1 - \mathrm{PVE}(\boldsymbol{c}_k)]$ can be interpreted as the extra proportion of $\boldsymbol{x}_k$'s variance that is explained by adding the distinctive-variation vector $\boldsymbol{d}_k$.

3.3 Selection of nuisance parameters

We discuss how to practically select the parameters $\{r_k\}_{k=1}^K$, $\mathcal{I}_0$, $\{r_k^*\}_{k=1}^K$ $\{\mathcal{I}_\Delta^{(\ell)}\}_{\ell\in\mathcal{I}_0}$, and $\{\mathrm{sign}(\alpha^{(\ell)})\}_{\ell\in\mathcal{I}_0}$.

Denote $\widehat{r}_k, \widehat{L}$, $\widehat{\mathcal{I}}_0$, $\widehat{r}_k^*$, $\widehat{\mathcal{I}}_\Delta^{(\ell)}$, $\widehat{\mathrm{sign}}(\alpha^{(\ell)})$ to be estimators of their true counterparts. We select $\{\widehat{r}_k\}_{k=1}^K$ by using the edge distribution method of Onatski (2010) that consistently estimates the rank for the factor model in (14) under mild conditions. To determine the other parameters, we use hypothesis tests based on the denoised data $\{\widehat{\mathbf{X}}_k\}_{k=1}^K$. Testing

procedures have been widely used in the literature of CCA (Bartlett, 1941; Lawley, 1959; Caliński and Krzyśko, 2005; Song et al., 2016) to select similar parameters.

Consider the selection of $L = \max\{\ell \in \{1, \ldots, r_f\} : \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) > 1\}$. Left-multiplying the both sides of $[\mathrm{cov}(\boldsymbol{f})\boldsymbol{\eta}^{(\ell)}]^{[\sum_{i=0}^{k-1} r_i : \sum_{i=1}^{k} r_i]} = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))\boldsymbol{\eta}^{(\ell)}]^{[\sum_{i=0}^{k-1} r_i : \sum_{i=1}^{k} r_i]}$ by $\boldsymbol{\eta}_k^{(\ell)}$ can obtain $\mathrm{cov}\left((\boldsymbol{\eta}_k^{(\ell)})^\top \boldsymbol{f}_k, \sum_{j\neq k}(\boldsymbol{\eta}_j^{(\ell)})^\top \boldsymbol{f}_j\right) = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f})) - 1]\,\|\boldsymbol{\eta}_k^{(\ell)}\|_F^2$ for all $k \le K$. Let $\widehat{L}$ be the largest $\ell \in [0, \mathrm{rank}\left(\widehat{\mathrm{cov}}(\boldsymbol{f})\right)]$ such that for at least one k, both $\mathrm{corr}(w^{(\ell)}, z_k^{(\ell)}) = 0$ and $\mathrm{corr}\left((\boldsymbol{\eta}_k^{(\ell)})^\top \boldsymbol{f}_k, \sum_{j\neq k}(\boldsymbol{\eta}_j^{(\ell)})^\top \boldsymbol{f}_j\right) = 0$ are rejected by a right-tailed test for zero correlation. The two tests indicate $\|\boldsymbol{\eta}_k^{(\ell)}\|_F \neq 0$ and $\mathrm{cov}\left((\boldsymbol{\eta}_k^{(\ell)})^\top \boldsymbol{f}_k, \sum_{j\neq k}(\boldsymbol{\eta}_j^{(\ell)})^\top \boldsymbol{f}_j\right) > 0$, respectively, thereby implying $\lambda_\ell(\mathrm{cov}(\boldsymbol{f})) - 1 = \mathrm{cov}\left((\boldsymbol{\eta}_k^{(\ell)})^\top \boldsymbol{f}_k, \sum_{j\neq k}(\boldsymbol{\eta}_j^{(\ell)})^\top \boldsymbol{f}_j\right)/\|\boldsymbol{\eta}_k^{(\ell)}\|_F^2 > 0$. We use the normal approximation test of DiCiccio and Romano (2017) for testing zero correlation.

To determine $\mathcal{I}_0 = \{\ell \in \{1, \ldots, L\} : \alpha^{(\ell)} \neq 0\}$, we retain index $\ell \le \widehat{L}$ in $\widehat{\mathcal{I}}_0$ if $\mathrm{corr}(w^{(\ell)}, z_k^{(\ell)}) = 0$ and $\mathrm{corr}(z_j^{(\ell)}, z_k^{(\ell)}) = 0$ are rejected respectively by the right-tailed and the two-tailed zero-correlation tests for all $k \le K$ and all $j \neq k$.

The rank estimate $\widehat{r}_k^*$ of $r_k^* = \mathrm{rank}(\mathrm{cov}(z_k^{\mathcal{I}_0}))$ is obtained by the two-step test of Chen and Fang (2019) for the rank of matrix $\mathrm{cov}(z_k^{\widehat{\mathcal{I}}_0})$.

We next select $\mathcal{I}_\Delta^{(\ell)} = \{(j,k) : \Delta_{jk}^{(\ell)} \ge 0, 1 \le j < k \le K\}$. An equivalent formula of $\Delta_{jk}^{(\ell)} = 0$ is $\cos\{\theta(z_{j,k}^{(\ell)}, z_{k,j}^{(\ell)})\} = 0$ with $z_{j,k}^{(\ell)} = z_j^{(\ell)} - 0.5[\cos\{\theta(w^{(\ell)}, z_j^{(\ell)})\} + \cos\{\theta(w^{(\ell)}, z_k^{(\ell)})\}]w^{(\ell)}$. For $\ell \in \widehat{\mathcal{I}}_0$, we exclude (j,k) from $\widehat{\mathcal{I}}_\Delta^{(\ell)}$ if $\widetilde{\Delta}_{jk}^{(\ell)} < 0$ and meanwhile $\mathrm{corr}(z_{j,k}^{(\ell)}, z_{k,j}^{(\ell)}) = 0$ is rejected by the two-tailed zero-correlation test.

Finally, consider to determine the sign of $\alpha^{(\ell)}$. Define $\alpha_+^{(\ell)} = \min\{\alpha_{jk}^{(\ell)} : \alpha_{jk}^{(\ell)} > 0, (j,k) \in \widehat{\mathcal{I}}_\Delta^{(\ell)}\}$ and $\alpha_-^{(\ell)} = \max\{\alpha_{jk}^{(\ell)} : \alpha_{jk}^{(\ell)} < 0, (j,k) \in \widehat{\mathcal{I}}_\Delta^{(\ell)}\}$, and define $\widehat{\alpha}_+^{(\ell)}$ and $\widehat{\alpha}_-^{(\ell)}$ in the same way by using $\widehat{\alpha}_{jk}^{(\ell)}$ instead. Let $\widehat{\mathrm{sign}}(\alpha^{(\ell)})$ be the sign of the existing one of $\widehat{\alpha}_+^{(\ell)}$ and $\widehat{\alpha}_-^{(\ell)}$ if the other does not exist. Otherwise, first test $|\alpha_+^{(\ell)}| - |\alpha_-^{(\ell)}| = 0$ by applying the bias-corrected and accelerated bootstrap interval (Efron and Tibshirani, 1993). Let $\widehat{\mathrm{sign}}(\alpha^{(\ell)}) = 1$ if zero is outside the bootstrap interval and $|\widehat{\alpha}_+^{(\ell)}| < |\widehat{\alpha}_-^{(\ell)}|$, and otherwise let $\widehat{\mathrm{sign}}(\alpha^{(\ell)}) = -1$.

4. Simulation studies

In this section, we evaluate the finite-sample performance of proposed D-GCCA estimation via simulations, comparing to the six competing methods mentioned in Section 1.

4.1 Simulation setups

We consider $K = 3$ datasets with signals $\{\boldsymbol{x}_k\}_{k=1}^3$ following the four simulation setups below, and generate signal-independent Gaussian noises $\{e_{ki}\}_{i=1}^{p_k} \overset{iid}{\sim} N(0, \sigma_{e_k}^2)$ that are independent across datasets. Simulations are conducted with sample size $n = 300$, variable dimension p_1 ranging from 100 to 1500, noise variance $\sigma_{e_1}^2$ from 0.25 to 9, and 1000 replications under each setting.

- Setup 1.1: Let $\boldsymbol{x}_1 \overset{d}{=} \boldsymbol{x}_2 \overset{d}{=} \boldsymbol{x}_3$ and $r_1 = r_2 = r_3 = 1$. Set standardized canonical variables $z_1^{(1)}, z_2^{(1)}, z_3^{(1)}$ to be jointly Gaussian with $\theta_z := \theta(z_j^{(1)}, z_k^{(1)})$ for all $j \neq k$. Let $\boldsymbol{\Lambda}_k = 500$

for $k = 1, 2, 3$. Randomly generate three unit vectors $\{\mathbf{V}_{xk}\}_{k=1}^3$ that are equal if with the same size and are fixed for all simulation replications of the same (p_1, p_2, p_3). Let $\boldsymbol{x}_k = \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2} z_k^{(1)}$. We vary θ_z from 10° to 70°, resulting in D-GCCA's $\{\text{PVE}(\boldsymbol{c}_k)\}_{k=1}^3$ all from 0.853 to 0.079 invariant to $\{p_k\}_{k=1}^3$; see Appendix B. Let $\sigma_{e_1}^2 = \sigma_{e_2}^2 = \sigma_{e_3}^2$.

- Setup 1.2: Fix variable dimensions $(p_2, p_3) = (300, 900)$ and noise variances $\sigma_{e_2}^2 = \sigma_{e_3}^2 = 1$. The other settings are the same as in Setup 1.1.

- Setup 2.1: Let $p_1 = p_2 = p_3$ and $r_1 = r_2 = r_3 = 5$. Design $\text{cov}(\boldsymbol{f})$ with eigenvalues $(3, 2.8, 2.25, 1.5, 1, 1, 1, 1, 0.635, 0.415, 0.4, 0, 0, 0, 0)$ such that, respectively for $\ell = 1, \ldots, 4$, $\{\theta(w^{(\ell)}, z_k^{(\ell)})\}_{k\le 3}$ are all approximately 0°, 15°, 30°, and 45°, and $\{\theta(z_j^{(\ell)}, z_k^{(\ell)})\}_{j<k\le 3}$ are all close to $0°, 25°, 50°$ and 75°. Matrix $\text{cov}(\boldsymbol{f})$ is given in Appendix B. Set $\boldsymbol{f}$ to be multivariate Gaussian with mean zero and covariance matrix $\text{cov}(\boldsymbol{f})$. Let $\boldsymbol{\Lambda}_k = \text{diag}(500, 400, 300, 200, 100)$ for all $k \le 3$. Randomly generate three matrices $\{\mathbf{V}_{xk}\}_{k=1}^3$, each with orthonormal columns, which are equal if with the same size and are fixed for all simulation replications of the same (p_1, p_2, p_3). Let $\boldsymbol{x}_k = \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\boldsymbol{f}_k$. D-GCCA has $(\text{PVE}(\boldsymbol{c}_1), \text{PVE}(\boldsymbol{c}_2), \text{PVE}(\boldsymbol{c}_3)) = (0.387, 0.324, 0.427)$ invariant to $\{p_k\}_{k=1}^3$. Let $\sigma_{e_1}^2 = \sigma_{e_2}^2 = \sigma_{e_3}^2$.

- Setup 2.2: Fix $(p_2, p_3) = (300, 900)$ and $\sigma_{e_2}^2 = \sigma_{e_3}^2 = 1$. The other settings are the same as in Setup 2.1.

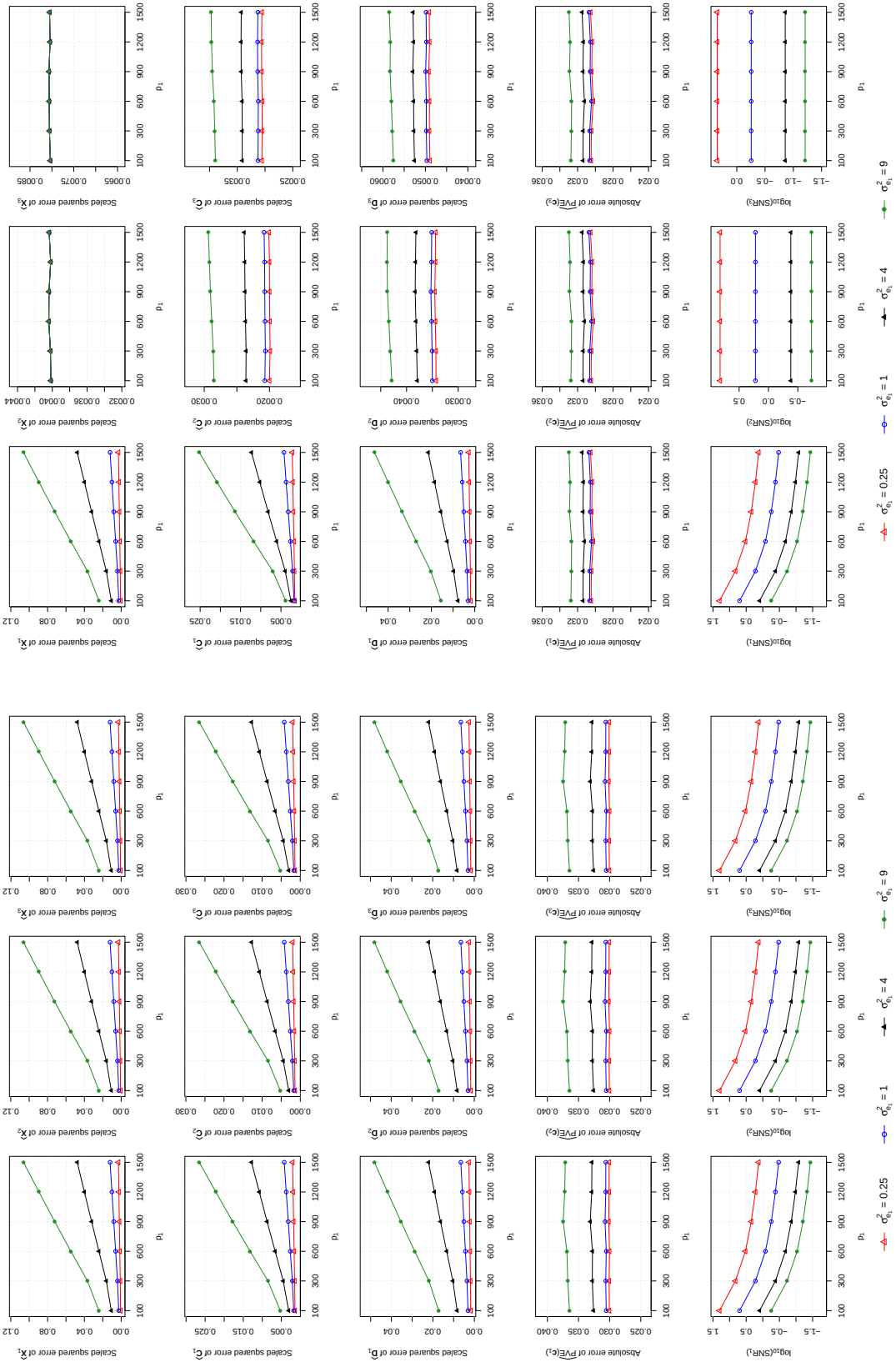
4.2 Finite-sample performance of D-GCCA estimators

We first apply the four error metrics in Theorem 5 to evaluate the performance of the D-GCCA estimation with true nuisance parameters $\{\{r_k, r_k^*\}_{k=1}^K, \mathcal{I}_0, \{\mathcal{I}_\Delta^{(\ell)}, \text{sign}(\alpha^{(\ell)})\}_{\ell\in\mathcal{I}_0}\}$. The practical selection of these nuisance parameters has been discussed in Section 3.3 and its performance is investigated later in this subsection. It is easily seen that $\text{SNR}_k = \text{tr}(\boldsymbol{\Lambda}_k)/(p_k\sigma_{e_k}^2)$ in the above simulation setups. For simplicity, we hence examine the trend of estimation errors in Theorem 5 with respect to $(p_k, \sigma_{e_k}^2)$ instead of (p_k, SNR_k).

Figure 1 shows the four estimation errors of D-GCCA in the Frobenius norm under Setups 1.1 and 1.2 with $\theta_z = 50°$. For Setup 1.1 in Figure 1(a), the average estimation errors are almost the same for the three identically distributed datasets, indicating the fair treatment of proposed estimation to each dataset. As expected in Theorem 5, the errors generally increase as either dimension p_1 or noise variance $\sigma_{e_1}^2$ grows, and the slow error trend of $\widehat{\text{PVE}}(\boldsymbol{c}_k) = \|\widehat{\mathbf{C}}_k\|_F^2/\|\widehat{\mathbf{X}}_k\|_F^2$ reflects its slow convergence rate. The errors are acceptable even for some cases when p_1 or $\sigma_{e_1}^2$ is large along with very low SNR_k. For example, the errors are smaller than 0.05 at $(p_1, \sigma_{e_1}^2) = (1500, 4)$ with $\text{SNR}_k = 0.083$.

In Figure 1(b) for Setup 1.2, the estimation result of the first dataset is similar to that in Figure 1(a). As for the second and third datasets with fixed variable dimensions and noise variances, when $(p_1, \sigma_{e_1}^2)$ the parameters of the first dataset grow, the signal matrix estimation is not affected, while the estimation errors of the other three quantities are observed with slight increasing trends. These results are consistent with those shown in Theorem 5.

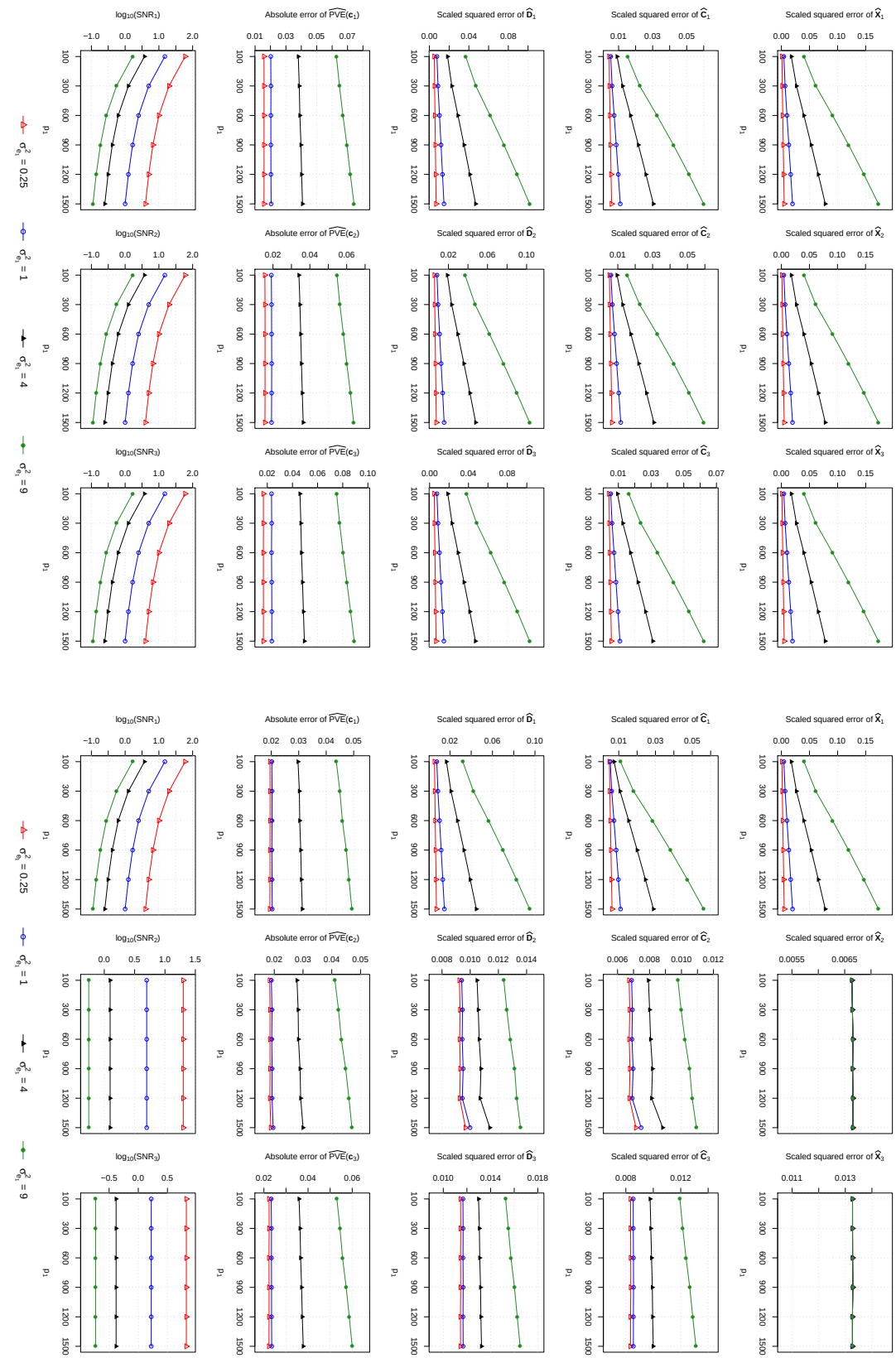
Figure 2 presents similar results as in Figure 1 but for higher-dimensional signal subspaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^3$ under Setups 2.1 and 2.2. The above result analysis generally holds



(a) Setup 1.1 with $\theta_z = 50°$

(b) Setup 1.2 with $\theta_z = 50°$

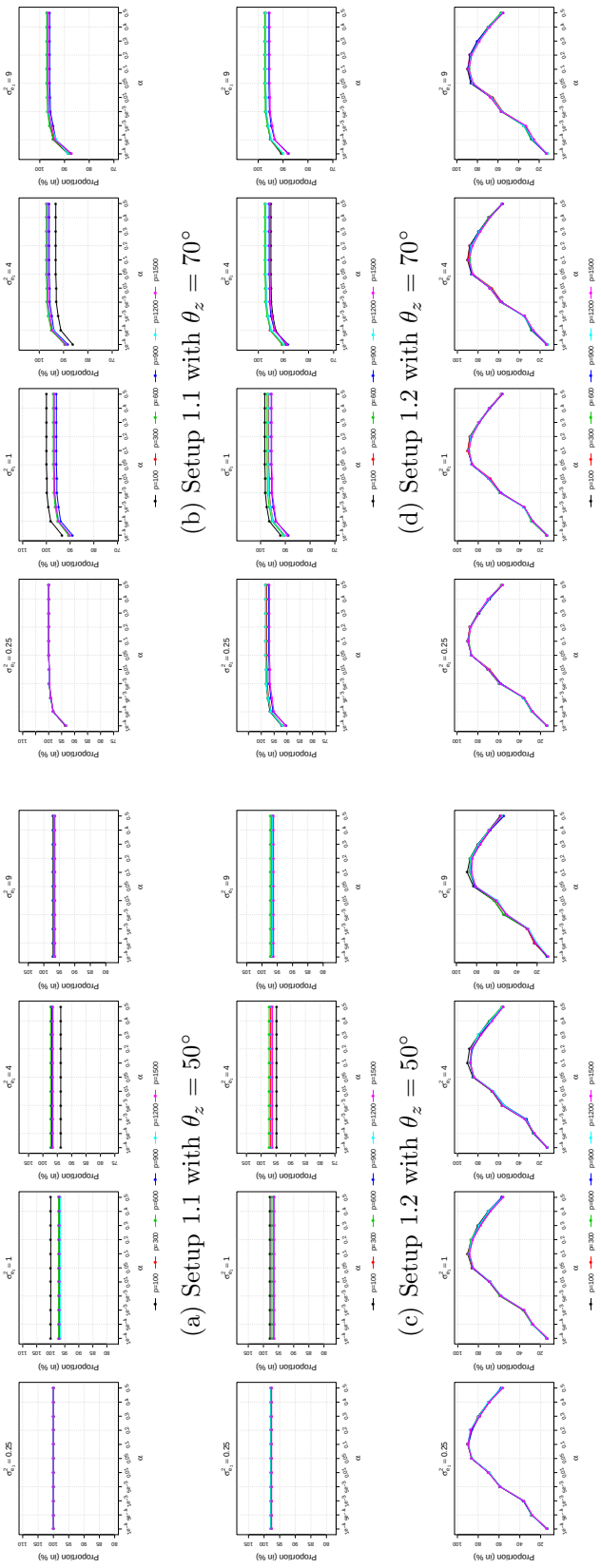
Figure 1: Average errors of D-GCCA estimates over 1000 replications in the Frobenius norm for Setup 1.1 and Setup 1.2 using true nuisance parameters.



(a) Setup 2.1

(b) Setup 2.2

Figure 2: Average errors of D-GCCA estimates over 1000 replications in the Frobenius norm for Setup 2.1 and Setup 2.2 using true nuisance parameters.



(a) Setup 1.1 with $\theta_z = 50°$

(b) Setup 1.1 with $\theta_z = 70°$

(c) Setup 1.2 with $\theta_z = 50°$

(d) Setup 1.2 with $\theta_z = 70°$

(e) Setup 2.1

(f) Setup 2.2

Figure 3: The proportion of 1000 simulation replications where all nuisance parameters of D-GCCA are correctly selected. The nuisance parameters are selected using the approach described in Section 3.3 with a significance level α uniformly applied to all tests.

for Setups 1.1 and 1.2 with other θ_z values and also for all the four setups with estimation errors in the spectral norm. These additional results are provided in Appendix B.

We also numerically evaluate the selection approach of nuisance parameters that is proposed in Section 3.3. Figures 3, 10 and 11 show the accuracy of the selection approach for the four simulation setups. For simplicity, we apply the same significance level α, ranging from 0.5 down to 0.0001, to all hypothesis tests involved in the selection approach. For Setups 1.1 and 1.2, $\alpha \in [0.0001, 0.5]$ and $\alpha \in [0.005, 0.5]$ perform the same well for $\theta_z \in [10°, 60°]$ and $\theta_z = 70°$, respectively, with accuracy values all above 90% and most around or above 95%. As for Setups 2.1 and 2.2, as shown in Figure 3 (e) and (f), when the significance level is 0.1, the accuracy achieves nearly 90% for most considered cases. There is no dramatic change when the significance level is down from 0.2 to 0.05. In practice, it is worth trying several significance levels to monitor the change of nuisance parameters, and also suggested to report the used significance level along with the obtained decomposition. One may also expect to potentially improve the accuracy by additionally using the Bagging technique (Hastie et al., 2009), that is, for each nuisance parameter applying the selection approach to a large number of resampled datasets and then combining the results by majority voting. We leave this to interested readers.

4.3 Comparison to related methods

We now compare the performance of D-GCCA and the six competing methods (JIVE, R.JIVE, AJIVE, COBE, OnPLS, and DISCO-SCA) under the four simulation setups.

Since the decompositions defined by the seven methods are different, it is unfair to compare the errors of their matrix estimates to D-GCCA's true matrices. Alternatively, under the general model given in (1) and (2), for each method we consider whether at least one orthogonal pair among $\{\boldsymbol{d}_k\}_{k=1}^K$ exists, and otherwise how large the common variation is retained among $\{\boldsymbol{d}_k\}_{k=1}^K$.

The orthogonality between $\boldsymbol{d}_j$ and $\boldsymbol{d}_k$ is equivalent to $\sum_{m=1}^{r_{d_j}} \sum_{\ell=1}^{r_{d_k}} [\mathrm{corr}(d_j^{(m)}, d_k^{(\ell)}) \neq 0] = 0$, where $\{d_k^{(\ell)}\}_{\ell=1}^{r_{d_k}}$ denote the latent factors of $\boldsymbol{d}_k$. We detect each $\mathrm{corr}(d_j^{(m)}, d_k^{(\ell)}) \neq 0$ using the normal approximation test (DiCiccio and Romano, 2017), with false discovery rate controlled at 0.05 (Benjamini and Hochberg, 1995) and the ℓ-th right-singular vector of $\widehat{\mathbf{D}}_k$ used as the n samples of $d_k^{(\ell)}$.

Let $\rho_\ell(\{\boldsymbol{x}_k\}_{k=1}^K)$ be the maximum of the objective function in (3). If no pairs in $\{\boldsymbol{d}_k\}_{k=1}^K$ are orthogonal, we use $\rho_1(\{\boldsymbol{d}_k\}_{k=1}^K) \in [1, K]$ to measure the amount of common variation retained among $\{\boldsymbol{d}_k\}_{k=1}^K$. From equation (6), we estimate $\rho_1(\{\boldsymbol{d}_k\}_{k=1}^K)$ by $\widehat{\rho}_1(\{\boldsymbol{d}_k\}_{k=1}^K) = \lambda_1(\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\top/n)$ with the matrix $\widehat{\mathbf{F}}$ that is defined in Section 3.1 but uses $\{\widehat{\mathbf{D}}_k\}_{k=1}^K$ here instead of $\{\widehat{\mathbf{X}}_k\}_{k=1}^K$.

Table 1 reports the comparison results for Setups 1.1 and 1.2 with $(p_1, \theta_z, \sigma_{e_1}^2) = (600, 50°, 1)$ and Setups 2.1 and 2.2 with $(p_1, \sigma_{e_1}^2) = (600, 1)$. We first observe that all simulation replications of R.JIVE for the four setups have at least one orthogonal pair among $\{\boldsymbol{d}_k\}_{k=1}^3$, but its scaled squared errors of signal matrix estimates are much larger than those of JIVE (its original version with no orthogonality constraint on $\{\boldsymbol{d}_k\}_{k=1}^K$) and our D-GCCA. This agrees with the design of R.JIVE, which can discard some signal as noise to ensure the orthogonality of $\{\boldsymbol{d}_k\}_{k=1}^K$. For Setups 1.1 and 1.2 with three one-dimensional signal subspaces $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^3$, our D-GCCA also has all its simulation replications satis-

Setup	Method	≥ 1 orth. pair	$\widehat{\rho}_1(\{\boldsymbol{d}_k\}_{k=1}^3)$	$\frac{\|\widehat{\mathbf{X}}_1-\mathbf{X}_1\|_F^2}{\|\mathbf{X}_1\|_F^2}$, $\frac{\|\widehat{\mathbf{X}}_2-\mathbf{X}_2\|_F^2}{\|\mathbf{X}_2\|_F^2}$, $\frac{\|\widehat{\mathbf{X}}_3-\mathbf{X}_3\|_F^2}{\|\mathbf{X}_3\|_F^2}$
	D-GCCA1	**100%**	1.10 (0.05)	0.006 (6.0e-4), 0.006 (5.9e-4), 0.006 (5.6e-4)
	D-GCCA2	**100%**	1.10 (0.05)	0.006 (1.0e-3), 0.006 (1.1e-3), 0.006 (1.6e-3)
Setup 1.1	JIVE	0%	2.22 (0.06)	0.014 (1.4e-3), 0.014 (1.4e-3), 0.014 (1.3e-3)
($p_1 = 600$,	R.JIVE	**100%**	1.00 (0.00)	**0.032(1.0e-2), 0.021(3.1e-3), 0.023(7.1e-3)**
$\theta_z = 50°$,	AJIVE	0% (zero $\widehat{\mathbf{C}}_k$s)	2.28 (0.05)	0.006 (6.0e-4), 0.006 (6.0e-4), 0.006 (5.6e-4)
$\sigma_{e_1}^2 = 1$)	COBE	0% (zero $\widehat{\mathbf{C}}_k$s)	2.28 (0.05)	0.006 (6.0e-4), 0.006 (6.0e-4), 0.006 (5.6e-4)
	OnPLS	1.1%	1.87 (0.05)	0.026 (2.3e-3), 0.026 (2.3e-3), 0.026 (2.2e-3)
	DISCO-SCA	0% (zero $\widehat{\mathbf{C}}_k$s)	3.00 (0.00)	0.014 (1.3e-3), 0.014 (1.3e-3), 0.014 (1.3e-3)
	D-GCCA1	**100%**	1.10 (0.05)	0.006 (6.0e-4), 0.004 (4.1e-4), 0.008 (7.3e-4)
	D-GCCA2	**100%**	1.10 (0.05)	0.006 (1.0e-3), 0.004 (7.9e-4), 0.008 (1.7e-3)
Setup 1.2	JIVE	0%	2.20 (0.06)	0.014 (1.4e-3), 0.009 (1.0e-3), 0.018 (1.6e-3)
($p_1 = 600$,	R.JIVE	**100%**	1.00 (0.00)	**0.033(1.0e-2), 0.014(2.3e-3), 0.029(6.7e-3)**
$\theta_z = 50°$,	AJIVE	0% (zero $\widehat{\mathbf{C}}_k$s)	2.28 (0.05)	0.006 (6.0e-4), 0.004 (4.1e-4), 0.008 (7.3e-4)
$\sigma_{e_1}^2 = 1$)	COBE	0% (zero $\widehat{\mathbf{C}}_k$s)	2.28 (0.05)	0.006 (6.0e-4), 0.004 (4.1e-4), 0.008 (7.3e-4)
	OnPLS	0.9%	1.83 (0.05)	0.026 (2.4e-3), 0.018 (1.6e-3), 0.026 (2.2e-3)
	DISCO-SCA	0% (zero $\widehat{\mathbf{C}}_k$s)	3.00 (0.00)	0.014 (1.3e-3), 0.008 (7.6e-4), 0.020 (1.8e-3)
	D-GCCA1	0%	**2.13 (0.05)**	0.010 (4.5e-4), 0.010 (4.5e-4), 0.010 (4.8e-4)
	D-GCCA2	0%	**2.14 (0.06)**	0.010 (4.5e-4), 0.010 (4.5e-4), 0.010 (4.8e-4)
Setup 2.1	JIVE	0%	2.52 (0.21)	0.016 (2.0e-3), 0.016 (2.2e-3), 0.016 (2.1e-3)
($p_1 = 600$,	R.JIVE	**100%**	1.00 (0.00)	**0.076(4.3e-2), 0.080(4.9e-2), 0.065(3.4e-2)**
$\sigma_{e_1}^2 = 1$)	AJIVE	0%	2.80 (0.02)	0.010 (4.4e-4), 0.010 (4.3e-4), 0.010 (4.7e-4)
	COBE	0%	2.80 (0.02)	0.010 (4.6e-4), 0.010 (4.6e-4), 0.010 (4.8e-4)
	OnPLS	0.1%	2.65 (0.18)	0.014 (1.7e-3), 0.014 (3.1e-3), 0.015 (1.8e-3)
	DISCO-SCA	NA	NA	NA
	D-GCCA1	0%	**2.13 (0.05)**	0.010 (4.5e-4), 0.007 (3.2e-4), 0.013 (6.1e-4)
	D-GCCA2	0%	**2.14 (0.06)**	0.010 (4.5e-4), 0.007 (3.2e-4), 0.013 (6.1e-4)
Setup 2.2	JIVE	0%	2.41 (0.26)	0.016 (2.3e-3), 0.010 (1.4e-3), 0.021 (3.1e-3)
($p_1 = 600$,	R.JIVE	**100%**	1.00 (0.00)	**0.064(4.0e-2), 0.063(5.0e-2), 0.079(4.3e-2)**
$\sigma_{e_1}^2 = 1$)	AJIVE	0%	2.80 (0.02)	0.010 (4.4e-4), 0.006 (3.0e-4), 0.013 (6.1e-4)
	COBE	0%	2.80 (0.02)	0.010 (4.6e-4), 0.007 (3.2e-4), 0.013 (6.2e-4)
	OnPLS	0.5%	2.51 (0.18)	0.015 (6.5e-3), 0.009 (1.3e-3), 0.020 (2.6e-3)
	DISCO-SCA	NA	NA	NA

Table 1: The proportions of replications with at least one orthogonal pair among $\{\boldsymbol{d}_k\}_{k=1}^3$, averages (SDs) of $\widehat{\rho}_1(\{\boldsymbol{d}_k\}_{k=1}^3)$, and averages (SDs) of scaled squared errors of signal matrix estimates over 1000 simulation replications. D-GCCA1: the D-GCCA using true nuisance parameters. D-GCCA2: the D-GCCA using nuisance parameters selected by the approach in Section 3.3. NA: not available due to out of the 24-hour time limit on a CPU core (up to 3.0GHz) per simulation replication. By the paired t-test, both D-GCCA1 and D-GCCA2 have significantly different mean $\widehat{\rho}_1(\{\boldsymbol{d}_k\}_{k=1}^3)$ values from those of all the other methods with p-values<1e-10.

fying the desirable orthogonality among $\{\boldsymbol{d}_k\}_{k=1}^3$, which is consistent with its decomposition in (10) for canonical variables. In contrast, the other five methods do not show the desirable orthogonality for all replications under the four setups. For Setups 2.1 and 2.2 with higher-dimensional signal subspaces, neither does D-GCCA own the desirable orthogonality, as explained in Remark 2 due to its relaxation into each subproblem (10), but D-GCCA still has significantly smaller mean $\widehat{\rho}_1(\{\boldsymbol{d}_k\}_{k=1}^K)$ values than those available for the five methods.

5. Real-world Data Examples

5.1 Application to TCGA breast cancer genomic datasets

We compare our D-GCCA with the state-of-the-art methods in analyzing the TCGA breast cancer genomic data (Koboldt et al., 2012). We consider three datasets on a common set of 664 tumor samples that contain mRNA expression data for the top 2930 variably expressed genes, miRNA expression data for 526 highly variant miRNAs, and DNA methylation data for 3067 most variable probes, respectively, following the preprocessing procedure of Lock and Dunson (2013). The tumor samples are categorized by the classic PAM50 model (Parker et al., 2009) into four intrinsic subtypes that are relevant with clinical outcomes, including 111 Basal-like, 56 HER2-enriched, 346 Luminal A, and 151 Luminal B tumors. The PAM50 intrinsic subtypes are defined by mRNA expression only. We investigate whether these intrinsic subtypes are also characterized by other data types such as DNA methylation and miRNA expression that represent different biological components. In particular, we study the relationship between the PAM50 intrinsic subtypes and the common and distinctive underlying mechanisms of the three genomic datasets by evaluating the ability of their corresponding matrices in (1) to separate the four intrinsic subtypes.

Each observed data matrix is row-centered by subtracting the average within each row. The nuisance parameters of our D-GCCA method are selected by using the approach described in Section 3.3. The selection approach yields the same decomposition by the choices 0.2 and 0.0001 for the significance level uniformly applied to all involved hypothesis tests. The values $(\text{rank}(\widehat{\mathbf{X}}_k), \text{rank}(\widehat{\mathbf{C}}_k), \text{rank}(\widehat{\mathbf{D}}_k), \|\widehat{\mathbf{C}}_k\|_F^2/\|\widehat{\mathbf{X}}_k\|_F^2)$ from the D-GCCA method are $(4, 2, 4, 0.239)$, $(3, 2, 3, 0.184)$ and $(3, 2, 3, 0.147)$ for the mRNA, miRNA, and DNA datasets, respectively. To quantify the subtype separation in a matrix, we adopt the SWISS score of Cabanski et al. (2010) that calculates the standardized within-subtype sum of squares: For a matrix $\mathbf{M} = (M_{ij})_{p\times n}$,

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

where $\bar{M}_{i,s(j)}$ is the average of the j-th sample's subtype on the i-th row, and $\bar{M}_{i,\cdot}$ is the average of the i-th row's elements. The lower score indicates better subtype separation.

Table 2 shows the SWISS scores computed for the D-GCCA method and also the six competing methods mentioned in Section 1. The denoised signal matrix $\widehat{\mathbf{X}}_k$ from all methods gains an improved ability on subtype separation with a smaller score, comparing to the noisy data matrix $\mathbf{Y}_k$. All methods, except AJIVE and COBE, discover nonzero common-variation matrices, and show a clear pattern of decreasing SWISS scores from $\widehat{\mathbf{D}}_k$ to $\widehat{\mathbf{X}}_k$

and then to $\widehat{\mathbf{C}}_k$. This pattern indicates that the four PAM50 intrinsic subtypes are more likely to be an inherent feature of the common mechanism underlying the three different genomic datasets. Moreover, our D-GCCA method has the lowest scores for estimated common-variation matrices when compared with the other methods. The result analysis remains the same even when our D-GCCA's $\widehat{\mathbf{X}}_k$s, which have the smallest SWISS scores among all signal estimates, are used as the input data for the other six methods.

The better SWISS scores of D-GCCA for common-variation matrix estimates indicate its enhanced ability to capture the common mechanism than the other methods, which benefits from our well designed orthogonality constraint on distinctive mechanisms. Table 3 further verifies this conclusion, and shows that significant nonzero correlations do not exist between D-GCCA's $\boldsymbol{d}_{\text{miRNA}}$ and $\boldsymbol{d}_{\text{DNA}}$ but account for over 15% among all pairs of $\boldsymbol{d}_k$s from the other methods except R.JIVE. However, R.JIVE enforces the orthogonality of $\boldsymbol{d}_k$s by sacrificing its unexplainable signal to be noise. This can be seen in Table 2, where R.JIVE has slightly lower SWISS scores for $\widehat{\mathbf{E}}_k$s than JIVE, its original version with no orthogonality constraint on $\boldsymbol{d}_k$s, and moreover has nonzero $\widehat{\mathbf{E}}_k$s when using low-rank D-GCCA's signal estimates as the input data.

Method	$\widehat{\mathbf{X}}_k$	$\widehat{\mathbf{C}}_k$	$\widehat{\mathbf{D}}_k$	$\widehat{\mathbf{E}}_k$
D-GCCA	48.0, 62.7, 73.6	**21.5**‡, **21.2**, **26.8**$^{\sharp}$	74.2, 84.9, 93.2	99.0, 98.4, 98.3
JIVE	74.0, 80.0, 82.5	65.6, 65.3, 58.9	86.1, 87.9, 92.1	99.8, 99.6, 99.7
R.JIVE	74.5, 74.7, 80.8	41.7, 38.1, 64.6	93.3, 99.7, 99.6	99.8, 97.0, 97.6
AJIVE	48.2, 62.7, 73.6	NA, NA, NA	48.2, 62.7, 73.6	99.0, 98.4, 98.3
COBE	48.2, 62.7, 73.6	NA, NA, NA	48.2, 62.7, 73.6	99.0, 98.4, 98.3
OnPLS	60.0, 70.8, 78.1	36.4, 34.1, 36.4	89.6, 95.8, 98.6	99.5, 98.9, 99.6
DISCO-SCA	56.7, 67.4, 75.0	52.6, 53.0, 48.5	99.0, 97.7, 99.3	99.4, 99.5, 99.6
JIVE*	48.0, 62.7, 73.6	35.0, 33.0, 50.8	89.0, 93.8, 97.3	NA, NA, NA
R.JIVE*	47.6, 60.2, 72.2	34.0, 28.5, 61.4	84.7, 98.7, 99.4	99.3, 84.7, 83.5
AJIVE*	48.0, 62.7, 73.6	NA, NA, NA	48.0, 62.7, 73.6	NA, NA, NA
COBE*	48.0, 62.7, 73.6	NA, NA, NA	48.0, 62.7, 73.6	NA, NA, NA
OnPLS*	48.0, 62.7, 73.6	22.6‡, 26.4, 30.5	75.1, 87.8, 94.0	NA, NA, NA
DISCO-SCA*	48.0, 62.7, 73.6	28.0, 28.0, 28.0$^{\sharp}$	77.9, 82.7, 94.9	NA, NA, NA
$\mathbf{Y}_k$	84.8, 87.8, 90.0			

Table 2: SWISS scores (in %) for TCGA breast cancer genomic datasets (k = mRNA, miRNA, DNA). Lower SWISS scores indicate better subtype separation. Methods suffixed with * use D-GCCA's $\widehat{\mathbf{X}}_k$s instead of $\mathbf{Y}_k$s as the input data. NA: not available due to a zero matrix estimate. All methods have SWISS($\widehat{\mathbf{X}}_k$) < SWISS($\mathbf{Y}_k$) for each k. Except AJIVE and COBE with $\widehat{\mathbf{C}}_k = \mathbf{0}$, all the other methods have SWISS($\widehat{\mathbf{C}}_k$) < SWISS($\widehat{\mathbf{X}}_k$) < SWISS($\widehat{\mathbf{D}}_k$) for each k. Our D-GCCA has the lowest SWISS($\widehat{\mathbf{C}}_k$) for all k. By the test of Cabanski et al. (2010), all above comparisons of SWISS scores are significantly different with p-values<0.001, except for the two annotated respectively by ‡ and ♯ with p-values>0.05.

Method	$\boldsymbol{d}_{\text{mRNA}}$ & $\boldsymbol{d}_{\text{miRNA}}$	$\boldsymbol{d}_{\text{mRNA}}$ & $\boldsymbol{d}_{\text{DNA}}$	$\boldsymbol{d}_{\text{miRNA}}$ & $\boldsymbol{d}_{\text{DNA}}$
D-GCCA	58.3%	58.3%	0%
JIVE	15.9%	21.0%	17.9%
R.JIVE	0%	0%	0%
AJIVE	75.0%	75.0%	77.8%
COBE	75.0%	75.0%	77.8%
OnPLS	41.3%	60.0%	36.1%
DISCO-SCA	62.5%	68.8%	56.3%
JIVE*	83.3%	75.0%	66.7%
R.JIVE*	0%	0%	0%
AJIVE*	75.0%	75.0%	77.8%
COBE*	75.0%	75.0%	77.8%
OnPLS*	83.3%	50.0%	25.0%
DISCO-SCA*	66.7%	83.3%	55.6%

Table 3: The proportions of significant nonzero correlations between distinctive latent factors across TCGA breast cancer genomic datasets. The proportion is computed by $\frac{1}{d_j d_k}\sum_{m=1}^{r_{d_j}}\sum_{\ell=1}^{r_{d_k}}[\text{corr}(d_j^{(m)}, d_k^{(\ell)}) \neq 0]$ for $j \neq k$, where $\{d_k^{(\ell)}\}_{\ell=1}^{r_{d_k}}$ are latent factors of $\boldsymbol{d}_k$, and $\text{corr}(d_j^{(m)}, d_k^{(\ell)}) \neq 0$ is detected by the normal approximation test (DiCiccio and Romano, 2017) with false discovery rate controlled at 0.05 (Benjamini and Hochberg, 1995) and the ℓ-th right-singular vector of $\widehat{\mathbf{D}}_k$ used as the n samples of $d_k^{(\ell)}$. Methods suffixed with * use D-GCCA's $\widehat{\mathbf{X}}_k$s instead of $\mathbf{Y}_k$s as the input data.

5.2 Application to HCP motor-task functional MRI

We consider the motor-task functional MRI data obtained from the HCP (Barch et al., 2013). During the image scanning, each of 1080 participants was asked by visual cues to either tap left or right fingers, or squeeze left or right toes, or move their tongue. From the acquired brain images, the HCP generated for every participant the z-statistic maps of the individual contrasts of the five tasks and also their average contrast against the fixation baseline. The average contrast represents the impact of the overall motor task. All the maps were computed at 91,282 grayordinates including 59,412 cortical surface vertices and 31,870 subcortical gray matter voxels. For each task, its z-statistic maps of all participants constitute a 91,282×1080 data matrix. We focus on the left-hand, right-hand, and overall motor tasks, and aim to discover the brain regions that are most affected by their common underlying mechanism.

The D-GCCA method is applied to the three data matrices of interest that are row-centered beforehand, with nuisance parameters selected by the approach discussed in Section 3.3. The selection approach yields the same decomposition by the choices 0.2 and 0.0001 for the significance level uniformly applied to all involved tests. All signal and common matrix estimates are rank-2. The proportion of each signal vector's variance explained by its common-variation vector, $\text{PVE}(\boldsymbol{c}_k) = \text{tr}\{\text{cov}(\boldsymbol{c}_k)\}/\,\text{tr}\{\text{cov}(\boldsymbol{x}_k)\}$, is estimated by $\|\widehat{\mathbf{C}}_k\|_F^2/\|\widehat{\mathbf{X}}_k\|_F^2$ with values 0.122, 0.120 and 0.128, respectively, for the left-hand, right-

hand and overall motor tasks. This quantity reflects the overall influence of the common underlying mechanism on the k-th considered motor task.

To assess the local influence of the common underlying mechanism on the i-th brain grayordinate of the k-th task, we use the variance ratio $\mathrm{var}(\boldsymbol{c}_k^{[i]})/\,\mathrm{var}(\boldsymbol{x}_k^{[i]})$ approximated by $\|\widehat{\mathbf{C}}_k^{[i,:]}\|_F^2/\|\widehat{\mathbf{X}}_k^{[i,:]}\|_F^2$. Figure 4 illustrates the estimated variance-ratio maps for the three tasks. In Figure 4 (a) for the left-hand task, we see that the common underlying mechanism has stronger impacts on the right cortical surface, particularly, the somatomotor cortex in the right green circle, whereas it affects more on the left subcortical regions such as the cerebellum shown in the first and last rows of the right part of the figure. The influence pattern is almost opposite for the right-hand task, and is nearly symmetric on the two sides of the brain for the overall motor task. The contralateral change in the somatomotor cortex and the cerebellum is consistent with their intrinsic functional connectivity shown in Buckner et al. (2011).

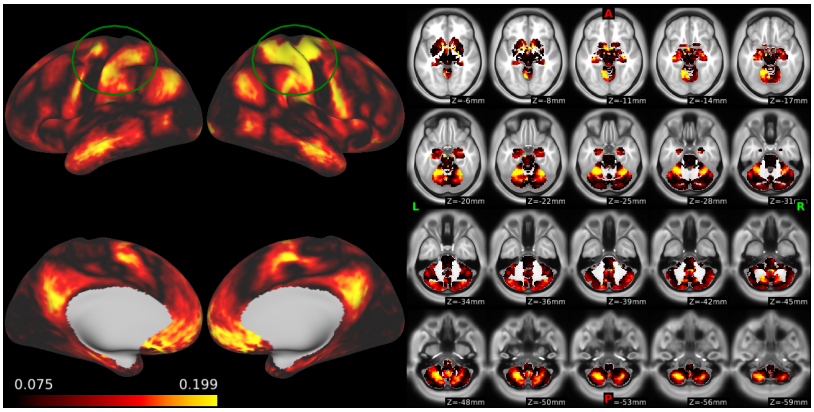
On this large-scale data, we also compare the computational performance of our D-GCCA and the six competing methods mentioned in Section 1. All methods were implemented separately on a computing node with two 10-core Intel Xeon E5-2690v2 3.0GHz CPUs, total 62GB memory, and 24-hour time limit. The three methods, JIVE (with 5.47 hours), R.JIVE (with 17.4 hours) and DISCO-SCA (out of 24 hours), all involving time-expensive iterative optimization, cannot converge within 5 hours. The OnPLS method ran out of memory due to computing the SVD of each large matrix $\mathbf{Y}_j\mathbf{Y}_k^\top$ for $j \neq k$. Both D-GCCA and AJIVE have closed-form expressions, and COBE uses a fast alternating optimization strategy. The computational time costs of the D-GCCA, AJIVE and COBE methods are 18.0, 180.5 and 25.3 seconds, respectively. However, the AJIVE and COBE methods were unable to identify nonzero common-variation matrices.

6. Conclusion

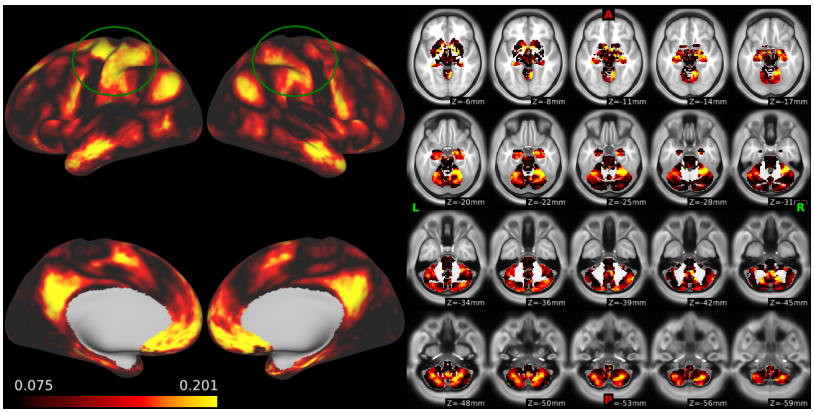
In this paper, we propose a novel decomposition method, called D-GCCA, to separate the common and distinctive variation structures of two or more datasets measured on the same objects. In contrast with existing methods, we build the decomposition on $(\mathcal{L}_0^2, \mathrm{cov})$ rather than the traditional $(\mathbb{R}^n, \cdot)$, and particularly impose a certain orthogonality constraint on the distinctive latent factors to better capture the common variation, along with a geometric interpretation from PCA for the associated common latent factors. Asymptotic result of proposed estimation under high-dimensional settings is established and supported by simulations. Moreover, the D-GCCA decomposition has a closed-form expression and thus is more computationally efficient, especially for large-scale datasets, than most existing methods with time-expensive iterative optimization. Simulated and real-world data show the advantages of D-GCCA over state-of-the-art methods in capturing the common variation and also in the computational time cost.

Acknowledgments

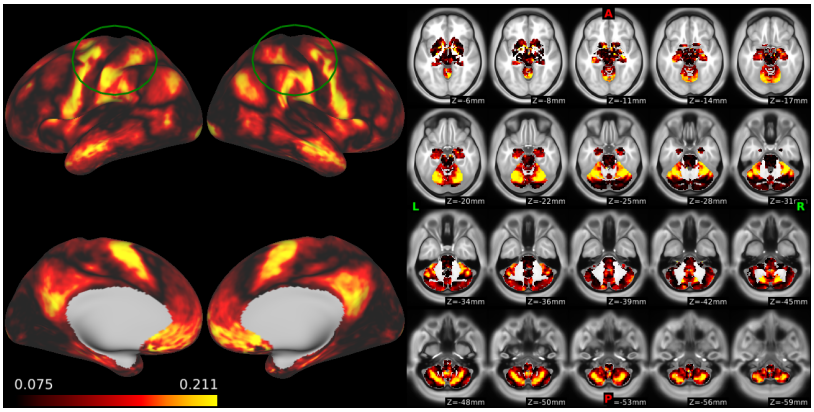
This research was partially supported by U.S. NIH grants MH086633 and MH116527.



(a) Left-hand task



(b) Right-hand task



(c) Overall motor task

Figure 4: Variance-ratio maps estimated by D-GCCA for the three HCP motor tasks. In each subfigure, the left part displays the cortical surface with the outer side shown in the first row and the inner side in the second row; the right part shows the subcortical area on 20 xy slides at the z axis. The somatomotor cortex is annotated by green circles.

Appendix A. Theoretical Proofs

Proof of Theorem 1. It is easily seen that $\sum_{k=1}^{K}\cos^2\theta(z_1,z_k)\geq 1$. If $w\perp\operatorname{span}(\boldsymbol{f}^\top)$, then $\sum_{k=1}^{K}\cos^2\theta(w,z_k)=0$, and thus such a w is not an optimal solution. When $w\not\perp\operatorname{span}(\boldsymbol{f}^\top)$, since $\cos\theta(w,z_k)=\cos\theta(w,w_0)\cos\theta(w_0,z_k)$, where w_0 denotes the projection of w onto $\operatorname{span}(\boldsymbol{f}^\top)$, we only need to consider $w\in\operatorname{span}(\boldsymbol{f}^\top)$. Then there exists a vector $\boldsymbol{b}=(b_1,\ldots,b_K)^\top$ such that $w=\boldsymbol{b}^\top\boldsymbol{f}$ and $\operatorname{cov}(w)=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}=1$. Let z_k^* be the projection of w onto $\operatorname{span}(\boldsymbol{x}_k^\top)$. We only need to consider z_k such that

$$z_k\begin{cases}=\text{ any standardized variable in }\operatorname{span}(\boldsymbol{x}_k^\top), & \text{if } z_k^*=0,\\ \propto z_k^*, & \text{if } z_k^*\neq 0.\end{cases}$$

Define $\boldsymbol{\Phi}_k=(\mathbf{0}_{r_k\times\sum_{j=1}^{k-1}r_j},\mathbf{I}_{r_k\times r_k},\mathbf{0}_{r_k\times\sum_{j=k+1}^{K}r_j})$. Then $\boldsymbol{f}_k=\boldsymbol{\Phi}_k\boldsymbol{f}$ and $\mathbf{I}_{\sum_{k=1}^{K}r_k\times\sum_{k=1}^{K}r_k}=\sum_{k=1}^{K}\boldsymbol{\Phi}_k^\top\boldsymbol{\Phi}_k$. Note that the inner product $\langle w,\boldsymbol{f}_k\rangle=\operatorname{cov}(w,\boldsymbol{f}_k)=\operatorname{cov}(\boldsymbol{b}^\top\boldsymbol{f},\boldsymbol{\Phi}_k\boldsymbol{f})=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{\Phi}_k^\top$, which is zero if $z_k^*=0$. We have

$$\begin{aligned}
&z_k^*=\langle w,\boldsymbol{f}_k\rangle\boldsymbol{f}_k=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{\Phi}_k^\top\boldsymbol{\Phi}_k\boldsymbol{f}, && (17)\\
&\operatorname{var}(z_k^*)=\langle w,\boldsymbol{f}_k\rangle\operatorname{cov}(\boldsymbol{f}_k)\langle w,\boldsymbol{f}_k\rangle^\top=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{\Phi}_k^\top\boldsymbol{\Phi}_k\operatorname{cov}(\boldsymbol{f})\boldsymbol{b},\\
&\operatorname{cov}(w,z_k^*)=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{\Phi}_k^\top\boldsymbol{\Phi}_k\operatorname{cov}(\boldsymbol{f})\boldsymbol{b},\\
&\operatorname{corr}^2(w,z_k^*)=\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{\Phi}_k^\top\boldsymbol{\Phi}_k\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}, && (18)
\end{aligned}$$

and then

$$\sum_{k=1}^{K}\cos^2\theta(w,z_k)=\sum_{k=1}^{K}\operatorname{corr}^2(w,z_k^*)=\boldsymbol{b}^\top\operatorname{cov}^2(\boldsymbol{f})\boldsymbol{b}. \tag{19}$$

Let $w^{(\ell)}=(\boldsymbol{b}^{(\ell)})^\top\boldsymbol{f}$. To maximize (19) with respect to $\boldsymbol{b}$ under the constraints $\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}=1$ and $\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(j)}=0$ for $j\leq\ell-1$, the associated Lagrange function from the method of Lagrange multipliers is

$$\mathcal{L}(\boldsymbol{b},l_1,\ldots,l_\ell)=\boldsymbol{b}^\top\operatorname{cov}^2(\boldsymbol{f})\boldsymbol{b}-l_\ell(\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}-1)-\sum_{j=1}^{\ell-1}l_j\boldsymbol{b}^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(j)}.$$

There exist $l_1^{(\ell)},\ldots,l_\ell^{(\ell)}$ such that $\nabla\mathcal{L}(\boldsymbol{b}^{(\ell)},l_1^{(\ell)},\ldots,l_\ell^{(\ell)})=\mathbf{0}$, which yields

$$\begin{cases}
2\operatorname{cov}^2(\boldsymbol{f})\boldsymbol{b}^{(\ell)}=2l_\ell^{(\ell)}\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(\ell)}+\sum_{j=1}^{\ell-1}l_j^{(\ell)}\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(j)}, & (20a)\\
(\boldsymbol{b}^{(\ell)})^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(\ell)}=1, & (20b)\\
(\boldsymbol{b}^{(\ell)})^\top\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(j)}=0,\quad\text{for}\quad j=1,\ldots,\ell-1. & (20c)
\end{cases}$$

When $\ell=1$, (20a) becomes $\operatorname{cov}^2(\boldsymbol{f})\boldsymbol{b}^{(1)}=l_1^{(1)}\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(1)}$. Then by (20b), we have $l_1^{(1)}=(\boldsymbol{b}^{(\ell)})^\top\operatorname{cov}^2(\boldsymbol{f})\boldsymbol{b}^{(\ell)}$. Thus, the maximum of (19) when $\ell=1$, i.e., the maximum of $l_1^{(1)}$ is $l_{f,1}\coloneqq\lambda_1(\operatorname{cov}(\boldsymbol{f}))$. We have $l_{f,1}^{-1/2}\operatorname{cov}(\boldsymbol{f})\boldsymbol{b}^{(1)}=\boldsymbol{\eta}^{(1)}$. Hence, $\boldsymbol{b}^{(1)}=l_{f,1}^{1/2}[\operatorname{cov}(\boldsymbol{f})]^\dagger\boldsymbol{\eta}^{(1)}+\boldsymbol{\zeta}$ for any vector $\boldsymbol{\zeta}$ satisfying $\mathbf{V}_f^\top\boldsymbol{\zeta}=\mathbf{0}$, where $\operatorname{cov}(\boldsymbol{f})=\mathbf{V}_f\boldsymbol{\Lambda}_f\mathbf{V}_f^\top$ is a compact SVD of $\operatorname{cov}(\boldsymbol{f})$,

and $[\mathrm{cov}(\boldsymbol{f})]^{\dagger} = \mathbf{V}_f \boldsymbol{\Lambda}_f^{-1} \mathbf{V}_f^{\top}$ is the pseudo-inverse of $\mathrm{cov}(\boldsymbol{f})$. Let $\boldsymbol{u} = \boldsymbol{\Lambda}_f^{-1/2} \mathbf{V}_f^{\top} \boldsymbol{f}$. Then $\boldsymbol{f} = \mathrm{cov}(\boldsymbol{f}, \boldsymbol{u}) \boldsymbol{u} = \mathbf{V}_f \boldsymbol{\Lambda}_f^{1/2} \boldsymbol{u}$. We have

$$\begin{aligned} w^{(1)} = (\boldsymbol{b}^{(1)})^{\top} \boldsymbol{f} &= (l_{f,1}^{1/2} (\boldsymbol{\eta}^{(1)})^{\top} [\mathrm{cov}(\boldsymbol{f})]^{\dagger} + \boldsymbol{\zeta}^{\top}) \mathbf{V}_f \boldsymbol{\Lambda}_f^{1/2} \boldsymbol{u} \\ &= l_{f,1}^{1/2} (\boldsymbol{\eta}^{(1)})^{\top} [\mathrm{cov}(\boldsymbol{f})]^{\dagger} \mathbf{V}_f \boldsymbol{\Lambda}_f^{1/2} \boldsymbol{u} \\ &= l_{f,1}^{-1/2} (\boldsymbol{\eta}^{(1)})^{\top} \mathrm{cov}(\boldsymbol{f}) [\mathrm{cov}(\boldsymbol{f})]^{\dagger} \mathbf{V}_f \boldsymbol{\Lambda}_f^{1/2} \boldsymbol{u} \\ &= l_{f,1}^{-1/2} (\boldsymbol{\eta}^{(1)})^{\top} \mathbf{V}_f \boldsymbol{\Lambda}_f^{1/2} \boldsymbol{u} \\ &= l_{f,1}^{-1/2} (\boldsymbol{\eta}^{(1)})^{\top} \boldsymbol{f}. \end{aligned}$$

Hence, we can simply let $\boldsymbol{b}^{(1)} = l_{f,1}^{-1/2} \boldsymbol{\eta}^{(1)}$. When $\ell = 2$, left-multiplying (20a) by $\boldsymbol{b}^{(1)}$ yields $l_1^{(2)} = 0$. Then (20) becomes

$$\begin{cases} \mathrm{cov}^2(\boldsymbol{f}) \boldsymbol{b}^{(2)} = l_2^{(2)} \mathrm{cov}(\boldsymbol{f}) \boldsymbol{b}^{(2)}, \\ (\boldsymbol{b}^{(2)})^{\top} \mathrm{cov}(\boldsymbol{f}) \boldsymbol{b}^{(2)} = 1, \\ (\boldsymbol{b}^{(2)})^{\top} \mathrm{cov}(\boldsymbol{f}) \boldsymbol{b}^{(1)} = 0. \end{cases}$$

Thus, we have $[\lambda_2(\mathrm{cov}(\boldsymbol{f}))]^{-1/2} \mathrm{cov}(\boldsymbol{f}) \boldsymbol{b}^{(2)} = \boldsymbol{\eta}^{(2)}$. Then using the same skill for obtaining $\boldsymbol{b}^{(1)}$, we can simply let $\boldsymbol{b}^{(2)} = [\lambda_2(\mathrm{cov}(\boldsymbol{f}))]^{-1/2} \boldsymbol{\eta}^{(2)}$ and have $\sum_{k=1}^{K} \cos^2 \theta(w^{(2)}, z_k^{(2)}) = \lambda_2(\mathrm{cov}(\boldsymbol{f}))$. Similarly, for $2 < \ell \leq r_f$, we can simply let $\boldsymbol{b}^{(\ell)} = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{-1/2} \boldsymbol{\eta}^{(\ell)}$ and have $\sum_{k=1}^{K} \cos^2 \theta(w^{(\ell)}, z_k^{(\ell)}) = \lambda_\ell(\mathrm{cov}(\boldsymbol{f}))$.

For $\ell \leq r_f$, by (17), the projection of $w^{(\ell)}$ onto space $\mathrm{span}(\boldsymbol{x}_k^{\top})$ is $z_k^{*(\ell)} = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{1/2} (\boldsymbol{\eta}_k^{(\ell)})^{\top} \boldsymbol{f}_k$ with $\mathrm{var}(z_k^{*(\ell)}) = \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) \|\boldsymbol{\eta}_k^{(\ell)}\|_F^2$. Thus,

$$z_k^{(\ell)} = \begin{cases} \text{any standardized variable in } \mathrm{span}(\boldsymbol{x}^{\top}), & \text{if } \boldsymbol{\eta}_k^{(\ell)} = \mathbf{0}, \\ \pm (\boldsymbol{\eta}_k^{(\ell)} / \|\boldsymbol{\eta}_k^{(\ell)}\|_F)^{\top} \boldsymbol{f}_k, & \text{if } \boldsymbol{\eta}_k^{(\ell)} \neq \mathbf{0}. \end{cases}$$

From equation (18), we have $\mathrm{cov}(w^{(\ell)}, z_k^{*(\ell)}) = \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) \|\boldsymbol{\eta}_k^{(\ell)}\|_F^2$. Then, $\cos \theta(w^{(\ell)}, z_k^{(\ell)}) = \pm [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{1/2} \|\boldsymbol{\eta}_k^{(\ell)}\|_F$.

To prove $\sum_{k=1}^{K} \mathrm{span}(\boldsymbol{x}_k^{\top}) = \mathrm{span}(\{w^{(\ell)}\}_{\ell=1}^{r_f})$, since $w^{(\ell)} \in \mathrm{span}(\boldsymbol{f}^{\top})$, we only need to show $\dim(\mathrm{span}(\{w^{(\ell)}\}_{\ell=1}^{r_f})) = \dim(\mathrm{span}(\boldsymbol{f}^{\top})) = r_f$, which is true because the r_f nonzero variables $\{w^{(\ell)}\}_{\ell=1}^{r_f}$ are orthogonal.

Now consider the revised $z_k^{(\ell)}$ in (8) for result (ii). By $\cos \theta(w^{(\ell)}, z_k^{(\ell)}) = [\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))]^{1/2} \|\boldsymbol{\eta}_k^{(\ell)}\|_F \geq 0$, we have $\theta(w^{(\ell)}, z_k^{(\ell)}) \in [0, \pi/2]$. Since $\mathrm{span}(\{z_k^{(\ell)}\}_{\ell=1}^{r_f})$ is the projection of $\mathrm{span}(\{w^{(\ell)}\}_{\ell=1}^{r_f})$ onto $\mathrm{span}(\boldsymbol{x}_k^{\top}) \subseteq \mathrm{span}(\{w^{(\ell)}\}_{\ell=1}^{r_f}) = \sum_{k=1}^{K} \mathrm{span}(\boldsymbol{x}_k^{\top})$, we have $\mathrm{span}(\{z_k^{(\ell)}\}_{\ell=1}^{r_f}) = \mathrm{span}(\boldsymbol{x}_k^{\top})$.

Next, consider result (iii). For some k and ℓ, since $\mathrm{span}(\{z_k^{(m)}\}_{m=1}^{\ell-1}) \neq \mathrm{span}(\boldsymbol{x}_k^{\top})$, there exists a unit-variance variable $v \in \mathrm{span}(\boldsymbol{x}_k^{\top})$ such that $v \perp \mathrm{span}(\{z_k^{(m)}\}_{m=1}^{\ell-1})$. Moreover, $v \perp w^{(m)}$ for all $m \leq \ell - 1$, because v is orthogonal to both the projection of $w^{(m)}$ onto $\mathrm{span}(\boldsymbol{x}_k^{\top})$ and the rejection of $w^{(m)}$ from $\mathrm{span}(\boldsymbol{x}_k^{\top})$. Thus, we just let $w^{(\ell)} = v$. Then, $\cos^2 \theta(w^{(\ell)}, z_k^{(\ell)}) = 1$. By $\sum_{k=1}^{K} \cos^2 \theta(w^{(\ell)}, z_k^{(\ell)}) = \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) \leq 1$, we have

$\sum_{j\neq k}\cos^2\theta(w^{(\ell)},z_j^{(\ell)})=0$, which implies $w^{(\ell)}\perp\sum_{1\leq j\neq k\leq K}\operatorname{span}(\boldsymbol{x}_j^\top)$. ■

Proof of Theorem 2. If $z_k^{(\ell)}=0$ for some k, it is easy to see $\alpha^{(\ell)}=0$. We only consider that for all $k\leq K$, $z_k^{(\ell)}\neq 0$, i.e., $\theta(w^{(\ell)},z_k^{(\ell)})\in[0,\pi/2)$. If $d_j^{(\ell)}\perp d_k^{(\ell)}$, then $\|d_j^{(\ell)}\|^2+\|d_k^{(\ell)}\|^2=\|z_j^{(\ell)}-z_k^{(\ell)}\|^2$, and consequently by the law of cosines we have

$$\begin{aligned}&\Big(\|z_j^{(\ell)}\|^2+\|c^{(\ell)}\|^2-2\|z_j^{(\ell)}\|\|c^{(\ell)}\|\cos\theta(z_j^{(\ell)},w^{(\ell)})\Big)\\&+\Big(\|z_k^{(\ell)}\|^2+\|c^{(\ell)}\|^2-2\|z_k^{(\ell)}\|\|c^{(\ell)}\|\cos\theta(z_k^{(\ell)},w^{(\ell)})\Big)\\&\quad=\|z_j^{(\ell)}\|^2+\|z_k^{(\ell)}\|^2-2\|z_j^{(\ell)}\|\|z_k^{(\ell)}\|\cos\theta(z_j^{(\ell)},z_k^{(\ell)})\end{aligned}$$

which gives

$$\alpha^{(\ell)}=\frac{1}{2}\left[\cos\theta(z_j^{(\ell)},w^{(\ell)})+\cos\theta(z_k^{(\ell)},w^{(\ell)})\pm(\Delta_{jk}^{(\ell)})^{1/2}\right].$$

Hence, the desired value of $\alpha^{(\ell)}$ is the one given in Theorem 2.

To prove the existence of $\alpha^{(\ell)}$, we only need to show that there exists a $\Delta_{jk}^{(\ell)}\geq 0$ with $j\neq k$. Denote $\lambda_\ell=\lambda_\ell(\operatorname{cov}(\boldsymbol{f}))$, and $\boldsymbol{\nu}_\ell=(\nu_{\ell,1},\ldots,\nu_{\ell,K})^\top$ with $\nu_{\ell,k}=\|\boldsymbol{\eta}_k^{(\ell)}\|_F$. We have

$$\operatorname{cov}(\boldsymbol{z}^{(\ell)})=\operatorname{diag}\big(\frac{(\boldsymbol{\eta}_1^{(\ell)})^\top}{\|\boldsymbol{\eta}_1^{(\ell)}\|_F},\ldots,\frac{(\boldsymbol{\eta}_K^{(\ell)})^\top}{\|\boldsymbol{\eta}_K^{(\ell)}\|_F}\big)\operatorname{cov}(\boldsymbol{f})\operatorname{diag}\big(\frac{\boldsymbol{\eta}_1^{(\ell)}}{\|\boldsymbol{\eta}_1^{(\ell)}\|_F},\ldots,\frac{\boldsymbol{\eta}_K^{(\ell)}}{\|\boldsymbol{\eta}_K^{(\ell)}\|_F}\big),$$

$$\operatorname{cov}(\boldsymbol{z}^{(\ell)})\boldsymbol{\nu}_\ell=\lambda_\ell\boldsymbol{\nu}_\ell,$$

$$\cos\theta(w^{(\ell)},z_k^{(\ell)})=\lambda_\ell^{1/2}\nu_{\ell,k},$$

and for all $j,k\leq K$,

$$\Delta_{jk}^{(\ell)}=\lambda_\ell\nu_{\ell,j}^2+\lambda_\ell\nu_{\ell,k}^2+2\lambda_\ell\nu_{\ell,j}\nu_{\ell,k}-4\operatorname{cov}(z_j^{(\ell)},z_k^{(\ell)}).$$

Then ,

$$
\begin{aligned}
&\sum_{j=1}^{K}\sum_{k=1}^{K}\cos\theta(w^{(\ell)},z_j^{(\ell)})\Delta_{jk}^{(\ell)}\cos\theta(w^{(\ell)},z_k^{(\ell)})\\
&=\sum_{j=1}^{K}\sum_{k=1}^{K}\operatorname{cov}(w^{(\ell)},z_j^{(\ell)})\Delta_{jk}^{(\ell)}\operatorname{cov}(w^{(\ell)},z_k^{(\ell)})\\
&=\sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_\ell^{1/2}\nu_{\ell,j}\left(\lambda_\ell\nu_{\ell,j}^2+\lambda_\ell\nu_{\ell,k}^2+2\lambda_\ell\nu_{\ell,j}\nu_{\ell,k}-4\operatorname{cov}(z_j^{(\ell)},z_k^{(\ell)})\right)\lambda_\ell^{1/2}\nu_{\ell,k}\\
&=\left[\sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_\ell^{1/2}\nu_{\ell,j}(\lambda_\ell\nu_{\ell,j}^2+\lambda_\ell\nu_{\ell,k}^2+2\lambda_\ell\nu_{\ell,j}\nu_{\ell,k})\lambda_\ell^{1/2}\nu_{\ell,k}\right]-4\lambda_\ell\boldsymbol{\nu}_\ell^\top\operatorname{cov}(\boldsymbol{z}^{(\ell)})\boldsymbol{\nu}_\ell\\
&=\left[\sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_\ell^{1/2}\nu_{\ell,j}(\lambda_\ell\nu_{\ell,j}^2+\lambda_\ell\nu_{\ell,k}^2+2\lambda_\ell\nu_{\ell,j}\nu_{\ell,k})\lambda_\ell^{1/2}\nu_{\ell,k}\right]-4\lambda_\ell^2\boldsymbol{\nu}_\ell^\top(\boldsymbol{\nu}_\ell\boldsymbol{\nu}_\ell^\top)\boldsymbol{\nu}_\ell\\
&=\sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_\ell^{1/2}\nu_{\ell,j}\left(\lambda_\ell\nu_{\ell,j}^2+\lambda_\ell\nu_{\ell,k}^2+2\lambda_\ell\nu_{\ell,j}\nu_{\ell,k}-4\lambda_\ell\nu_{\ell,j}\nu_{\ell,k}\right)\lambda_\ell^{1/2}\nu_{\ell,k}\\
&=\sum_{j=1}^{K}\sum_{k=1}^{K}\cos\theta(w^{(\ell)},z_j^{(\ell)})(\lambda_\ell^{1/2}\nu_{\ell,j}-\lambda_\ell^{1/2}\nu_{\ell,k})^2\cos\theta(w^{(\ell)},z_k^{(\ell)})\\
&\geq 0. \qquad (21)
\end{aligned}
$$

For all $k<K$, $\cos\theta(w^{(\ell)},z_k^{(\ell)})>0$ for $\theta(w^{(\ell)},z_k^{(\ell)})\in[0,\pi/2)$, and moreover, we have $\Delta_{kk}^{(\ell)}=4\cos^2\theta(w^{(\ell)},z_k^{(\ell)})-4\leq 0$. Hence, by (21), we have at least one $\Delta_{jk}^{(\ell)}\geq 0$ with $j\neq k$. ■

Proof of Theorem 3. When $K=2$, by Lemma 2 in Kettenring (1971), L is equal to the number of positive canonical correlations between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Then following the constructions of these two decomposition methods, the proof is easy to complete. The details are omitted. ■

Proof of Theorem 4. Let $\widetilde{\boldsymbol{f}}_k^\top$ be another orthonormal basis of $\operatorname{span}(\boldsymbol{x}_k^\top)$. Then, there exists an orthogonal matrix $\mathbf{O}_k$ such that $\widetilde{\boldsymbol{f}}_k=\mathbf{O}_k\boldsymbol{f}_k$. Define $\widetilde{\boldsymbol{f}}=(\widetilde{\boldsymbol{f}}_1^\top,\ldots,\widetilde{\boldsymbol{f}}_K^\top)^\top$. We have $\widetilde{\boldsymbol{f}}=\mathbf{O}\boldsymbol{f}$ and $\operatorname{cov}(\widetilde{\boldsymbol{f}})=\mathbf{O}\operatorname{cov}(\boldsymbol{f})\mathbf{O}^\top$ with $\mathbf{O}=\operatorname{diag}(\mathbf{O}_1,\ldots,\mathbf{O}_K)$. Hence, $\lambda_\ell(\operatorname{cov}(\widetilde{\boldsymbol{f}}))=\lambda_\ell(\operatorname{cov}(\boldsymbol{f}))$ for $\ell\leq\sum_{k=1}^K r_k$. Denote $\widetilde{\boldsymbol{\eta}}^{(\ell)}=((\widetilde{\boldsymbol{\eta}}_1^{(\ell)})^\top,\ldots,(\widetilde{\boldsymbol{\eta}}_K^{(\ell)})^\top)^\top$, with $(\widetilde{\boldsymbol{\eta}}_k^{(\ell)})^\top\in\mathbb{R}^{r_k}$, to be a normalized eigenvector of $\operatorname{cov}(\widetilde{\boldsymbol{f}})$ corresponding to $\lambda_\ell(\operatorname{cov}(\widetilde{\boldsymbol{f}}))$ for $\ell\leq L$. Now, from the assumption that $\lambda_1(\operatorname{cov}(\boldsymbol{f})),\ldots,\lambda_L(\operatorname{cov}(\boldsymbol{f}))$ are distinct, we have $\widetilde{\boldsymbol{\eta}}^{(\ell)}=\pm\mathbf{O}\boldsymbol{\eta}^{(\ell)}$ and $\widetilde{\boldsymbol{\eta}}_k^{(\ell)}=\pm\mathbf{O}_k\boldsymbol{\eta}_k^{(\ell)}$. Denote $\widetilde{w}^{(\ell)}$, $\widetilde{z}_k^{(\ell)}$, $\widetilde{\alpha}^{(\ell)}$ and $\widetilde{c}^{(\ell)}$ to be the counterparts of $w^{(\ell)}$, $z_k^{(\ell)}$, $\alpha^{(\ell)}$ and $c^{(\ell)}$ that are defined in (5), (8) and (11) by using $\widetilde{\boldsymbol{f}}$ and $\widetilde{\boldsymbol{\eta}}^{(\ell)}$ instead of $\boldsymbol{f}$ and $\boldsymbol{\eta}^{(\ell)}$. We have $\widetilde{w}^{(\ell)}=\pm w^{(\ell)}$, $\widetilde{z}_k^{(\ell)}=\pm z_k^{(\ell)}$, $\widetilde{\alpha}^{(\ell)}=\alpha^{(\ell)}$ due to the formula in Theorem 2, and then

$\widetilde{c}^{(\ell)} = \pm c^{(\ell)}$. Let $\widetilde{\boldsymbol{z}}_k^{\mathcal{I}_0} = (\widetilde{z}_k^{(\ell)})_{\ell\in\mathcal{I}_0}^\top$ and $\widetilde{\boldsymbol{c}}^{\mathcal{I}_0} = (\widetilde{c}^{(\ell)})_{\ell\in\mathcal{I}_0}^\top$. There exists a diagonal matrix $\mathbf{D}$ with diagonal entries being either 1 or -1 such that $\widetilde{\boldsymbol{z}}_k^{\mathcal{I}_0} = \mathbf{D}\boldsymbol{z}_k^{\mathcal{I}_0}$ and $\widetilde{\boldsymbol{c}}^{\mathcal{I}_0} = \mathbf{D}\boldsymbol{c}^{\mathcal{I}_0}$. Then,

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{x}_k, \widetilde{\boldsymbol{z}}_k^{\mathcal{I}_0})[\operatorname{cov}(\widetilde{\boldsymbol{z}}_k^{\mathcal{I}_0})]^\dagger \widetilde{\boldsymbol{c}}^{\mathcal{I}_0} &= \operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\mathbf{D}[\mathbf{D}\operatorname{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\mathbf{D}]^\dagger \mathbf{D}\boldsymbol{c}^{\mathcal{I}_0} \\
&= \operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\mathbf{D}[\mathbf{D}\mathbf{V}_{zk}\boldsymbol{\Lambda}_{zk}\mathbf{V}_{zk}^\top\mathbf{D}]^\dagger \mathbf{D}\boldsymbol{c}^{\mathcal{I}_0} \\
&= \operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\mathbf{D}[\mathbf{D}\mathbf{V}_{zk}\boldsymbol{\Lambda}_{zk}^{-1}\mathbf{V}_{zk}^\top\mathbf{D}]\mathbf{D}\boldsymbol{c}^{\mathcal{I}_0} \\
&= \operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})[\operatorname{cov}(\boldsymbol{z}_k)]^\dagger \boldsymbol{c}^{\mathcal{I}_0} = \boldsymbol{c}_k.
\end{aligned}$$

The proof is complete. ■

Proof of Theorem 5. First of all, it is worth mentioning that $\widehat{\mathbf{X}}_k$ is rank-r_k with probability tending to 1. This is because we have

$$\lambda_{r_k}(\widehat{\operatorname{cov}}(\boldsymbol{x}_k)) \geq (1 - o_P(1))\, \lambda_{r_k}(\operatorname{cov}(\boldsymbol{x}_k))$$

from (S.17) in the supplement of Shu et al. (2019). Due to their Lemma S.1, in the rest of the proof we simply assume that $\widehat{\mathbf{X}}_k$ is rank-r_k. By their (S.13) and (S.14), there exists a constant $\kappa_x > 0$ such that

$$\kappa_x + o_P(1) \leq \frac{\|\mathbf{X}_k\|_2}{[n\,\lambda_1(\operatorname{cov}(\boldsymbol{x}_k))]^{1/2}} \leq \frac{\|\mathbf{X}_k\|_F}{[n\,\lambda_1(\operatorname{cov}(\boldsymbol{x}_k))]^{1/2}} \leq r_k^{1/2} + o_P(1). \tag{22}$$

From their (S.15), we have

$$\begin{aligned}
&\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_2 \leq \|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_F \\
&\lesssim_P \min\left\{ \Big\lceil \frac{\lambda_1(\operatorname{cov}(\boldsymbol{x}_k))}{n} \Big\rceil^{1/2} + (p_k \log p_k)^{1/2}, [n\,\lambda_1(\operatorname{cov}(\boldsymbol{x}_k))]^{1/2} \right\}.
\end{aligned} \tag{23}$$

Here and in the following text, we write $A \lesssim_P B$ if and only if $A = O_P(B)$. From (S.7) of Shu et al. (2019), we have $\lambda_1(\operatorname{cov}(\boldsymbol{x}_k)) \asymp \lambda_{r_k}(\operatorname{cov}(\boldsymbol{x}_k))$. By Weyl's inequality (see Theorem 3.3.16(a) in Horn and Johnson (1994)) as well as Assumption 1 (i) and (v), $\kappa_1 \leq \lambda_{k,p_k} = \lambda_{k,(r_k+1)+(p_k-r_k)-1} - \lambda_{r_k+1}(\operatorname{cov}(\boldsymbol{x}_k)) \leq \lambda_{p_k-r_k}(\operatorname{cov}(\boldsymbol{e}_k)) \leq \lambda_1(\operatorname{cov}(\boldsymbol{e}_k)) = \|\operatorname{cov}(\boldsymbol{e}_k)\|_2 \leq \|\operatorname{cov}(\boldsymbol{e}_k)\|_\infty \leq s_0$. Thus,

$$\frac{\lambda_1(\operatorname{cov}(\boldsymbol{x}_k))}{p_k} \asymp \frac{\operatorname{tr}(\operatorname{cov}(\boldsymbol{x}_k))}{\operatorname{tr}(\operatorname{cov}(\boldsymbol{e}_k))} = \mathrm{SNR}_k\,. \tag{24}$$

By (22), (23) and (24), we obtain

$$\max\left\{ \frac{\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_2^2}{\|\mathbf{X}_k\|_2^2}, \frac{\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_F^2}{\|\mathbf{X}_k\|_F^2} \right\} \lesssim_P \min\left\{ \frac{1}{n^2} + \frac{\log p_k}{n\,\mathrm{SNR}_k}, 1 \right\}. \tag{25}$$

Simply choose $\boldsymbol{f}_k = \boldsymbol{\Lambda}_{xk}^{-1/2}\mathbf{V}_{xk}^\top\boldsymbol{x}_k$, where $\operatorname{cov}(\boldsymbol{x}_k) = \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}\mathbf{V}_{xk}^\top$ is a compact SVD. Then, we have $\boldsymbol{z}_k^{\mathcal{I}_0} = \mathbf{H}_k\boldsymbol{f}_k = \mathbf{H}_k\boldsymbol{\Lambda}_{xk}^{-1/2}\mathbf{V}_{xk}^\top\boldsymbol{x}_k$ with $\mathbf{H}_k = (\boldsymbol{\eta}_k^{(\ell)}/\|\boldsymbol{\eta}_k^{(\ell)}\|_F)_{\ell\in\mathcal{I}_0}^\top$. From (13), it follows that we can write the common-variation matrix $\mathbf{C}_k$ as

$$\mathbf{C}_k = \operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0})\{\operatorname{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\}^\dagger \mathbf{C}^{\mathcal{I}_0}, \tag{26}$$

where the three components are formulated by $\mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{\mathcal{I}_0}) = \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top$, $\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0}) = \mathbf{H}_k\mathbf{H}_k^\top$, and $\mathbf{C}^{\mathcal{I}_0} = \mathbf{ANF}$ with $\mathbf{A} = \mathrm{diag}\{(\alpha^{(\ell)}[\lambda_\ell\{\mathrm{cov}(\boldsymbol{f})\}]^{-1/2})_{\ell\in\mathcal{I}_0}\}$, $\mathbf{N} = (\boldsymbol{\eta}^{(\ell)})_{\ell\in\mathcal{I}_0}^\top$, and $\mathbf{F} = (\mathbf{F}_1^\top, \ldots, \mathbf{F}_K^\top)^\top$ in which $\mathbf{F}_k = \boldsymbol{\Lambda}_{xk}^{-1/2}\mathbf{V}_{xk}^\top\mathbf{X}_k$.

Since K is a constant and each $\mathrm{span}(\boldsymbol{x}_k^\top)$ is a fixed space independent of n and $\{p_k\}_{k=1}^K$, we have that $r_1, \ldots, r_K$ are constants and there exist positive constants κ_z, κ_η and κ_{zz} such that $\min_{k\le K}\lambda_{r_k^*}(\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})) > \kappa_z$, $\min_{k\le K,\ell\in\mathcal{I}_0}\|\boldsymbol{\eta}_k^{(\ell)}\|_F > \kappa_\eta$, and $\min_{(j,k)\in\mathcal{I}_\Delta^{(\ell)},\ell\in\mathcal{I}_0}|\cos\theta(z_j^{(\ell)}, z_k^{(\ell)})| > \kappa_{zz}$.

From Shu et al. (2019), using their (S.8), (S.30) and the first inequality on page 10 of their supplement, we have that for all $j, k \le K$,

$$\lambda_1(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) \lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k)), \tag{27}$$

$$\|\widehat{\mathbf{V}}_{xk}\widehat{\boldsymbol{\Lambda}}_{xk}^{1/2} - \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\|_2 \lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}, \tag{28}$$

and

$$\begin{aligned}\|\widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\|_F &\le [\max(r_j, r_k)]^{1/2}\|\widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\|_2 \\ &\lesssim_P \min\left\{n^{-1/2} + \left(\frac{p_j\log p_j}{n\,\lambda_1(\mathrm{cov}(\boldsymbol{x}_j))}\right)^{1/2} + \left(\frac{p_k\log p_k}{n\,\lambda_1(\mathrm{cov}(\boldsymbol{x}_k))}\right)^{1/2}, 1\right\},\end{aligned}$$

where $\widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) = n^{-1}\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_k^\top$ is a submatrix of $\widehat{\mathrm{cov}}(\boldsymbol{f})$. Then,

$$\begin{aligned}\|\widehat{\mathrm{cov}}(\boldsymbol{f}) - \mathrm{cov}(\boldsymbol{f})\|_F &= \Big(\sum_{1\le j,k\le K}\|\widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\|_F^2\Big)^{1/2} \\ &\lesssim_P \min\left\{n^{-1/2} + \sum_{k=1}^K\Big(\frac{p_k\log p_k}{n\,\lambda_1(\mathrm{cov}(\boldsymbol{x}_k))}\Big)^{1/2}, 1\right\} \\ &=: \delta_f.\end{aligned} \tag{29}$$

By the uniqueness given in Theorem 4, we let $\boldsymbol{\eta}^{(\ell)}$ satisfy $(\boldsymbol{\eta}^{(\ell)})^\top\widehat{\boldsymbol{\eta}}^{(\ell)} \ge 0$ for all $\ell \in \mathcal{I}_0$. By Corollary 1 in Yu et al. (2015), $\delta_\eta = o(1)$, and the condition that $\{\lambda_\ell(\mathrm{cov}(f))\}_{\ell=1}^L$ are distinct, we have

$$\begin{aligned}&\max_{\ell\in\mathcal{I}_0}\|\widehat{\boldsymbol{\eta}}^{(\ell)} - \boldsymbol{\eta}^{(\ell)}\|_F \\ &\lesssim_P \frac{\delta_f}{\min_{\ell\in\mathcal{I}_0}\{\lambda_{\ell-1}(\mathrm{cov}(\boldsymbol{f})) - \lambda_\ell(\mathrm{cov}(\boldsymbol{f})), \lambda_\ell(\mathrm{cov}(\boldsymbol{f})) - \lambda_{\ell+1}(\mathrm{cov}(\boldsymbol{f}))\}} \\ &\lesssim_P \delta_\eta.\end{aligned} \tag{30}$$

Since $\delta_\eta = o(1)$ and $\min_{k\le K,\ell\in\mathcal{I}_0}\|\boldsymbol{\eta}_k^{(\ell)}\|_F > \kappa_\eta$, then by (30) we have

$$\min_{k\le K,\ell\in\mathcal{I}_0}\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F \ge \kappa_\eta - o_P(1), \tag{31}$$

and thus

$$\begin{aligned}\|\widehat{\mathbf{H}}_k-\mathbf{H}_k\|_2 &\le \|\widehat{\mathbf{H}}_k-\mathbf{H}_k\|_F \lesssim_P L^{1/2}\max_{\ell\in\mathcal{I}_0}\left\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}/\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F-\boldsymbol{\eta}_k^{(\ell)}/\|\boldsymbol{\eta}_k^{(\ell)}\|_F\right\|_F\\ &\lesssim_P L^{1/2}\max_{\ell\in\mathcal{I}_0}\left\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}(\|\boldsymbol{\eta}_k^{(\ell)}\|_F-\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F)+(\widehat{\boldsymbol{\eta}}_k^{(\ell)}-\boldsymbol{\eta}_k^{(\ell)})\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F\right\|_F/(\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F\|\boldsymbol{\eta}_k^{(\ell)}\|_F)\\ &\lesssim_P 2L^{1/2}\max_{\ell\in\mathcal{I}_0}\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}-\boldsymbol{\eta}_k^{(\ell)}\|_F/\|\boldsymbol{\eta}_k^{(\ell)}\|_F\\ &\lesssim_P \delta_\eta. \end{aligned}\tag{32}$$

We will frequently use the following matrix inequality:

$$\|\widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_2-\mathbf{M}_1\mathbf{M}_2\|_2\le\begin{cases}\|\widehat{\mathbf{M}}_1\|_2\|\widehat{\mathbf{M}}_2-\mathbf{M}_2\|_2+\|\mathbf{M}_2\|_2\|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_2,\\ \|\widehat{\mathbf{M}}_2\|_2\|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_2+\|\mathbf{M}_1\|_2\|\widehat{\mathbf{M}}_2-\mathbf{M}_2\|_2.\end{cases}\tag{33}$$

Then together with $\max_{k\le K}\{\|\mathbf{H}_k\|_F,\|\widehat{\mathbf{H}}_k\|_F\}\le L^{1/2}$, we have

$$\|\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\|_2=\|\widehat{\mathbf{H}}_k\widehat{\mathbf{H}}_k^\top-\mathbf{H}_k\mathbf{H}_k^\top\|_2\le(\|\widehat{\mathbf{H}}_k\|_2+\|\mathbf{H}_k\|_2)\|\widehat{\mathbf{H}}_k-\mathbf{H}_k\|_2\lesssim_P\delta_\eta.\tag{34}$$

Recall that $\min_{k\le K}\lambda_{r_k^*}(\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0}))>\kappa_z$. Let $\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})=\mathbf{V}_{zk}\boldsymbol{\Lambda}_{zk}\mathbf{V}_{zk}^\top$ be its compact SVD, where $\boldsymbol{\Lambda}_{zk}$ has nonincreasing diagonal elements. Let $\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}=0$ for $j>\widetilde{r}_k:=\mathrm{rank}(\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0}))$, and $\boldsymbol{\Lambda}_{zk}^{[j,j]}=0$ for $j>r_k^*$. By Weyl's inequality (see Theorem 3.3.16(c) in Horn and Johnson (1994)), for all j,

$$|\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}-\boldsymbol{\Lambda}_{zk}^{[j,j]}|\le\|\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\|_2\lesssim_P\delta_\eta.$$

Hence,

$$\widehat{\boldsymbol{\Lambda}}_{zk}^{[\check{r}_k,\check{r}_k]}\ge\boldsymbol{\Lambda}_{zk}^{[\check{r}_k,\check{r}_k]}-O_P(\delta_\eta)\ge\kappa_z-o_P(1)\tag{35}$$

and

$$\max_{j>r_k^*}\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}\lesssim_P\delta_\eta.$$

Then,

$$\begin{aligned}&\|\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\|_2\\ &=\left\|\sum_{j=1}^{\check{r}_k}\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}\widehat{\mathbf{V}}_{zk}^{[,j]}(\widehat{\mathbf{V}}_{zk}^{[,j]})^\top-\sum_{j=1}^{r_k^*}\boldsymbol{\Lambda}_{zk}^{[j,j]}\mathbf{V}_{zk}^{[,j]}(\mathbf{V}_{zk}^{[,j]})^\top\right\|_2\\ &\le\left\|\sum_{j=1}^{\widetilde{r}_k}\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}\widehat{\mathbf{V}}_{zk}^{[,j]}(\widehat{\mathbf{V}}_{zk}^{[,j]})^\top-\sum_{j=1}^{r_k^*}\boldsymbol{\Lambda}_{zk}^{[j,j]}\mathbf{V}_{zk}^{[,j]}(\mathbf{V}_{zk}^{[,j]})^\top\right\|_2+\sum_{j=\check{r}_k+1}^{\widetilde{r}_k}\|\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}\widehat{\mathbf{V}}_{zk}^{[,j]}(\widehat{\mathbf{V}}_{zk}^{[,j]})^\top\|_2\\ &=\|\widetilde{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\|_2+\sum_{j=\check{r}_k+1}^{\widetilde{r}_k}\widehat{\boldsymbol{\Lambda}}_{zk}^{[j,j]}\\ &\lesssim_P\delta_\eta+\max(\widetilde{r}_k-r_k^*,0)\delta_\eta\\ &\lesssim_P\delta_\eta.\end{aligned}\tag{36}$$

By Theorem 2.1 in Meng and Zheng (2010), (35), and $\min_{k\le K}\lambda_{r_k^*}(\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0}))>\kappa_z$,

$$\begin{aligned}
&\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger-[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2\\
&\le\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger-[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_F\\
&\le\max\left\{\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2^2,\left\|[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2^2\right\}\left\|\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\right\|_F\\
&\le\max\left\{\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2^2,\left\|[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2^2\right\}L^{1/2}\left\|\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})-\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})\right\|_2\\
&\lesssim_P\delta_\eta. \qquad (37)
\end{aligned}$$

By (33), (32), and (28), we have

$$\begin{aligned}
&\|\widehat{\mathbf{V}}_{xk}\widehat{\mathbf{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top-\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top\|_2\\
&\le\|\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\|_2\|\widehat{\mathbf{H}}_k-\mathbf{H}_k\|_2+\|\widehat{\mathbf{H}}_k\|_2\|\widehat{\mathbf{V}}_{xk}\widehat{\mathbf{\Lambda}}_{xk}^{1/2}-\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\|_2\\
&\lesssim_P\lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))\delta_\eta+\lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}.
\end{aligned}$$

Using (33) again together with the above inequality, (37), and (35) yields

$$\begin{aligned}
&\left\|\widehat{\mathbf{V}}_{xk}\widehat{\mathbf{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger-\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2\\
&\le\left\|\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\mathbf{H}_k\right\|_2\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger-[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2\\
&\quad+\left\|[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\right\|_2\left\|\widehat{\mathbf{V}}_{xk}\widehat{\mathbf{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top-\mathbf{V}_{xk}\mathbf{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top\right\|_2\\
&\lesssim_P\lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))\delta_\eta+\lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}. \qquad (38)
\end{aligned}$$

By Weyl's inequality (see Theorem 3.3.16(c) in Horn and Johnson (1994)) and (29), for all $\ell\in\mathcal{I}_0$ we have

$$|\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f}))-\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))|\le\|\widehat{\mathrm{cov}}(\boldsymbol{f})-\mathrm{cov}(\boldsymbol{f})\|_2\lesssim_P\delta_f.$$

Then by $\delta_f\le\delta_\eta=o(1)$ and $\lambda_L(\mathrm{cov}(\boldsymbol{f}))>1$, for all $\ell\in\mathcal{I}_0$ we have

$$\begin{aligned}
&\left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))-\lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\right|\\
&=\left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))+\lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\right|^{-1}|\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{f}))-\lambda_\ell(\mathrm{cov}(\boldsymbol{f}))|\\
&\le\lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f}))\,\|\widehat{\mathrm{cov}}(\boldsymbol{f})-\mathrm{cov}(\boldsymbol{f})\|_2\\
&\lesssim_P\delta_f=o(1). \qquad (39)
\end{aligned}$$

Thus, for all $\ell\in\mathcal{I}_0$,

$$\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))\ge\lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))-\left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))-\lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\right|\ge1-o_P(1),$$

and then

$$\begin{aligned}
&\left|\lambda_\ell^{-1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f})) - \lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f}))\right| \\
&= \left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f})) - \lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\right| \lambda_\ell^{-1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))\, \lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f})) \\
&\lesssim_P \delta_f. \qquad (40)
\end{aligned}$$

For all $k \le K$ and $\ell \in \mathcal{I}_0$, by (33), $\lambda_1(\mathrm{cov}(\boldsymbol{f})) \le \mathrm{tr}(\mathrm{cov}(\boldsymbol{f})) \le \sum_{k=1}^K r_k$, (30), $\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F \le \|\widehat{\boldsymbol{\eta}}^{(\ell)}\|_F = 1$, and (39), we obtain

$$\begin{aligned}
&\left|\widehat{\cos}\theta(w^{(\ell)}, z_k^{(\ell)}) - \cos\theta(w^{(\ell)}, z_k^{(\ell)})\right| \\
&= \left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f}))\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F - \lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\|\boldsymbol{\eta}_k^{(\ell)}\|_F\right| \\
&\le |\,\lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))|\left|\|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F - \|\boldsymbol{\eta}_k^{(\ell)}\|_F\right| + \|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F \left|\lambda_\ell^{1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f})) - \lambda_\ell^{1/2}(\mathrm{cov}(\boldsymbol{f}))\right| \\
&\lesssim_P \delta_\eta + \delta_f \\
&\lesssim_P \delta_\eta. \qquad (41)
\end{aligned}$$

For all $\ell \in \mathcal{I}_0$ and $j, k \le K$,

$$\cos\theta(z_j^{(\ell)}, z_k^{(\ell)}) = \frac{(\boldsymbol{\eta}_j^{(\ell)})^\top \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\boldsymbol{\eta}_k^{(\ell)}}{\|\boldsymbol{\eta}_j^{(\ell)}\|_F \|\boldsymbol{\eta}_k^{(\ell)}\|_F}.$$

By (33), (31), (29), and (32),

$$\begin{aligned}
&\left\|(\widehat{\boldsymbol{\eta}}_j^{(\ell)})^\top \|\widehat{\boldsymbol{\eta}}_j^{(\ell)}\|_F^{-1} \widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - (\boldsymbol{\eta}_j^{(\ell)})^\top \|\boldsymbol{\eta}_j^{(\ell)}\|_F^{-1} \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\right\|_2 \\
&\le \left\|(\widehat{\boldsymbol{\eta}}_j^{(\ell)})^\top \|\widehat{\boldsymbol{\eta}}_j^{(\ell)}\|_F^{-1}\right\|_2 \left\|\widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\right\|_2 \\
&\quad + \left\|\mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\right\|_2 \left\|(\widehat{\boldsymbol{\eta}}_j^{(\ell)})^\top \|\widehat{\boldsymbol{\eta}}_j^{(\ell)}\|_F^{-1} - (\boldsymbol{\eta}_j^{(\ell)})^\top \|\boldsymbol{\eta}_j^{(\ell)}\|_F^{-1}\right\|_2 \\
&\lesssim_P \delta_f + \delta_\eta \\
&\lesssim_P \delta_\eta,
\end{aligned}$$

and then,

$$\begin{aligned}
&\left|\widehat{\cos}\theta(z_j^{(\ell)}, z_k^{(\ell)}) - \cos\theta(z_j^{(\ell)}, z_k^{(\ell)})\right| \\
&\le \left\|\widehat{\boldsymbol{\eta}}_k^{(\ell)} \|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F^{-1}\right\|_2 \left\|(\widehat{\boldsymbol{\eta}}_j^{(\ell)})^\top \|\widehat{\boldsymbol{\eta}}_j^{(\ell)}\|_F^{-1} \widehat{\mathrm{cov}}(\boldsymbol{f}_j, \boldsymbol{f}_k) - (\boldsymbol{\eta}_j^{(\ell)})^\top \|\boldsymbol{\eta}_j^{(\ell)}\|_F^{-1} \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\right\|_2 \\
&\quad + \left\|(\boldsymbol{\eta}_j^{(\ell)})^\top \|\boldsymbol{\eta}_j^{(\ell)}\|_F^{-1} \mathrm{cov}(\boldsymbol{f}_j, \boldsymbol{f}_k)\right\|_2 \left\|\widehat{\boldsymbol{\eta}}_k^{(\ell)} \|\widehat{\boldsymbol{\eta}}_k^{(\ell)}\|_F^{-1} - \boldsymbol{\eta}_k^{(\ell)} \|\boldsymbol{\eta}_k^{(\ell)}\|_F^{-1}\right\|_2 \\
&\lesssim_P \delta_\eta. \qquad (42)
\end{aligned}$$

By (41) and (42), for $\ell \in \mathcal{I}_0$ we have

$$\begin{aligned}
&\left|\widetilde{\Delta}_{jk}^{(\ell)} - \Delta_{jk}^{(\ell)}\right| \\
&\le \left|[\widehat{\cos}\theta(w^{(\ell)}, z_j^{(\ell)}) + \widehat{\cos}\theta(w^{(\ell)}, z_k^{(\ell)})]^2 - [\cos\theta(w^{(\ell)}, z_j^{(\ell)}) + \cos\theta(w^{(\ell)}, z_k^{(\ell)})]^2\right| \\
&\quad + 4\left|\widehat{\cos}\theta(z_j^{(\ell)}, z_k^{(\ell)}) - \cos\theta(z_j^{(\ell)}, z_k^{(\ell)})\right| \\
&\le 4\left|[\widehat{\cos}\theta(w^{(\ell)}, z_j^{(\ell)}) + \widehat{\cos}\theta(w^{(\ell)}, z_k^{(\ell)})] - [\cos\theta(w^{(\ell)}, z_j^{(\ell)}) + \cos\theta(w^{(\ell)}, z_k^{(\ell)})]\right| \\
&\quad + 4\left|\widehat{\cos}\theta(z_j^{(\ell)}, z_k^{(\ell)}) - \cos\theta(z_j^{(\ell)}, z_k^{(\ell)})\right| \\
&\le 8 \max_{1\le k\le K}\left|\widehat{\cos}\theta(w^{(\ell)}, z_k^{(\ell)}) - \cos\theta(w^{(\ell)}, z_k^{(\ell)})\right| + 4\left|\widehat{\cos}\theta(z_j^{(\ell)}, z_k^{(\ell)}) - \cos\theta(z_j^{(\ell)}, z_k^{(\ell)})\right| \\
&\lesssim_P \delta_\eta. \qquad (43)
\end{aligned}$$

Now consider $\ell \in \mathcal{I}_0$ and $(j,k) \in \mathcal{I}_\Delta^{(\ell)}$. We have that $\widehat{\Delta}_{jk}^{(\ell)}$ and $\Delta_{jk}^{(\ell)}$ are nonnegative. Thus,

$$\left|(\widehat{\Delta}_{jk}^{(\ell)})^{1/2} - (\Delta_{jk}^{(\ell)})^{1/2}\right| \le \left|\widehat{\Delta}_{jk}^{(\ell)} - \Delta_{jk}^{(\ell)}\right|^{1/2} \le \left|\widetilde{\Delta}_{jk}^{(\ell)} - \Delta_{jk}^{(\ell)}\right|^{1/2} \lesssim_P \delta_\eta^{1/2}. \qquad (44)$$

From (41) and (44),

$$|\widehat{\alpha}_{jk}^{(\ell)} - \alpha_{jk}^{(\ell)}| \lesssim_P \delta_\eta + \delta_\eta^{1/2} \lesssim_P \delta_\eta^{1/2}.$$

Recall that $\min_{(j,k)\in\mathcal{I}_\Delta^{(\ell)}, \ell\in\mathcal{I}_0} |\cos\theta(z_j^{(\ell)}, z_k^{(\ell)})| > \kappa_{zz}$. By (42) and $\delta_\eta = o(1)$, with probability tending to 1 we have that $\widehat{\cos}\theta(z_j^{(\ell)}, z_k^{(\ell)}) \cdot \cos\theta(z_j^{(\ell)}, z_k^{(\ell)}) > 0$ and thus $\widehat{\alpha}_{jk}^{(\ell)}\alpha_{jk}^{(\ell)} > 0$. Without loss of generality, we assume $\alpha^{(\ell)} > 0$. Let $\mathcal{I}_+^{(\ell)} = \{(j,k) \in \mathcal{I}_\Delta^{(\ell)} : \alpha_{jk}^{(\ell)} > 0\}$, then $\alpha^{(\ell)} = \min\{\alpha_{jk}^{(\ell)} : \alpha_{jk}^{(\ell)} > 0, (j,k) \in \mathcal{I}_+^{(\ell)}\}$. With probability tending to 1, $\widehat{\alpha}^{(\ell)} = \min\{\widehat{\alpha}_{jk}^{(\ell)} : \widehat{\alpha}_{jk}^{(\ell)} > 0, (j,k) \in \mathcal{I}_+^{(\ell)}\}$. Due to Lemma S.1 in Shu et al. (2019), we simply assume $\widehat{\alpha}^{(\ell)} = \min\{\widehat{\alpha}_{jk}^{(\ell)} : \widehat{\alpha}_{jk}^{(\ell)} > 0, (j,k) \in \mathcal{I}_+^{(\ell)}\}$ in the rest of the proof. Without loss of generality, denote $\alpha_{12}^{(\ell)} = \alpha^{(\ell)}$. If $\widehat{\alpha}_{12}^{(\ell)} = \widehat{\alpha}^{(\ell)}$, then $|\widehat{\alpha}^{(\ell)} - \alpha^{(\ell)}| \lesssim_P \delta_\eta^{1/2}$. Otherwise, without loss of generality we denote $\widehat{\alpha}_{23}^{(\ell)} = \widehat{\alpha}^{(\ell)}$. If $|\alpha_{23}^{(\ell)} - \alpha_{12}^{(\ell)}| = O(\delta_\eta^{1/2})$, then $|\alpha_{12}^{(\ell)} - \widehat{\alpha}_{23}^{(\ell)}| \le |\alpha_{12}^{(\ell)} - \alpha_{23}^{(\ell)}| + |\alpha_{23}^{(\ell)} - \widehat{\alpha}_{23}^{(\ell)}| \lesssim_P \delta_\eta^{1/2}$. Otherwise, we have $\alpha_{23}^{(\ell)} - \alpha_{12}^{(\ell)} \gg \delta_\eta^{1/2}$, and then $\widehat{\alpha}_{23}^{(\ell)} - \widehat{\alpha}_{12}^{(\ell)} \ge (1 - o_P(1))(\alpha_{23}^{(\ell)} - \alpha_{12}^{(\ell)})$, which contradicts $\widehat{\alpha}_{12}^{(\ell)} > \widehat{\alpha}_{23}^{(\ell)} = \widehat{\alpha}^{(\ell)}$. Hence,

$$|\widehat{\alpha}^{(\ell)} - \alpha^{(\ell)}| \lesssim_P \delta_\eta^{1/2}. \qquad (45)$$

By (33), (40) and (45), for all $\ell \in \mathcal{I}_0$,

$$\begin{aligned}
&\left|\widehat{\alpha}^{(\ell)}\,\lambda_\ell^{-1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f})) - \alpha^{(\ell)}\,\lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f}))\right| \\
&\le \widehat{\alpha}^{(\ell)}\left|\lambda_\ell^{-1/2}(\widehat{\mathrm{cov}}(\boldsymbol{f})) - \lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f}))\right| + \lambda_\ell^{-1/2}(\mathrm{cov}(\boldsymbol{f}))|\widehat{\alpha}^{(\ell)} - \alpha^{(\ell)}| \\
&\lesssim_P \delta_f + \delta_\eta^{1/2} \\
&\lesssim_P \delta_\eta^{1/2}.
\end{aligned}$$

Then together with (33) and (30) gives

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{N}} - \mathbf{A}\mathbf{N}\|_2 \le \|\mathbf{A}\|_2\|\widehat{\mathbf{N}} - \mathbf{N}\|_F + \|\widehat{\mathbf{N}}\|_F\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 \lesssim_P \delta_\eta^{1/2}. \tag{46}$$

From the inequalities respectively below (S.12) and (S.22) in the supplement of Shu et al. (2019), we obtain

$$n^{-1}\|\mathbf{F}_k\|_F^2 = r_k + O_P(n^{-1/2})$$

and

$$\|\widehat{\mathbf{F}}_k - \mathbf{F}_k\|_F \le r_k^{1/2}\|\widehat{\mathbf{F}}_k - \mathbf{F}_k\|_2 \lesssim_P \min\Big\{1 + [p_k\,\lambda_1^{-1}(\mathrm{cov}(\boldsymbol{x}_k))\log p_k]^{1/2}, n^{1/2}\Big\} =: \delta_k.$$

Hence,

$$\|\mathbf{F}\|_F = \big(\sum_{k=1}^K \|\mathbf{F}_k\|_F^2\big)^{1/2} = O_P(n^{1/2}) \tag{47}$$

and

$$\|\widehat{\mathbf{F}} - \mathbf{F}\|_F = \Big(\sum_{k=1}^K \|\widehat{\mathbf{F}}_k - \mathbf{F}_k\|_F^2\Big)^{1/2} \lesssim_P \sum_{k=1}^K \delta_k. \tag{48}$$

By (33), (47), (46) and (48), we obtain

$$\begin{aligned}\|\widehat{\mathbf{C}}^{\mathcal{I}_0} - \mathbf{C}^{\mathcal{I}_0}\|_2 &\le \|\mathbf{F}\|_F\|\widehat{\mathbf{A}}\widehat{\mathbf{N}} - \mathbf{A}\mathbf{N}\|_2 + \|\widehat{\mathbf{A}}\|_2\|\widehat{\mathbf{N}}\|_F\|\widehat{\mathbf{F}} - \mathbf{F}\|_F \\ &\lesssim_P n^{1/2}\delta_\eta^{1/2} + \sum_{k=1}^K \delta_k.\end{aligned} \tag{49}$$

Using (33), (47), (38), (27), (35) and (49) yields

$$\begin{aligned}&\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_2 \\ &\le \|\mathbf{A}\mathbf{N}\mathbf{F}\|_2\Big\|\widehat{\mathbf{V}}_{xk}\widehat{\boldsymbol{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger - \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\Big\|_2 \\ &\quad + \Big\|\widehat{\mathbf{V}}_{xk}\widehat{\boldsymbol{\Lambda}}_{xk}^{1/2}\widehat{\mathbf{H}}_k^\top[\widehat{\mathrm{cov}}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\Big\|_2\|\widehat{\mathbf{C}}^{\mathcal{I}_0} - \mathbf{C}^{\mathcal{I}_0}\|_2 \\ &\lesssim_P n^{1/2}[\lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))\delta_\eta + \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}] \\ &\quad + \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))\Big[n^{1/2}\delta_\eta^{1/2} + \sum_{k=1}^K \delta_k\Big] \\ &\lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{1/2}\delta_\eta^{1/2}.\end{aligned} \tag{50}$$

By $\mathrm{rank}(\mathbf{M}_1\mathbf{M}_2) \le \min(\mathrm{rank}(\mathbf{M}_1), \mathrm{rank}(\mathbf{M}_2))$ and $\mathrm{rank}(\mathbf{M}_1 - \mathbf{M}_2) \le \mathrm{rank}(\mathbf{M}_1) + \mathrm{rank}(\mathbf{M}_2)$ for any real matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ with compatible sizes, we have $\mathrm{rank}(\widehat{\mathbf{C}}_k - \mathbf{C}_k) \le 2L$. Thus,

$$\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_F \le \|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_2[\mathrm{rank}(\widehat{\mathbf{C}}_k - \mathbf{C}_k)]^{1/2} \lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{1/2}\delta_\eta^{1/2}. \tag{51}$$

By (50), (51) and (22), we obtain

$$\max\left\{\frac{\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_2}{\|\mathbf{X}_k\|_2}, \frac{\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_F}{\|\mathbf{X}_k\|_F}\right\} \lesssim_P \delta_\eta^{1/2}. \tag{52}$$

By $\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\| \le \|\widehat{\mathbf{X}}_k - \mathbf{X}_k\| + \|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|$ for both the Frobenius norm and the spectral norm, (25) and (52), we obtain

$$\max\left\{\frac{\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\|_2}{\|\mathbf{X}_k\|_2}, \frac{\|\widehat{\mathbf{D}}_k - \mathbf{D}_k\|_F}{\|\mathbf{X}_k\|_F}\right\} \lesssim_P \delta_\eta^{1/2}.$$

Now consider the estimated proportion of explained variance. Note that $\|\widehat{\mathbf{X}}_k\|_F^2/n = \mathrm{tr}(\widehat{\mathbf{X}}_k\widehat{\mathbf{X}}_k^\top/n) = \mathrm{tr}(\widehat{\mathrm{cov}}(\boldsymbol{x}_k))$. By inequality (S.16) of Shu et al. (2019),

$$\begin{aligned}\left|\frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2 - \mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))\right| &= \left|\,\mathrm{tr}(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) - \mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))\right| \\ &\le \sum_{\ell=1}^{r_k}\left|\,\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) - \lambda_\ell(\mathrm{cov}(\boldsymbol{x}_k))\right| \\ &\lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}, \end{aligned} \tag{53}$$

and by their (S.17),

$$\frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2 = \mathrm{tr}(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) = \sum_{\ell=1}^{r_k}\lambda_\ell(\widehat{\mathrm{cov}}(\boldsymbol{x}_k)) \ge r_k(1-o_P(1))\,\lambda_{r_k}(\mathrm{cov}(\boldsymbol{x}_k)). \tag{54}$$

Since (51) and

$$\big\|\mathbf{C}_k\big\|_F \le L^{1/2}\big\|\mathbf{C}_k\big\|_2 = L^{1/2}\big\|\mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\mathbf{ANF}\big\|_2 \lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{1/2},$$

we obtain

$$\big\|\widehat{\mathbf{C}}_k\big\|_F \le \|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_F + \big\|\mathbf{C}_k\big\|_F \lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))n^{1/2}. \tag{55}$$

Then,

$$\begin{aligned}\left|\frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 - \frac{1}{n}\|\mathbf{C}_k\|_F^2\right| &= \frac{1}{n}\left|\|\widehat{\mathbf{C}}_k\|_F - \|\mathbf{C}_k\|_F\right|(\|\widehat{\mathbf{C}}_k\|_F + \|\mathbf{C}_k\|_F) \\ &\le \frac{1}{n}\|\widehat{\mathbf{C}}_k - \mathbf{C}_k\|_F(\|\widehat{\mathbf{C}}_k\|_F + \|\mathbf{C}_k\|_F) \\ &\lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k))\delta_\eta^{1/2}. \end{aligned} \tag{56}$$

From the central limit theorem,

$$\left\|\frac{1}{n}\mathbf{FF}^\top - \mathrm{cov}(\boldsymbol{f})\right\|_2 \le \sum_{k=1}^K r_k\left\|\frac{1}{n}\mathbf{FF}^\top - \mathrm{cov}(\boldsymbol{f})\right\|_{\max} \lesssim_P n^{-1/2}.$$

Let $\mathbf{Q}_k = \mathbf{V}_{xk}\boldsymbol{\Lambda}_{xk}^{1/2}\mathbf{H}_k^\top[\mathrm{cov}(\boldsymbol{z}_k^{\mathcal{I}_0})]^\dagger\mathbf{AN}$, then $\|\mathbf{Q}_k\|_2 \lesssim_P \lambda_1^{1/2}(\mathrm{cov}(\boldsymbol{x}_k))$. By Weyl's inequality (see Theorem 3.3.16(c) in Horn and Johnson (1994)),

$$\begin{aligned}\max_{\ell\le L}\left|\lambda_\ell(\frac{1}{n}\mathbf{C}_k\mathbf{C}_k^\top) - \lambda_\ell(\mathrm{cov}(\boldsymbol{c}_k))\right| &\le \left\|\frac{1}{n}\mathbf{C}_k\mathbf{C}_k^\top - \mathrm{cov}(\boldsymbol{c}_k)\right\|_2 \\ &= \left\|\mathbf{Q}_k\frac{1}{n}\mathbf{FF}^\top\mathbf{Q}_k^\top - \mathbf{Q}_k\,\mathrm{cov}(\boldsymbol{f})\mathbf{Q}_k^\top\right\|_2 \\ &\le \|\mathbf{Q}_k\|_2\left\|\frac{1}{n}\mathbf{FF}^\top - \mathrm{cov}(\boldsymbol{f})\right\|_2\|\mathbf{Q}_k^\top\|_2 \\ &\lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}. \end{aligned}$$

Then applying the same skill used for (53) yields

$$\left|\frac{1}{n}\|\mathbf{C}_k\|_F^2 - \mathrm{tr}(\mathrm{cov}(\boldsymbol{c}_k))\right| \le \sum_{\ell=1}^{L}\left|\lambda_\ell(\frac{1}{n}\mathbf{C}_k\mathbf{C}_k^\top) - \lambda_\ell(\mathrm{cov}(\boldsymbol{c}_k))\right| \lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k))n^{-1/2}. \tag{57}$$

Combining (56) and (57) with the triangle inequality gives

$$\left|\frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 - \mathrm{tr}(\mathrm{cov}(\boldsymbol{c}_k))\right| \lesssim_P \lambda_1(\mathrm{cov}(\boldsymbol{x}_k))\delta_\eta^{1/2}. \tag{58}$$

From (33), (53), (54), (55), (58) and (24), we have

$$\begin{aligned}
&\left|\frac{\frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2}{\frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2} - \frac{\mathrm{tr}(\mathrm{cov}(\boldsymbol{c}_k))}{\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))}\right| \\
&\le \left|\frac{1}{\frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2} - \frac{1}{\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))}\right| \cdot \frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 + \left|\frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 - \mathrm{tr}(\mathrm{cov}(\boldsymbol{c}_k))\right| \frac{1}{\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))} \\
&\le \frac{\left|\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k)) - \frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2\right|}{\frac{1}{n}\|\widehat{\mathbf{X}}_k\|_F^2\,\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))} \cdot \frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 + \left|\frac{1}{n}\|\widehat{\mathbf{C}}_k\|_F^2 - \mathrm{tr}(\mathrm{cov}(\boldsymbol{c}_k))\right| \frac{1}{\mathrm{tr}(\mathrm{cov}(\boldsymbol{x}_k))} \\
&\lesssim_P \delta_\eta^{1/2}.
\end{aligned}$$

The proof is complete. ■

Appendix B. Additional Simulation Results

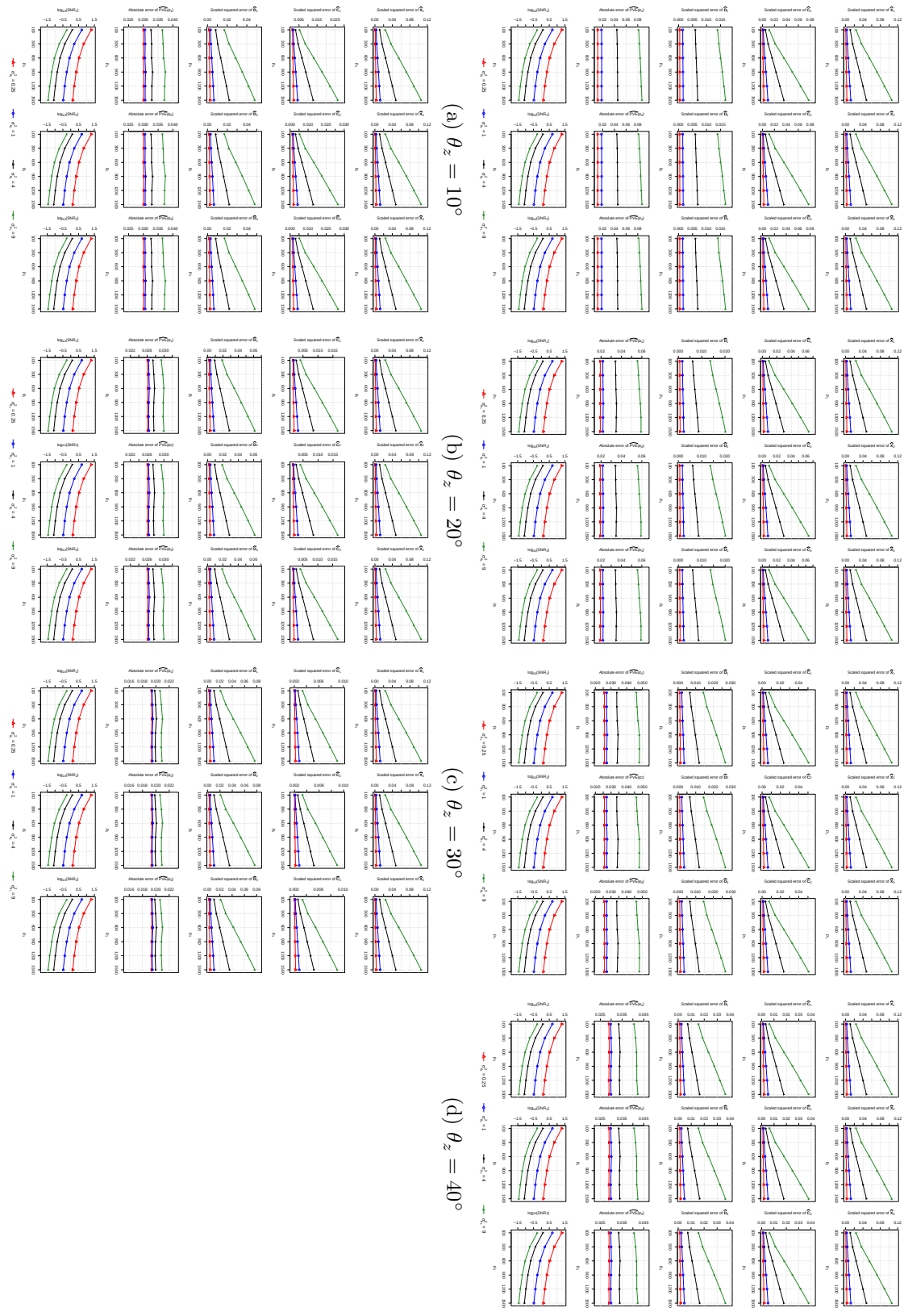
In Setup 1.1, the angle $\theta_z = 10°, 20°, 30°, 40°, 50°, 60°, 70°$ corresponds to $\mathrm{PVE}(\boldsymbol{c}_k) = 0.853, 0.702, 0.552, 0.409, 0.279, 0.167, 0.079$ for all $k \in \{1, 2, 3\}$. In Setup 2.1, the covariance matrix $\mathrm{cov}(\boldsymbol{f}) \in \mathbb{R}^{15\times 15}$ has blocks

$$\mathrm{cov}(\boldsymbol{f}_1, \boldsymbol{f}_2) = \begin{bmatrix} 0.02498103503160578 & -0.3734791596502449 & -0.1482674122573037 & -0.3913807076061239 & -0.05845072081373771 \\ 0.1298912403724416 & -0.2915966482089937 & -0.703223066831662 & -0.286977394728156 & -0.07037562289439672 \\ -0.4691315902716665 & -0.02216628581934877 & -0.05789731182102772 & -0.1224434530178697 & 0.7359965879693088 \\ -0.005270967060252731 & -0.1916047000827934 & 0.1572469950904809 & -0.1862928969932901 & 0.0648022978041196 \\ 0.3309749556233325 & 0.2910731038141944 & -0.2222302484678626 & 0.4183644600274041 & -0.09116219316544609 \end{bmatrix},$$

$$\mathrm{cov}(\boldsymbol{f}_1, \boldsymbol{f}_3) = \begin{bmatrix} -0.1652455953442644 & 0.07288409202801582 & 0.4797927991048995 & -0.1974810941368655 & 0.2123320697504773 \\ -0.3889488816571995 & 0.05377416249857463 & 0.5653871787847853 & 0.03845218160536631 & -0.2069628634535125 \\ 0.4125592431747815 & -0.7372033575312142 & 0.2721804829221633 & -0.0862772040030661 & -0.2227478031028198 \\ -0.02345535210198419 & -0.1075518721538277 & 0.1394751370539585 & -0.1625882523272944 & 0.3301641568167817 \\ -0.3328426143159536 & -0.09361178321406048 & -0.4483940610130605 & 0.3455811570541347 & -0.09767404221183135 \end{bmatrix},$$

$$\mathrm{cov}(\boldsymbol{f}_2, \boldsymbol{f}_3) = \begin{bmatrix} -0.1234093117538375 & 0.2223022967058531 & -0.3593383789512091 & 0.04344070064196999 & 0.2617381817815529 \\ -0.09993460814692552 & -0.008819786526375878 & -0.4039397802979183 & 0.2933537865045707 & -0.2650032054127345 \\ 0.5075563895372593 & -0.1098865559264541 & -0.4771360952896037 & -0.1119099874049149 & 0.2079731636733454 \\ -0.08232391689469482 & -0.01395485249078317 & -0.5724368834706903 & 0.3121430368957581 & -0.1821568224740747 \\ 0.3937761144502051 & -0.6998227270213208 & 0.1161733947993463 & -0.04568041770157075 & -0.1795827017135321 \end{bmatrix},$$

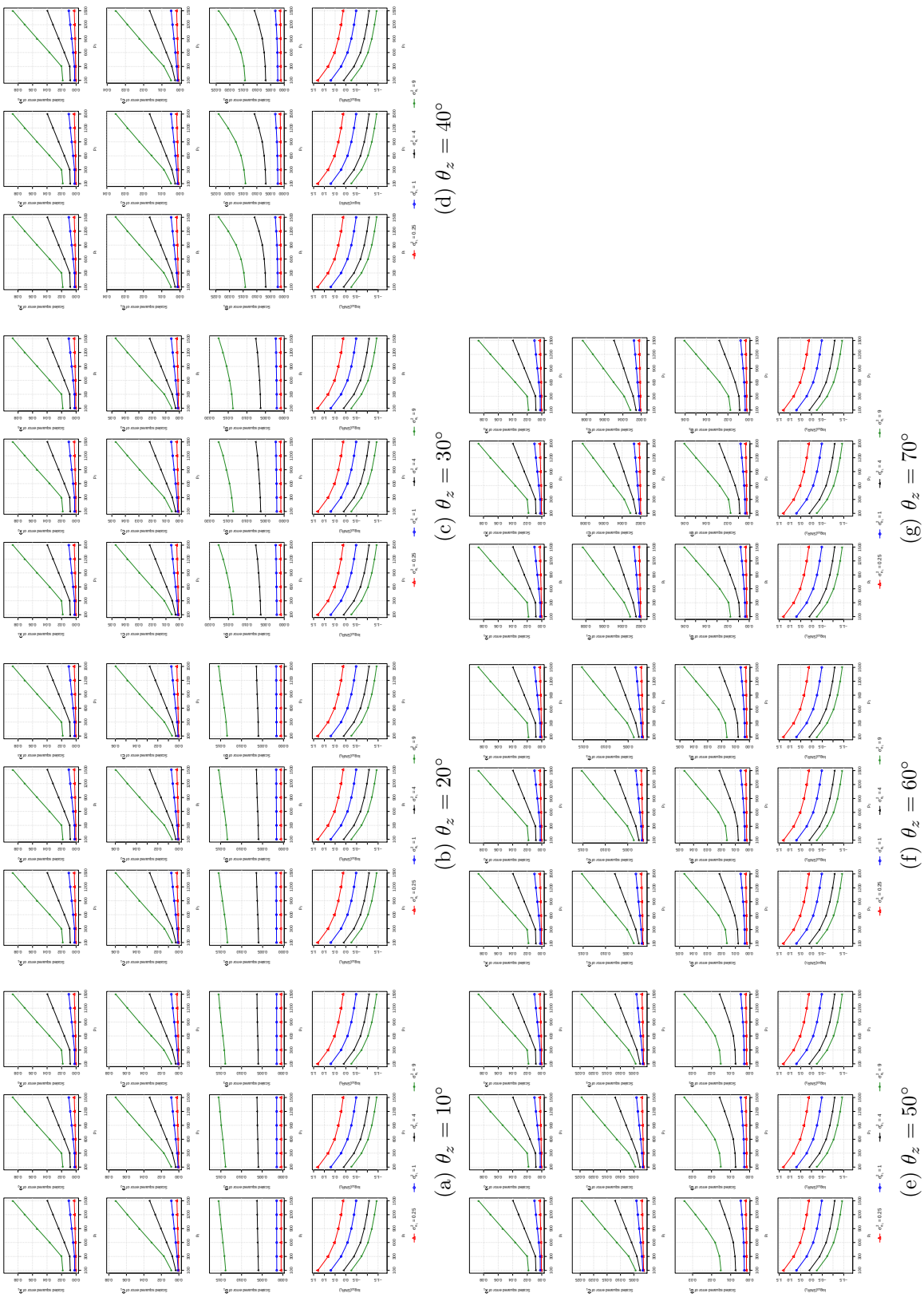
and $\mathrm{cov}(\boldsymbol{f}_k) = \mathbf{I}_{5\times 5}$ for $k = 1, 2, 3$. Figures 5–11 show the additional simulation results for Setups 1.1-2.2. The result analysis described in Section 4.2 also holds here.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$

(e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

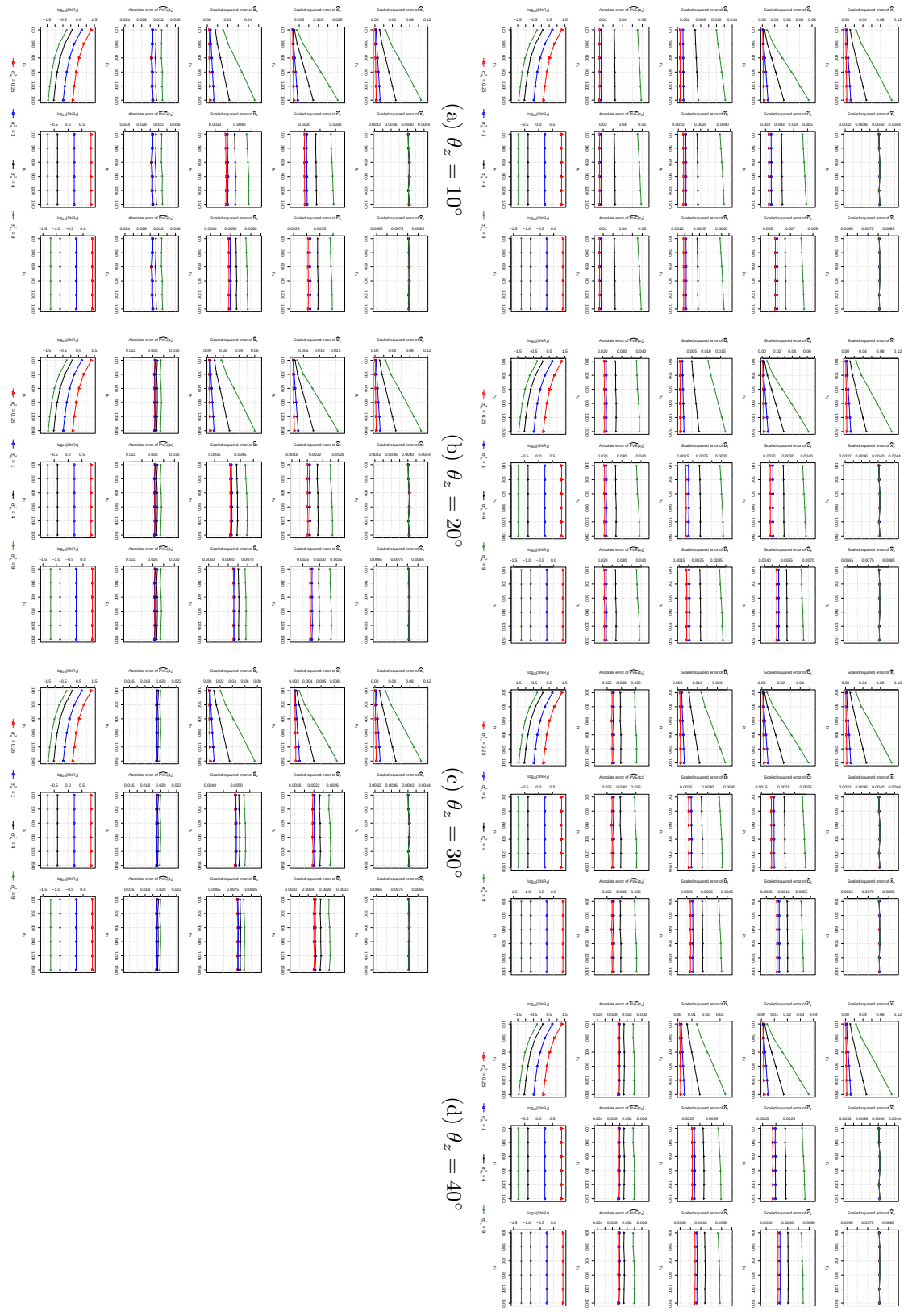
Figure 5: Average errors of D-GCCA estimates over 1000 replications in the Frobenius norm for Setup 1.1 using true nuisance parameters.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$

(e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

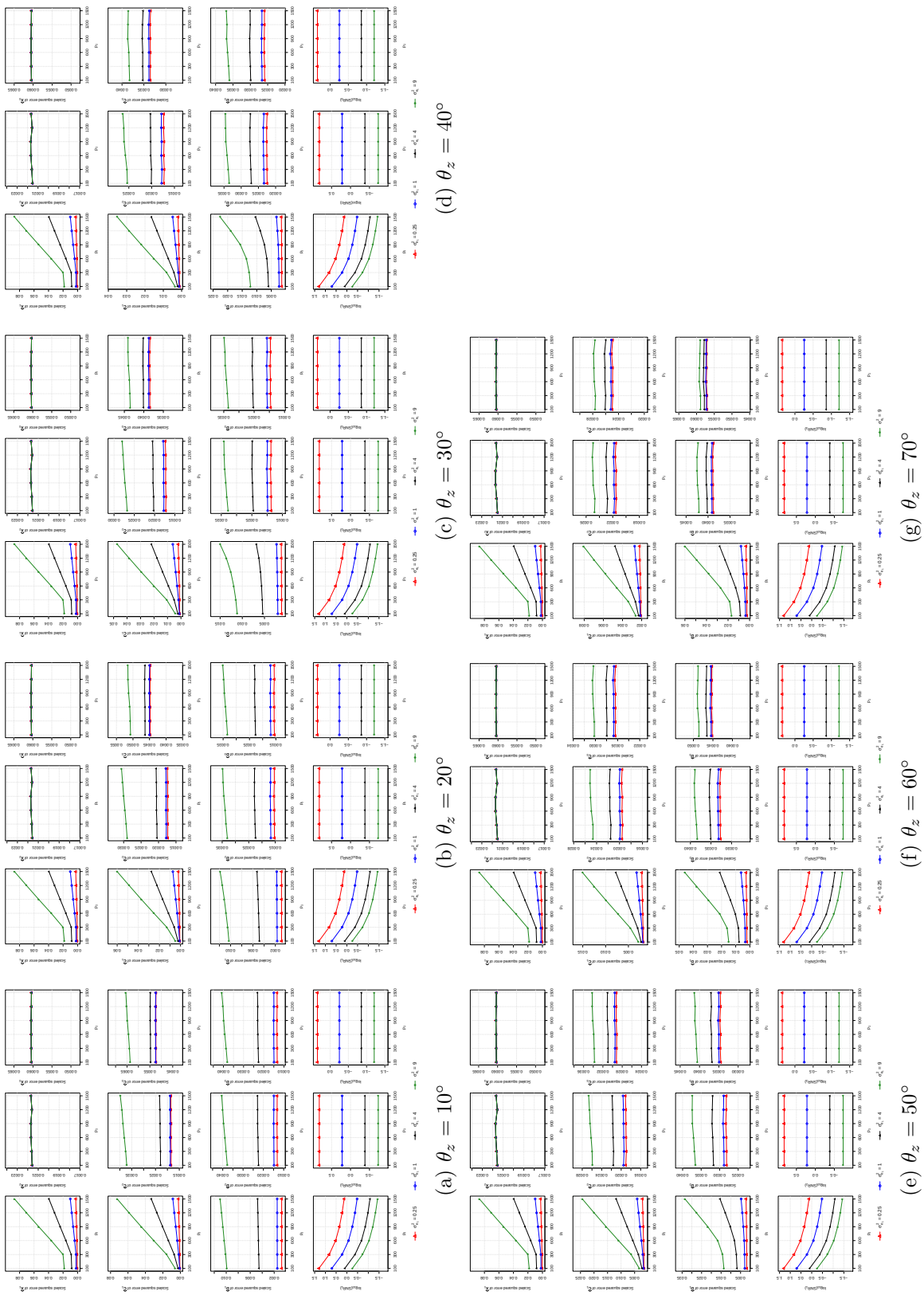
Figure 6: Average errors of D-GCCA estimates over 1000 replications in the spectral norm for Setup 1.1 using true nuisance parameters.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$

(e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

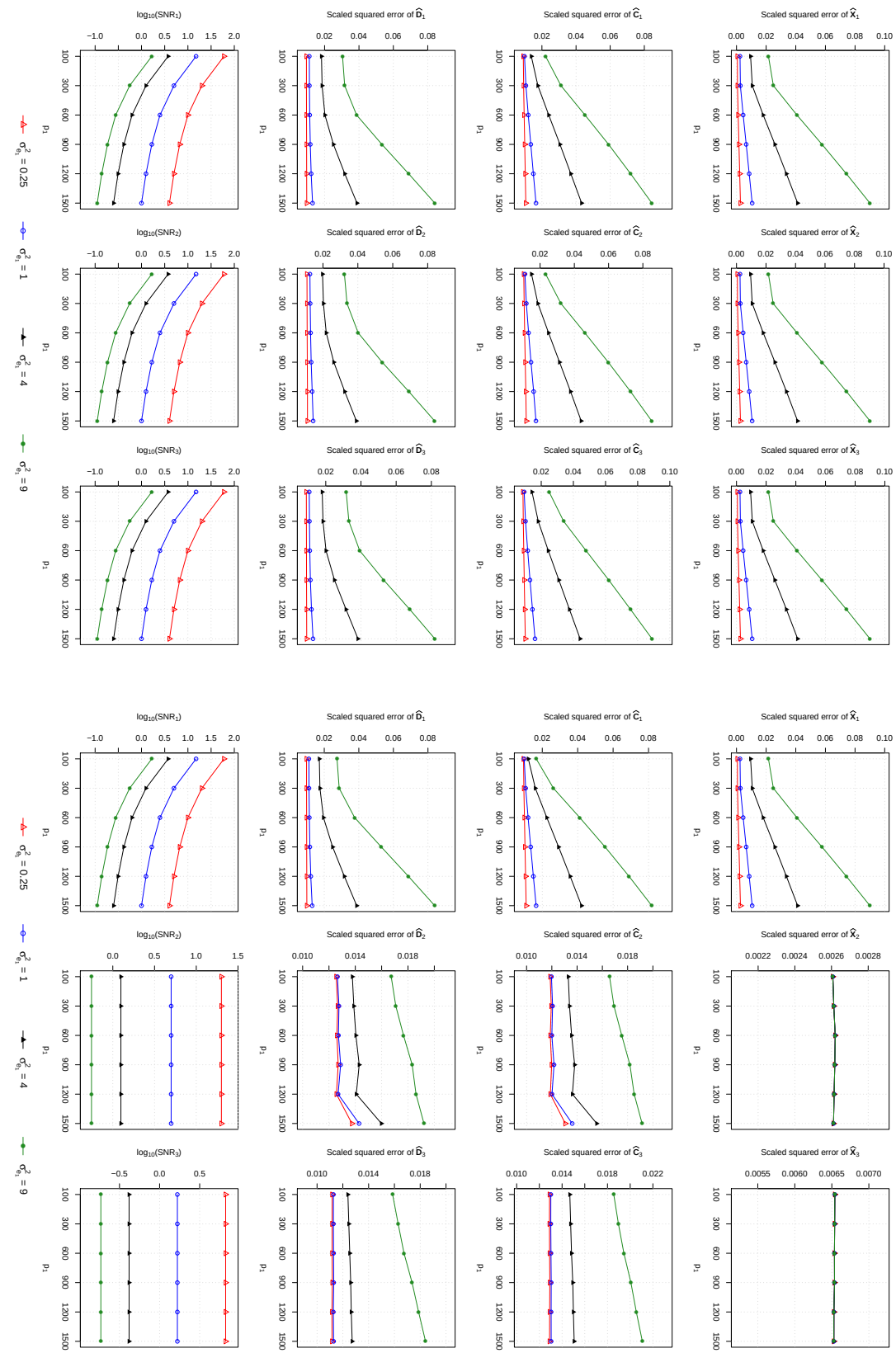
Figure 7: Average errors of D-GCCA estimates over 1000 replications in the Frobenius norm for Setup 1.2 using true nuisance parameters.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$

(e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

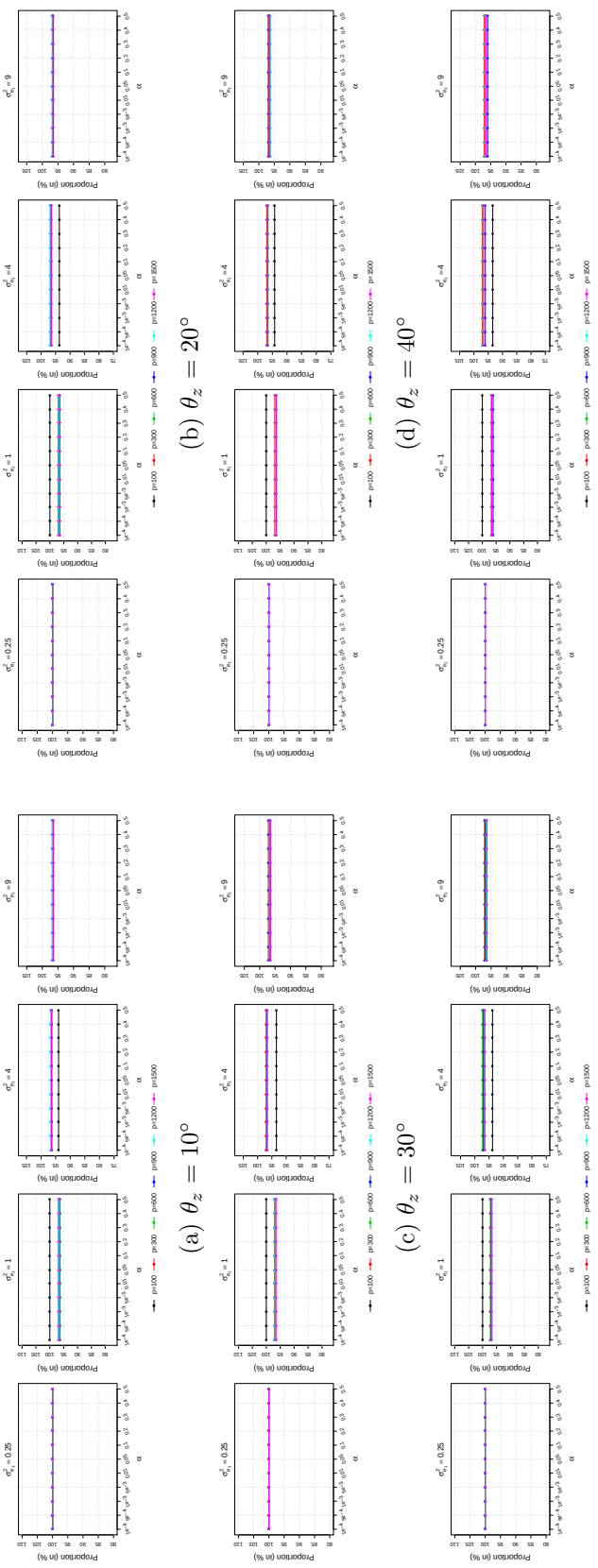
Figure 8: Average errors of D-GCCA estimates over 1000 replications in the spectral norm for Setup 1.2 using true nuisance parameters.



(a) Setup 2.1

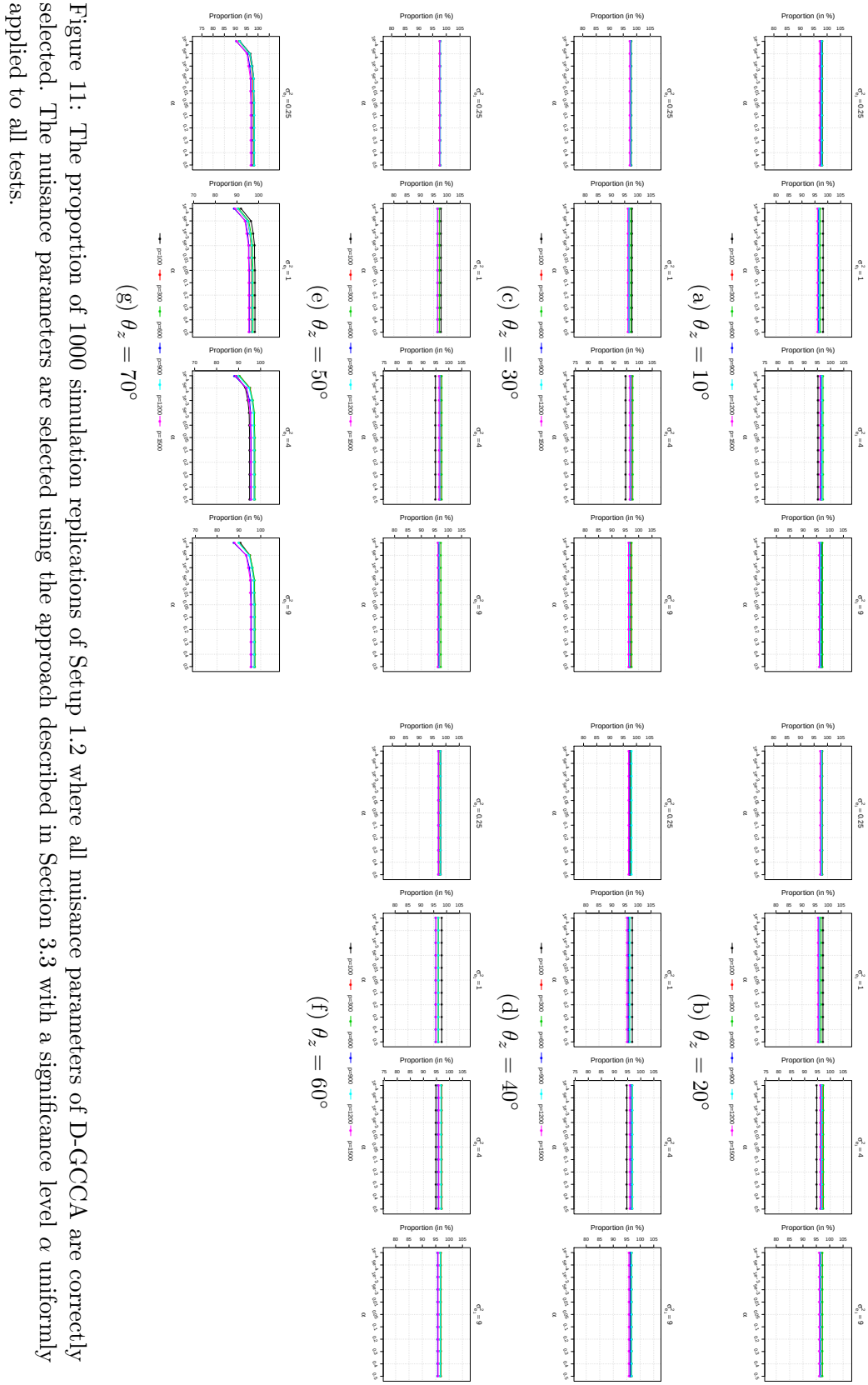
(b) Setup 2.2

Figure 9: Average errors of D-GCCA estimates over 1000 replications in the spectral norm for Setup 2.1 and Setup 2.2 using true nuisance parameters.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$ (e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

Figure 10: The proportion of 1000 simulation replications of Setup 1.1 where all nuisance parameters of D-GCCA are correctly selected. The nuisance parameters are selected using the approach described in Section 3.3 with a significance level α uniformly applied to all tests.



(a) $\theta_z = 10°$ (b) $\theta_z = 20°$ (c) $\theta_z = 30°$ (d) $\theta_z = 40°$ (e) $\theta_z = 50°$ (f) $\theta_z = 60°$ (g) $\theta_z = 70°$

Figure 11: The proportion of 1000 simulation replications of Setup 1.2 where all nuisance parameters of D-GCCA are correctly selected. The nuisance parameters are selected using the approach described in Section 3.3 with a significance level α uniformly applied to all tests.

References

D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

M. S. Bartlett. The statistical significance of canonical correlations. *Biometrika*, 32:29–37, 1941.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1): 289–300, 1995.

A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics*, 41(3):1055–1084, 2013.

R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. Thomas Yeo. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(5):2322–2345, 2011.

C. R. Cabanski, Y. Qi, X. Yin, E. Bair, M. C. Hayward, C. Fan, J. Li, M. D. Wilkerson, J. S. Marron, C. M. Perou, and D. N. Hayes. SWISS MADE: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*, 5(3):e9905, 2010.

T. Caliński and M. Krzyśko. A closed testing procedure for canonical correlations. *Communications in Statistics - Theory and Methods*, 34(5):1105–1116, 2005.

J. D. Campbell, C. Yau, R. Bowlby, Y. Liu, K. Brennan, H. Fan, A. M. Taylor, C. Wang, V. Walter, R. Akbani, et al. Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell Reports*, 23(1):194–212, 2018.

J. D. Carroll. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th Annual Convention of the American Psychological Association*, pages 227–228, 1968.

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

Q. Chen and Z. Fang. Improved inference on the rank of a matrix. *Quantitative Economics*, 10:1787–1824, 2019.

G. Ciriello, M. L. Gatza, A. H. Beck, M. D. Wilkerson, S. K. Rhie, A. Pastore, H. Zhang, M. McLellan, C. Yau, C. Kandoth, and R. Bowlby. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.

K. L. Crawford, S. C. Neu, and A. W. Toga. The image and data archive at the laboratory of neuro imaging. *Neuroimage*, 124:1080–1083, 2016.

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

B. Draper, M. Kirby, J. Marks, T. Marrinan, and C. Peterson. A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32, 2014.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, Series B*, 75(4):603–680, 2013.

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

A. K. Gupta and T. Varga. Rank of a quadratic form in an elliptically contoured matrix random variable. *Statistics & probability letters*, 12(2):131–134, 1991.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 2009.

K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

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

H. Hotelling. Analysis of a complex statistical variable into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

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

H. Huang. Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, 18(45):1–21, 2017.

M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt. The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, 130:453–459, 2017.

J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

N. Kishore Kumar and J. Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

D. C. Koboldt, R. S. Fulton, M. D. McLellan, Heather S., Joelle Kalicki-Veizer, J. F. McMichael, L. L. Fulton, D. J. Dooling, D. Li, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

D. N. Lawley. Tests of significance in canonical analysis. *Biometrika*, 46(1/2):59–66, 1959.

Y. Li, F. Wu, and A. Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19:325–340, 2018.

E. F. Lock and D. B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20): 2610–2616, 2013.

E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542, 2013.

T. Löfstedt and J. Trygg. OnPLS–a novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, 25(8):441–455, 2011.

L. Meng and B. Zheng. The optimal perturbation bounds of the Moore-Penrose inverse under the Frobenius norm. *Linear Algebra and its Applications*, 432(4):956–963, 2010.

R. R. Nadakuditi and J. W. Silverstein. 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(3):468–480, 2010.

M. J. O'Connell and E. F. Lock. R.JIVE for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879, 2016.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, and J. F. Quackenbush. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009.

U. Saeed, J. Compagnone, R. I. Aviv, A. P. Strafella, S. E. Black, A. E. Lang, and M. Masellis. Imaging biomarkers in parkinson's disease and parkinsonian syndromes: current and emerging concepts. *Translational Neurodegeneration*, 6(1):8, Mar 2017.

M. Schouteden, K. Van Deun, T. F. Wilderjans, and I. Van Mechelen. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behavior Research Methods*, 46(2):576–587, 2014.

H. Shu, X. Wang, and H. Zhu. D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, page DOI: 10.1080/01621459.2018.1543599, 2019.

A. K. Smilde, I. Mge, T. Ns, T. Hankemeier, M. A. Lips, H. A. L. Kiers, E. Acar, and R. Bro. Common and distinct components in data fusion. *Journal of Chemometrics*, 31 (7):e2900, 2017. ISSN 1099-128X.

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

M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

M. van de Velden. On generalized canonical correlation analysis. In *Proceedings of the 58th World Statistical Congress*, pages 758–765, 2011.

F. M. van der Kloet, P. Sebastián-León, A. Conesa, A. K. Smilde, and J. A. Westerhuis. Separating common from distinctive variation. *BMC bioinformatics*, 17(5):S195, 2016.

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, 2012.

W. Wang and J. Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of Statistics*, 45(3):1342–1374, 2017.

M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, et al. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

Y. Yin, Z. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4): 509–521, 1988.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

G. Zhou, A. Cichocki, Y. Zhang, and D. P. Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2426 – 2439, 2016.

J. Zhou and X. He. Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, 36(4):1649–1668, 2008.