Electronic Journal of Statistics
Vol. 16 (2022) 2475–2517
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2008

CDPA: Common and distinctive pattern analysis between high-dimensional datasets*†

Hai Shu

Department of Biostatistics, School of Global Public Health, New York University
e-mail: hs120@nyu.edu

Zhe Qu

Department of Mathematics, School of Science and Engineering, Tulane University

Abstract: A representative model in integrative analysis of two high-dimensional correlated datasets is to decompose each data matrix into a low-rank common matrix generated by latent factors shared across datasets, a low-rank distinctive matrix corresponding to each dataset, and an additive noise matrix. Existing decomposition methods claim that their common matrices capture the common pattern of the two datasets. However, their so-called common pattern only denotes the common latent factors but ignores the common pattern between the two coefficient matrices of these common latent factors. We propose a new unsupervised learning method, called the common and distinctive pattern analysis (CDPA), which appropriately defines the two types of data patterns by further incorporating the common and distinctive patterns of the coefficient matrices. A consistent estimation approach is developed for high-dimensional settings, and shows reasonably good finite-sample performance in simulations. Our simulation studies and real data analysis corroborate that the proposed CDPA can provide better characterization of common and distinctive patterns and thereby benefit data mining.

Keywords and phrases: Canonical variable, data integration, factor pattern, graph matching, mixing channel, principal vector.

Received April 2021.

Contents

*Dr. Shu's research was partially supported by the grant R21AG070303 from the National Institutes of Health and a startup fund from New York University.

†The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, New York University, or Tulane University.

1. Introduction

Modern biomedical studies often collect multiple types of large-scale datasets on a common set of objects [8, 23]. For example, The Cancer Genome Atlas (TCGA) [18] collected for tumor samples the multi-platform genomic data such as mRNA expression and DNA methylation; the Human Connectome Project (HCP) [52] acquired multi-modal brain imaging data, including structural MRI and functional MRI, from healthy adults. The use of multiple data types can allow us to enhance understanding the mechanisms underlying complex diseases like cancers [26, 5] and neurodegenerative diseases [55, 43], or to improve the performance in various learning tasks such as clustering and classification [51, 46].

The most straightforward approach to the integrative analysis of multi-type datasets is to concatenate all their data matrices into one matrix and then implement standard data analysis tools. One such example is the simultaneous component analysis (SCA) [47], which applies the principal component analysis (PCA) to the concatenated data matrix and thus is also known as SUM-PCA. These methods are simple to implement, but they are unable to explore or interpret the relationships among datasets. As pioneers to overcome this drawback, the canonical correlation analysis (CCA) [20, 33] and its various generalizations [6, 24, 49] measure the correlations and extract the most correlated components among datasets. The CCA methods only account for correlated features and fail

to reveal a more detailed relationship on the common and distinctive patterns across datasets.

A family of data integration methods has emerged recently to identify and separate the common and distinctive variations across datasets, including orthogonal n-block partial least squares (OnPLS) [31], distinctive and common components with SCA (DISCO-SCA) [44], common orthogonal basis extraction (COBE) [58], joint and individual variation explained (JIVE) [30], angle-based JIVE (AJIVE) [13], and decomposition-based CCA (D-CCA) [45]. Consider the case with two datasets. All these methods decompose each data matrix into a low-rank *common matrix* generated by latent factors shared across datasets,[1] a low-rank *distinctive matrix* corresponding to each dataset, and an additive noise matrix. Despite different constraints in the decomposition, these methods refer the common pattern of the two datasets to the common latent factors, but ignore the common pattern between the two coefficient matrices of these common latent factors. It may be more appropriate to name their "common" and "distinctive" matrices as *common-source* and *distinctive-source matrices*.

We propose a new unsupervised learning method, called the common and distinctive pattern analysis (CDPA), to improve the delineation of the common and distinctive patterns between two correlated datasets. The CDPA method defines the common pattern by incorporating both the common latent factors and the common pattern of their coefficient matrices, and determines each distinctive pattern as the residual part of the corresponding signal dataset. In factor analysis [17], a coefficient matrix of latent factors is called a *factor pattern matrix*, containing the factor loadings (i.e., coefficients) that represent the contributions of latent factors to the signal data. A coefficient matrix is also known as a *mixing channel* in signal processing [39, 41] which introduces correlations into the uncorrelated source variables to generate the output data. Hence, the two coefficient matrices of the common latent factors for two correlated datasets may contain common and distinctive patterns of the ways in which these common latent factors form their corresponding common-source matrices. Such common and distinctive patterns in the two coefficient matrices are also important and should be separated into the common and distinctive patterns of the two datasets. Our defined common-pattern and distinctive-pattern matrices together with the aforementioned common-source and distinctive-source matrices constitute a more comprehensive picture that depicts the relationship of two datasets.

Three challenging issues arise in the construction and estimation of common-pattern and distinctive-pattern matrices: (i) There exists the row matching problem of the two coefficient matrices, or equivalently the variable pairing problem of the two datasets, if the rows of either observed data matrix can be arbitrarily ordered independent of the other matrix; (ii) The common pattern of the two coefficient matrices must be identified; (iii) Recovering the high-dimensional

[1] The common matrices of OnPLS may have different sets of latent factors. As a post-processing step [51], the same set of latent factors can be obtained as an orthonormal basis of the vector space spanned by all these sets of latent factors. The common matrices remain unchanged after this post-processing step.

common-pattern and distinctive-pattern matrices confronts the curse of dimensionality where the unknown large covariance matrices may not be consistently estimated by the traditional sample covariance matrices [56]. We successfully convert the row matching problem (i) into the classic graph matching problem [32]. We extract the common pattern in (ii) by our extended analogy of the state-of-the-art D-CCA. To address the challenge (iii), we develop consistent estimators of proposed common-pattern and distinctive-pattern matrices under the high-dimensional spiked covariance model [12, 53, 45], which has been widely used in various fields, such as signal processing [36], machine learning [21], and economics [7].

The rest of this article is organized as follows. Section 2 introduces the CCA and D-CCA methods as preliminaries. Our CDPA method and its consistent estimation are established in Section 3. Section 4 examines the finite-sample performance of proposed estimators via simulations. We also compare CDPA with six D-CCA-type methods through simulated data in Section 4 and through two real-data examples from HCP and TCGA in Section 5. Section 6 concludes with discussion. All theoretical proofs and additional simulation and real-data results are provided in Appendices. A Python package for the proposed CDPA method is available at https://github.com/shu-hai/CDPA.

2. Preliminaries

Let $\mathbf{Y}_k \in \mathbb{R}^{p_k \times n}$ for $k \in \{1, 2\}$ be the k-th dataset obtained on a common set of n objects, where p_k is the number of variables. The decomposition model considered in aforementioned existing methods (e.g., D-CCA) is

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k \in \mathbb{R}^{p_k \times n} \tag{1}$$

for which the n columns of each matrix are assumed to be independent and identically distributed (i.i.d.) copies of the corresponding mean-zero random vector in

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

where $\{\mathbf{X}_k, \boldsymbol{x}_k\}_{k=1}^2$ and $\{\mathbf{E}_k, \boldsymbol{e}_k\}_{k=1}^2$ are signals and noises, respectively, $\{\mathbf{C}_k\}_{k=1}^2$ and $\{\boldsymbol{c}_k\}_{k=1}^2$ are common-source matrices and random vectors that are generated from the common latent factors of the two datasets, and $\mathbf{D}_k$ and $\boldsymbol{d}_k$ are the distinctive-source matrix and random vector from distinctive latent factors of the k-th dataset. Write each k-th common-source random vector by $\boldsymbol{c}_k = \mathbf{B}_k(c_1, \ldots, c_{L_{12}})^\top$, where $c_1, \ldots, c_{L_{12}}$ are the common latent factors and $\mathbf{B}_k$ is their coefficient matrix. The common pattern of $\mathbf{B}_1$ and $\mathbf{B}_2$ is not considered by the existing methods, which motivates our current research.

We start with signal vectors $\{\boldsymbol{x}_k\}_{k=1}^2$ for simplicity, and introduce the CCA and D-CCA methods in the two following subsections. The signal estimation is deferred to Section 3.3.

Notation. For any matrix $\mathbf{M} = (M_{ij})_{1 \le i \le p, 1 \le j \le n} \in \mathbb{R}^{p \times n}$, denote the ℓ-th largest singular value and the ℓ-th largest eigenvalue (if $p = n$) by $\sigma_\ell(\mathbf{M})$

and $\lambda_\ell(\mathbf{M})$ respectively, the spectral norm $\|\mathbf{M}\|_2 = \sigma_1(\mathbf{M})$, the Frobenius norm $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^n M_{ij}^2}$, the matrix $\mathcal{L}^\infty$ norm $\|\mathbf{M}\|_\infty = \max_{1\le i\le p} \sum_{j=1}^n |M_{ij}|$, and the max norm $\|\mathbf{M}\|_{\max} = \max_{i,j} |\mathbf{M}_{ij}|$. Let $\mathbf{M}^{[s:t,u:v]}$, $\mathbf{M}^{[s:t,:]}$ and $\mathbf{M}^{[:,u:v]}$ denote 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. Write the Moore-Penrose pseudoinverse and the column space of $\mathbf{M}$ by $\mathbf{M}^\dagger$ and $\mathrm{colsp}(\mathbf{M})$, respectively. Let $[\mathbf{M}_1; ...; \mathbf{M}_L] = [\mathbf{M}_1^\top, \ldots, \mathbf{M}_L^\top]^\top$ for matrices $\mathbf{M}_1, \ldots, \mathbf{M}_L$ with the same number of columns. Denote the j-th entry of a vector $\boldsymbol{v} \in \mathbb{R}^p$ by $\boldsymbol{v}^{[j]}$. Write $\mathrm{span}(\boldsymbol{v}^\top) = \mathrm{span}(\{\boldsymbol{v}^{[j]}\}_{j=1}^p) = \{\sum_{j=1}^p a_j \boldsymbol{v}^{[j]} : \forall a_j \in \mathbb{R}\}$. For vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L$ of same length, let $[\boldsymbol{v}_\ell]_{\ell=1}^L = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L)$. Denote the angle between two elements v_1 and v_2 in an inner product space by $\theta(v_1, v_2)$. Let $(\mathcal{L}_0^2, \mathrm{cov})$ be the inner product space composed of all real-valued random variables with zero mean and finite variance, and endowed with the covariance operator as the inner product. Let both $x := y$ and $y =: x$ mean that x is defined by y. Write $a \propto b$ if $a = \kappa b$ for some constant $\kappa \in \mathbb{R}$. Denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For signal vectors $\{\boldsymbol{x}_k\}_{k=1}^2$, denote $\boldsymbol{\Sigma}_k = \mathrm{cov}(\boldsymbol{x}_k)$, $\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, $r_k = \mathrm{rank}(\boldsymbol{\Sigma}_k)$, $r_{\min} = r_1 \wedge r_2$, $r_{\max} = r_1 \vee r_2$, and $r_{12} = \mathrm{rank}(\boldsymbol{\Sigma}_{12})$. Throughout the paper, our asymptotic arguments are by default under $n \to \infty$.

2.1. Canonical correlation analysis

The CCA method [20] sequentially finds the most correlated variables, called *canonical variables*, between the two subspaces $\{\mathrm{span}(\boldsymbol{x}_k^\top)\}_{k=1}^2$ in $(\mathcal{L}_0^2, \mathrm{cov})$. For $1 \le \ell \le r_{12}$, the ℓ-th pair of canonical variables are defined as

$$\begin{aligned} \{z_{1\ell}, z_{2\ell}\} \in \underset{\{z_k\}_{k=1}^2}{\arg\max}\ & \mathrm{corr}(z_1, z_2) \quad \text{subject to} \\ & \mathrm{var}(z_k) = 1 \text{ and } z_k \in \mathrm{span}(\boldsymbol{x}_k^\top) \setminus \mathrm{span}(\{z_{km}\}_{m=1}^{\ell-1}), \end{aligned} \tag{3}$$

where $\mathrm{span}(\boldsymbol{x}_k^\top) \setminus \mathrm{span}(\{z_{km}\}_{m=1}^0) := \mathrm{span}(\boldsymbol{x}_k^\top)$, and for $\ell > 1$, $\mathrm{span}(\boldsymbol{x}_k^\top) \setminus \mathrm{span}(\{z_{km}\}_{m=1}^{\ell-1})$ denotes the orthogonal complement of $\mathrm{span}(\{z_{km}\}_{m=1}^{\ell-1})$ in $\mathrm{span}(\boldsymbol{x}_k^\top)$. The correlation $\rho_\ell := \mathrm{corr}(z_{1\ell}, z_{2\ell})$ is called the ℓ-th *canonical correlation* of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Augment $\{z_{k\ell}\}_{\ell=1}^{r_{12}}$ with any $(r_k - r_{12})$ standardized variables to be $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kr_k})^\top$ such that $\boldsymbol{z}_k^\top$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{x}_k^\top)$. We have the bi-orthogonality [45] that

$$\mathrm{cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathrm{diag}(\rho_1, \ldots, \rho_{r_{12}}, \mathbf{0}_{(r_1 - r_{12}) \times (r_2 - r_{12})}). \tag{4}$$

The augmented canonical variables $\{\boldsymbol{z}_k\}_{k=1}^2$ can be obtained by $\boldsymbol{z}_k = \mathbf{U}_{\theta k}^\top \boldsymbol{z}_k^*$, where $\boldsymbol{z}_k^* = \boldsymbol{\Lambda}_k^{-1/2} \mathbf{V}_k^\top \boldsymbol{x}_k$, $\boldsymbol{\Sigma}_k = \mathbf{V}_k \boldsymbol{\Lambda}_k \mathbf{V}_k^\top$ is a compact singular value decomposition (SVD) with $\boldsymbol{\Lambda}_k = \mathrm{diag}(\sigma_1(\boldsymbol{\Sigma}_k), \ldots, \sigma_{r_k}(\boldsymbol{\Sigma}_k))$, and $\boldsymbol{\Theta} := \mathrm{cov}(\boldsymbol{z}_1^*, \boldsymbol{z}_2^*) = \mathbf{U}_{\theta 1} \boldsymbol{\Lambda}_\theta \mathbf{U}_{\theta 2}^\top$ is a full SVD with $\boldsymbol{\Lambda}_\theta = \mathrm{diag}(\rho_1, \ldots, \rho_{r_{12}}, \mathbf{0}_{(r_1 - r_{12}) \times (r_2 - r_{12})})$. Note that in the inner product space $(\mathcal{L}_0^2, \mathrm{cov})$, $\cos\theta(\cdot, \cdot) = \mathrm{corr}(\cdot, \cdot)$ and $\|\cdot\| = \sqrt{\mathrm{var}(\cdot)}$.

A similar method to CCA is the principal angle analysis (PAA) [3], which investigates the closeness of any two subspaces, denoted by F and G, in the Euclidean dot product space $(\mathbb{R}^p, \cdot)$. For $1 \le \ell \le q := \min\{\dim(F), \dim(G)\}$, the ℓ-th *principal angle* $\theta_\ell \in [0, \pi/2]$ between F and G is defined by

$$\begin{aligned} &\cos\theta_\ell = \max_{\boldsymbol{u}\in F}\max_{\boldsymbol{v}\in G} \boldsymbol{u}^\top \boldsymbol{v} = \boldsymbol{u}_\ell^\top \boldsymbol{v}_\ell \quad \text{subject to} \\ &\|\boldsymbol{u}\|_F = \|\boldsymbol{v}\|_F = 1, \text{ and } \boldsymbol{u}^\top \boldsymbol{u}_j = \boldsymbol{v}^\top \boldsymbol{v}_j = 0 \text{ for } j = 1, \ldots, \ell - 1. \end{aligned} \tag{5}$$

The vectors $\{\boldsymbol{u}_\ell, \boldsymbol{v}_\ell\}$ are called the ℓ-th pair of *principal vectors* of F and G. Let $\mathbf{Q}_F$ and $\mathbf{Q}_G$ be the matrices whose columns form the orthonormal bases of F and G, respectively. The principal angles and principal vectors can be obtained by

$$\cos\theta_\ell = \sigma_\ell(\mathbf{Q}_F^\top \mathbf{Q}_G),\ (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q) = \mathbf{Q}_F \mathbf{U}_Q,\ (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q) = \mathbf{Q}_G \mathbf{V}_Q, \tag{6}$$

where $\mathbf{Q}_F^\top \mathbf{Q}_G = \mathbf{U}_Q \operatorname{diag}\{\sigma_1(\mathbf{Q}_F^\top \mathbf{Q}_G), \ldots, \sigma_q(\mathbf{Q}_F^\top \mathbf{Q}_G)\} \mathbf{V}_Q^\top$ is a SVD of $\mathbf{Q}_F^\top \mathbf{Q}_G$.

The PAA and CCA methods are essentially the same except their respective inner product spaces $(\mathbb{R}^p, \cdot)$ and $(\mathcal{L}_0^2, \mathrm{cov})$. The principal vectors and the cosines of principal angles of PAA correspond to the canonical variables and the canonical correlations of CCA. The cosines of principal angles are also called canonical correlations in PAA [3]. Similar to (4), the bi-orthogonality between different pairs of principal vectors also holds.

2.2. Decomposition-based canonical correlation analysis

For random vectors $\{\boldsymbol{x}_k\}_{k=1}^2$, the D-CCA method [45] aims to decompose each $\boldsymbol{x}_k$ into a common-source vector $\boldsymbol{c}_k$ and a distinctive-source vector $\boldsymbol{d}_k$ by

$$\boldsymbol{x}_k = \boldsymbol{c}_k + \boldsymbol{d}_k \tag{7}$$

subject to three desirable constraints in $(\mathcal{L}_0^2, \mathrm{cov})$:

$$\begin{cases} \operatorname{span}(\boldsymbol{c}_1^\top) = \operatorname{span}(\boldsymbol{c}_2^\top), \\ \operatorname{span}(\boldsymbol{d}_1^\top) \perp \operatorname{span}(\boldsymbol{d}_2^\top), \\ \operatorname{span}([\boldsymbol{x}_1; \boldsymbol{x}_2]^\top) = \operatorname{span}([\boldsymbol{c}_1; \boldsymbol{c}_2; \boldsymbol{d}_1; \boldsymbol{d}_2]^\top). \end{cases} \tag{8}$$

To this end, guided by the bi-orthogonality (4) of augmented canonical variables $\boldsymbol{z}_k^\top = [z_{k\ell}]_{\ell=1}^{r_k}$, $k = 1, 2$, D-CCA divides the decomposition problem (7) of $\operatorname{span}([\boldsymbol{x}_1; \boldsymbol{x}_2]^\top)$ into $r_{\max}$ subproblems, each within one of the mutually orthogonal subspaces $\left\{ \operatorname{span}(\{z_{k\ell}\}_{k=1}^2) \right\}_{\ell=1}^{r_{\max}}$ as

$$z_{k\ell} = c_\ell + d_{k\ell}, \tag{9}$$

where $z_{k\ell} = 0$ for $\ell > r_k$, and $c_\ell = 0$ for $\ell > r_{\min}$. For $\ell \le r_{\min}$, the common variable c_ℓ is defined by

$$c_\ell \propto \operatorname*{arg\,max}_{w \in (\mathcal{L}_0^2, \mathrm{cov})} \left\{ \cos^2\theta(z_{1\ell}, w) + \cos^2\theta(z_{2\ell}, w) \right\} \tag{10}$$

such that

$$\begin{cases} d_{1\ell} \perp d_{2\ell}, & (11) \\ \|c_\ell\| \text{ increases as } \theta_{z\ell} := \theta(z_{1\ell}, z_{2\ell}) \text{ decreases on } [0, \pi/2]. & (12) \end{cases}$$

Constraint (12) equivalently says that $\|c_\ell\|$ indicates the correlation strength of $z_{1\ell}$ and $z_{2\ell}$. The unique solution of (10) subject to (11) and (12) is

$$c_\ell = \left(1 - \sqrt{\frac{1-\cos\theta_{z\ell}}{1+\cos\theta_{z\ell}}}\right) \frac{z_{1\ell}+z_{2\ell}}{2} = \left[1 - \tan\left(\frac{\theta_{z\ell}}{2}\right)\right] \frac{z_{1\ell}+z_{2\ell}}{2}. \quad (13)$$

Figure 1 (a) geometrically illustrates the solution (13) with ℓ omitted in the subscriptions.

Combining the solutions of subproblems yields the D-CCA decomposition: for $k = 1, 2$,

$$\boldsymbol{x}_k = \sum_{\ell=1}^{r_k} \boldsymbol{\beta}_{k\ell} z_{k\ell} = \sum_{\ell=1}^{r_{12}} \boldsymbol{\beta}_{k\ell} c_\ell + \sum_{\ell=1}^{r_k} \boldsymbol{\beta}_{k\ell} d_{k\ell} =: \boldsymbol{c}_k + \boldsymbol{d}_k \quad (14)$$

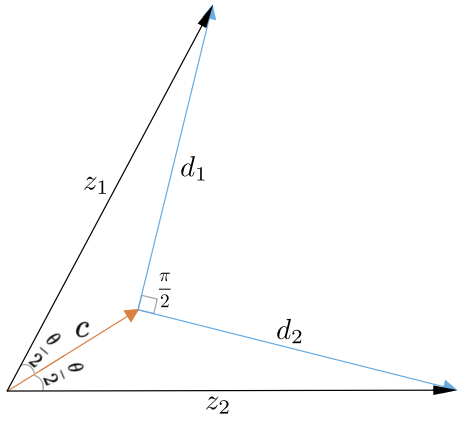
with $\boldsymbol{\beta}_{k\ell} = \mathrm{cov}(\boldsymbol{x}_k, z_{k\ell})$. Here, $\{c_\ell\}_{\ell=1}^{r_{12}}$ are the *common latent factors* of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\{d_{k\ell}\}_{\ell=1}^{r_k}$ are the *distinctive latent factors* of $\boldsymbol{x}_k$. Figure 1 (b) shows the decomposition structure of D-CCA.

3. Common and distinctive pattern analysis

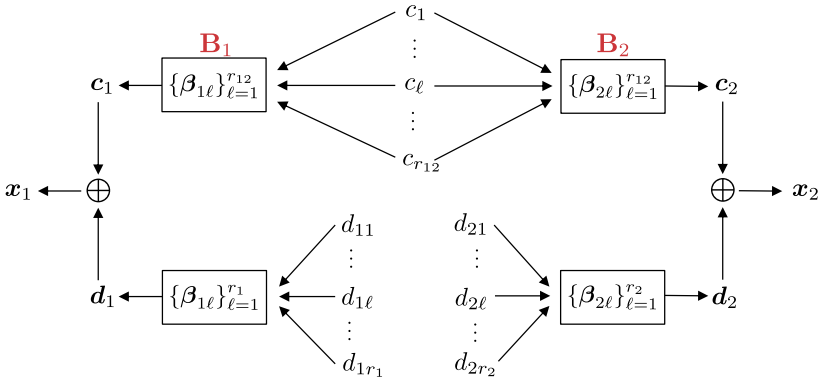
The CDPA method aims to more comprehensively define the common and distinctive patterns of two datasets by incorporating the common and distinctive patterns of the two coefficient matrices of common latent factors. We use a graph matching approach to match the unpaired rows between the coefficient matrices. Consistent estimators are established for the CDPA-defined common-pattern and distinctive-pattern matrices.

3.1. Common and distinctive patterns

As shown in Figure 1 (b), D-CCA only focuses on the common latent factors $\{c_\ell\}_{\ell=1}^{r_{12}}$ of $\{\boldsymbol{x}_k\}_{k=1}^2$, and ignores the common pattern of their coefficient matrices $\mathbf{B}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{kr_{12}})$ for $k = 1, 2$. So do its five previous methods mentioned in Section 1. In factor analysis [17], the coefficient matrix $\mathbf{B}_k$ of latent factors $\{c_\ell\}_{\ell=1}^{r_{12}}$ in $\boldsymbol{c}_k = \mathbf{B}_k([c_\ell]_{\ell=1}^{r_{12}})^\top$ is called their factor pattern matrix, and the entry $\mathbf{B}_k^{[i_k,\ell]}$ is the factor loading on c_ℓ for variable $\boldsymbol{c}_k^{[i_k]} = \sum_{\ell=1}^{r_{12}} \mathbf{B}_k^{[i_k,\ell]} c_\ell$, representing the contribution of c_ℓ in the linear combination of $\{c_\ell\}_{\ell=1}^{r_{12}}$ to forming $\boldsymbol{c}_k^{[i_k]}$. In signal processing, $\mathbf{B}_k$ is called a mixing channel [39, 41], which introduces correlations into the uncorrelated input sources $\{c_\ell\}_{\ell=1}^{r_{12}}$ to generate the output signal $\boldsymbol{c}_k$ that has $\mathrm{cov}(\boldsymbol{c}_k) = \mathbf{B}_k \,\mathrm{diag}(\mathrm{var}(c_1), \ldots, \mathrm{var}(c_{r_{12}}))\mathbf{B}_k^\top$. Thus, $\mathbf{B}_1$ and $\mathbf{B}_2$ may possess common and distinctive patterns of the respective ways in



(a) The geometry of D-CCA for two standardized random variables.



(b) The D-CCA decomposition for two signal random vectors.

FIG 1. *The D-CCA decomposition structure. In subfigure (a), the distinctive variables d_1 and d_2 are orthogonal (i.e., uncorrelated); the norm (i.e., standard deviation) of the common variable c indicates the correlation strength of the two standardized variables z_1 and z_2. In subfigure (b), D-CCA refers the common pattern of $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ to the common latent factors $\{c_\ell\}_{\ell=1}^{r_{12}}$, but ignores the common pattern of their coefficient matrices $\mathbf{B}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{kr_{12}})$ for $k = 1, 2$.*

which the common latent factors $\{c_\ell\}_{\ell=1}^{r_{12}}$ constitute $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. In CDPA, we define a common-pattern vector $\boldsymbol{c}$ for $\{\boldsymbol{x}_k\}_{k=1}^2$ which takes into account both the common latent sources $\{c_\ell\}_{\ell=1}^{r_{12}}$ and the common pattern of their mixing channels $\{\mathbf{B}_k\}_{k=1}^2$. The distinctive-pattern vector of signal $\boldsymbol{x}_k$ is then defined as the residual part of the signal after removing $\boldsymbol{c}$.

In the process $\boldsymbol{c}_k = \mathbf{B}_k([c_\ell]_{\ell=1}^{r_{12}})^\top = \sum_{\ell=1}^{r_{12}} \boldsymbol{\beta}_{k\ell} c_\ell$, the ℓ-th column $\boldsymbol{\beta}_{k\ell}$ of the mixing channel $\mathbf{B}_k$ is the sub-channel transmitting c_ℓ, and the linear mixture

of sub-channel outputs $\{\boldsymbol{\beta}_{k\ell}c_\ell\}_{\ell=1}^{r_{12}}$ reflects the "mixing" performance of the channel $\mathbf{B}_k$. We disentangle the common and distinctive latent structures for the two sub-channel spaces $\{\text{colsp}(\mathbf{B}_k)\}_{k=1}^2$ in a similar way as D-CCA does for the two signal spaces $\{\text{span}(\boldsymbol{x}_k^\top)\}_{k=1}^2$.

Two issues need to be solved before the analysis. First, the sub-channel vectors $\{\boldsymbol{\beta}_{k\ell}\}_{k\le 2,\ell\le r_{12}}$ may have unequal lengths p_1 and p_2, Without loss of generality, we let $p_1 \ge p_2$ throughout the paper. When $p_1 > p_2$, we zero-pad $\mathbf{B}_2$ to be a $p_1 \times r_{12}$ matrix $\mathbf{B}_{2A} = [\mathbf{B}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$. This zero padding is equivalent to adding $(p_1 - p_2)$ zeros into $\boldsymbol{x}_2$. In other words, we are now equivalently studying the patterns between $\boldsymbol{x}_1$ and $\boldsymbol{x}_{2A} = [\boldsymbol{x}_2; \mathbf{0}_{(p_1-p_2)\times 1}]$. Second, sometimes the rows between $\mathbf{B}_1$ and $\mathbf{B}_{2A}$ or equivalently the entries between $\boldsymbol{x}_1$ and $\boldsymbol{x}_{2A}$ are not one-to-one matched due to their arbitrary ordering. For this scenario, we match their rows by permuting the rows of $\mathbf{B}_{2A}$ with a permutation matrix $\mathbf{P}$. The permutation can be defined so that $\text{colsp}(\mathbf{B}_1)$ and $\text{colsp}(\mathbf{P}\mathbf{B}_{2A})$ are closest to each other by maximizing $\sum_{\ell=1}^{r_{12}} \cos^2\theta_{B\ell}$, where $\theta_{B\ell}$ is their ℓ-th principal angle. This row-matching procedure will be discussed in detail in Section 3.2. For the generalization of our results to other row-matching criteria, we assume that the permutation matrix $\mathbf{P}$ is prespecified in the following text. For notational simplicity, we use the superscript "$\diamond$" to indicate adding zero padding and $\mathbf{P}$ to the given vector or matrix if necessary, for example, $(\boldsymbol{x}_1^\diamond, \boldsymbol{x}_2^\diamond, \mathbf{B}_1^\diamond, \mathbf{B}_2^\diamond) = (\boldsymbol{x}_1, \mathbf{P}\boldsymbol{x}_{2A}, \mathbf{B}_1, \mathbf{P}\mathbf{B}_{2A})$.

We now consider the latent structure of the two sub-channel spaces $\{\text{colsp}(\mathbf{B}_k^\diamond)\}_{k=1}^2$ by using an analogy of D-CCA on $(\mathbb{R}^{p_1}, \cdot)$, where constraints (7)-(12) are translated for the columns of $\{\mathbf{B}_k^\diamond\}_{k=1}^2$ and CCA is replaced by PAA. Let $\theta_{B\ell}$ and $\{\boldsymbol{v}_{B_1\ell}, \boldsymbol{v}_{B_2\ell}\}$ be the ℓ-th principal angle and the ℓ-th pair of principal vectors of $\{\text{colsp}(\mathbf{B}_k^\diamond)\}_{k=1}^2$. There are r_{12} such pairs since $\mathbf{B}_k = \mathbf{V}_k\boldsymbol{\Lambda}_k^{1/2}\mathbf{U}_{\theta k}^{[:,1:r_{12}]}$ is a rank-r_{12} matrix. We define the common and distinctive components of $\{\boldsymbol{v}_{B_1\ell}, \boldsymbol{v}_{B_2\ell}\}$ using a decomposition similar to that in (9) and (13):

$$\boldsymbol{c}_{B\ell} = \left(1 - \sqrt{\frac{1-\cos\theta_{B\ell}}{1+\cos\theta_{B\ell}}}\right)\frac{(\boldsymbol{v}_{B_1\ell} + \boldsymbol{v}_{B_2\ell})}{2} \quad \text{and} \boldsymbol{d}_{B_k\ell} = \boldsymbol{v}_{B_k\ell} - \boldsymbol{c}_{B\ell} \tag{15}$$

for $k = 1, 2$ and $\ell = 1, \ldots, r_{12}$. Because the principal vectors $(\boldsymbol{v}_{B_k1}, \ldots, \boldsymbol{v}_{B_kr_{12}}) =: \mathbf{V}_{B_k}$ form an orthonormal basis of $\text{colsp}(\mathbf{B}_k^\diamond)$, the mixing-channel matrix can be written as

$$\mathbf{B}_k^\diamond = \mathbf{V}_{B_k}(\mathbf{V}_{B_k}^\top\mathbf{B}_k^\diamond) = \left([\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}} + [\boldsymbol{d}_{B_k\ell}]_{\ell=1}^{r_{12}}\right)(\mathbf{V}_{B_k}^\top\mathbf{B}_k^\diamond). \tag{16}$$

The part of $\boldsymbol{x}_k^\diamond$ that contains the common latent factors (source variables) $\{c_\ell\}_{\ell=1}^{r_{12}}$ and the common mixing-channel basis $\{\boldsymbol{c}_{B\ell}\}_{\ell=1}^{r_{12}}$ is

$$\boldsymbol{c}_k^* := [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}\mathbf{V}_{B_k}^\top\mathbf{B}_k^\diamond([c_\ell]_{\ell=1}^{r_{12}})^\top. \tag{17}$$

The difference between $\boldsymbol{c}_1^*$ and $\boldsymbol{c}_2^*$ is the matrices $\mathbf{S}_k := \mathbf{V}_{B_1}^\top\mathbf{B}_k^\diamond$, $k = 1, 2$ in the middle of their formulas, which contain the weights dually owned by

$\{\boldsymbol{c}_{B\ell}\}_{\ell=1}^{r_{12}}$ and $\{c_\ell\}_{\ell=1}^{r_{12}}$. We define the common part of the two dual weight matrices $\{\mathbf{S}_k\}_{k=1}^2$ as

$$\mathbf{S} = \underset{\mathbf{M}\in\mathbb{R}^{r_{12}\times r_{12}}}{\arg\min} \sum_{k=1}^{2} \left\| \mathbf{M} - [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\mathbf{S}_k \right\|_F^2 = \frac{1}{2}\sum_{k=1}^{2} [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\mathbf{S}_k. \tag{18}$$

To avoid overweighting a dataset when signals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have different scales, we weight $\mathbf{S}_k$ by the scale factor $[\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}$ in (18). This is equivalent to rescaling each $\boldsymbol{x}_k$ by the factor $[\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}$ at the very beginning as in [30]. Our definition of $\mathbf{S}$ in (18) is motivated by the *consensus configuration* in generalized procrustes analysis [15, 16] which minimizes the sum of squared Euclidean distances to transformed configurations of interest (i.e., the scaled $\{\mathbf{S}_k\}_{k=1}^2$ in our case). This minimization is equivalent to that of the sum of Kullback-Leibler divergences [35, Lemma 17.4.3] and yields a closed-form solution.

We combine the three types of common parts $\{\boldsymbol{c}_{B\ell}\}_{\ell=1}^{r_{12}}$, $\{c_\ell\}_{\ell=1}^{r_{12}}$ and $\mathbf{S}$ to define the *common-pattern vector* of the scaled signals $\boldsymbol{x}_k^S := [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\boldsymbol{x}_k^\diamond$, $k = 1, 2$ as

$$\boldsymbol{c} = \mathbf{B}_c([c_\ell]_{\ell=1}^{r_{12}})^\top = [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}\mathbf{S}([c_\ell]_{\ell=1}^{r_{12}})^\top = \frac{1}{2}\sum_{k=1}^{2}[\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\boldsymbol{c}_k^*, \tag{19}$$

where $\mathbf{B}_c = [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}\mathbf{S}$ is defined as the common pattern of $\mathbf{B}_1^\diamond$ and $\mathbf{B}_2^\diamond$.

For each individiual unscaled signal vector $\boldsymbol{x}_k^\diamond$, we rescale $\boldsymbol{c}$ to be $\boldsymbol{c}^{(k)} = [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{1/2}\boldsymbol{c}$ and express the CDPA decomposition as

$$\boldsymbol{x}_k^\diamond = \boldsymbol{c}_k^\diamond + \boldsymbol{d}_k^\diamond =: (\boldsymbol{c}^{(k)} + \boldsymbol{h}_k) + \boldsymbol{d}_k^\diamond = \boldsymbol{c}^{(k)} + (\boldsymbol{h}_k + \boldsymbol{d}_k^\diamond) =: \boldsymbol{c}^{(k)} + \boldsymbol{\delta}_k. \tag{20}$$

For signal vector $\boldsymbol{x}_k^\diamond$, the vector $\boldsymbol{h}_k$ represents the distinctive pattern retained within the common-source vector $\boldsymbol{c}_k^\diamond$, and the vector $\boldsymbol{\delta}_k$ characterizes the *total* distinctive pattern by incorporating both $\boldsymbol{h}_k$ and the distinctive-source vector $\boldsymbol{d}_k^\diamond$. We denote $\{\mathbf{C}, \mathbf{C}^{(k)}, \mathbf{H}_k, \boldsymbol{\Delta}_k\}$ to be the corresponding sample matrices of $\{\boldsymbol{c}, \boldsymbol{c}^{(k)}, \boldsymbol{h}_k, \boldsymbol{\delta}_k\}$ associated with $\mathbf{X}_k$.

Definition 1. *We define the common-pattern vector of $\{\boldsymbol{x}_1^\diamond, \boldsymbol{x}_2^\diamond\}$ (more precisely, $\{\boldsymbol{x}_1^S, \boldsymbol{x}_2^S\}$) as the vector $\boldsymbol{c}$ given in (19), and the scaled common-pattern vector for $\boldsymbol{x}_k^\diamond$ as $\boldsymbol{c}^{(k)} = [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{1/2}\boldsymbol{c}$. The distinctive-pattern vector of $\boldsymbol{x}_k^\diamond$ is $\boldsymbol{\delta}_k = \boldsymbol{x}_k^\diamond - \boldsymbol{c}^{(k)}$. As the sample matrices of $\boldsymbol{c}$, $\{\boldsymbol{c}^{(k)}\}_{k=1}^2$ and $\{\boldsymbol{\delta}_k\}_{k=1}^2$, matrices $\mathbf{C}$, $\{\mathbf{C}^{(k)}\}_{k=1}^2$ and $\{\boldsymbol{\Delta}_k\}_{k=1}^2$ are called the common-pattern, the scaled common-pattern, and distinctive-pattern matrices of $\{\mathbf{X}_k\}_{k=1}^2$, respectively.*

The population CDPA decomposition is summarized in Algorithm 1 and its uniqueness is given in Theorem 1.

Theorem 1. *Given any $p_1 \times p_1$ permutation matrix $\mathbf{P}$, the common-pattern vector $\boldsymbol{c}$ defined in (19) for $(\boldsymbol{x}_1^\diamond, \boldsymbol{x}_2^\diamond) = (\boldsymbol{x}_1, \mathbf{P}\boldsymbol{x}_{2A})$ is unique, regardless of the non-unique choices of canonical variables $\{z_{1\ell}, z_{2\ell}\}_{\ell=1}^{r_{12}}$ and principal vectors $\{\boldsymbol{v}_{B_1\ell}, \boldsymbol{v}_{B_2\ell}\}_{\ell=1}^{r_{12}}$.*

Algorithm 1 Population CDPA

Input: Signal vectors $\boldsymbol{x}_k \in \mathbb{R}^{p_k}, k = 1, 2$

1: Obtain common latent factors $[c_\ell]_{\ell=1}^{r_{12}}$ and coefficient matrix $\mathbf{B}_k$ by the D-CCA in (14);
2: Add zero rows to $\mathbf{B}_1$ or $\mathbf{B}_2$ if dimensions $p_1 \neq p_2$;
3: Match the rows of $\mathbf{B}_1$ and $\mathbf{B}_2$ by the graph-matching based approach (Section 3.2), if the variables in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not paired;
4: Compute the common mixing-channel basis $\{\boldsymbol{c}_{B\ell}\}_{\ell=1}^{r_{12}}$ by (15) for $\{\text{colsp}(\mathbf{B}_k)\}_{k=1}^2$;
5: Compute the common dual-weight matrix $\mathbf{S}$ by (18);
6: Obtain the common-pattern vector of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as $\boldsymbol{c} = [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}} \mathbf{S} ([c_\ell]_{\ell=1}^{r_{12}})^\top$;
7: Rescale $\boldsymbol{c}$ to be the scaled common-pattern vector $\boldsymbol{c}^{(k)} = [\text{tr}(\boldsymbol{\Sigma}_k)]^{1/2} \boldsymbol{c}$;
8: Obtain the distinctive-pattern vector of $\boldsymbol{x}_k$ as $\boldsymbol{\delta}_k = \boldsymbol{x}_k - \boldsymbol{c}^{(k)}$;

Output: Common-pattern vector $\boldsymbol{c}$, scaled common-pattern vectors $\{\boldsymbol{c}^{(k)}\}_{k=1}^2$, distinctive-pattern vectors $\{\boldsymbol{\delta}_k\}_{k=1}^2$

Remark 1. *Since $\boldsymbol{c}$ is the common-pattern vector of the scaled signal vectors $\boldsymbol{x}_1^S$ and $\boldsymbol{x}_2^S$, $\text{tr}\{\text{cov}(\boldsymbol{c})\} = \frac{\text{tr}\{\text{cov}(\boldsymbol{c})\}}{\frac{1}{2}\sum_{k=1}^2 \text{tr}\{\text{cov}(\boldsymbol{x}_k^S)\}}$ represents the proportion of the average variance of $\boldsymbol{x}_1^S$ and $\boldsymbol{x}_2^S$ explained by $\boldsymbol{c}$, which reflects the similarity strength of the two signal vectors.*

Remark 2. *The common-pattern vector $\boldsymbol{c}$ differs only in its sign for $\{\boldsymbol{x}_1, \mathbf{P}\boldsymbol{x}_{2A}\}$ and $\{-\boldsymbol{x}_1, -\mathbf{P}\boldsymbol{x}_{2A}\}$, but is usually quite different for $\{\boldsymbol{x}_1, \mathbf{P}\boldsymbol{x}_{2A}\}$ and $\{\boldsymbol{x}_1, -\mathbf{P}\boldsymbol{x}_{2A}\}$. We assume the sign of each entry in $\boldsymbol{y}_k$ or $\boldsymbol{x}_k$ cannot be arbitrarily changed, but the sign of $\boldsymbol{y}_k$ or equivalently that of $\boldsymbol{x}_k$ may change. The assumption is generally true if each dataset represents a data type. For example, let $\boldsymbol{y}_2$ be mRNA expression data and its entry $\boldsymbol{y}_2^{[i]}$ measure the mRNA expression level on the i-th gene. The arbitrary entry-wise sign changes can result in two different measurements applied to $\boldsymbol{y}_2$. Regarding the different $\boldsymbol{c}$'s due to the sign change (if allowed) of entirely $\boldsymbol{y}_2$ or $\boldsymbol{x}_2$, we suggest to choose the one with larger variance $\text{tr}\{\text{cov}(\boldsymbol{c})\}$ or, in practice, larger $\frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2 = \text{tr}(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top)$, where $\widehat{\mathbf{C}}$ is the estimate of $\mathbf{C}$ that will be introduced in Section 3.3. It will be shown later in Theorem 2 that $\frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2 \xrightarrow{P} \text{tr}\{\text{cov}(\boldsymbol{c})\}$ under mild conditions. The confidence interval (CI) of $\frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2$ can be constructed by bootstrapping samples [11] once the ranks $\{r_1, r_2, r_{12}\}$ and the permutation matrix $\mathbf{P}$ are determined.*

3.2. Row matching of coefficient matrices

When the rows of coefficient matrices $\mathbf{B}_1$ and $\mathbf{B}_{2A}$ are not one-to-one matched, we match them by permuting the rows of $\mathbf{B}_{2A}$ with the following permutation matrix

$$\mathbf{P}_* = \underset{\mathbf{P} \in \Pi_{p_1}}{\arg\max} \sum_{\ell=1}^{r_{12}} \cos^2 \theta_{B\ell}, \tag{21}$$

where $\theta_{B\ell}$ is the ℓ-th principal angle of $\text{colsp}(\mathbf{B}_1)$ and $\text{colsp}(\mathbf{P}\mathbf{B}_{2A})$, and Π_{p_1} is the set of all $p_1 \times p_1$ permutation matrices. This optimization is equivalent to minimizing the Frobenius distance $(2\sum_{\ell=1}^{r_{12}} \sin^2 \theta_{B\ell})^{1/2}$. Commonly-used distances between vector spaces [9] also include the geodesic distance

$(\sum_{\ell=1}^{r_{12}} \theta_{B\ell}^2)^{1/2}$, the Martin distance $(\log \prod_{\ell=1}^{r_{12}} \cos^{-2} \theta_{B\ell})^{1/2}$, the Asimov distance θ_{B1}, and the gap distance $\sin \theta_{B1}$. The four alternative distances appear more difficult to be minimized, but our criterion based on the Frobenius distance can be converted to the famous graph matching problem [32].

Specifically, by equations in (6), the optimization problem in (21) is equivalent to

$$\begin{aligned} \mathbf{P}_* &= \underset{\mathbf{P}\in\Pi_{p_1}}{\arg\max}\, \mathrm{tr}\left(\mathbf{Q}_1^\top \mathbf{P}\mathbf{Q}_{2A}(\mathbf{Q}_1^\top \mathbf{P}\mathbf{Q}_{2A})^\top\right) \\ &= \underset{\mathbf{P}\in\Pi_{p_1}}{\arg\max}\, \mathrm{tr}\left(\mathbf{Q}_1\mathbf{Q}_1^\top \mathbf{P}\mathbf{Q}_{2A}\mathbf{Q}_{2A}^\top \mathbf{P}^\top\right), \end{aligned} \tag{22}$$

where $\mathbf{Q}_k \in \mathbb{R}^{p_k \times r_{12}}$ is a matrix whose columns are an orthonormal basis of $\mathrm{colsp}(\mathbf{B}_k)$, which can be the r_{12} left singular vectors of $\mathbf{B}_k$, and $\mathbf{Q}_{2A} = [\mathbf{Q}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$ whose columns are still an orthonormal basis of $\mathbf{B}_{2A}$. Let $\mathbf{M}_1 = \mathbf{Q}_1\mathbf{Q}_1^\top$ and $\mathbf{M}_2 = \mathbf{Q}_{2A}\mathbf{Q}_{2A}^\top$. For $k = 1, 2$, let $\mathbf{M}_k^+$ be the matrix obtained by all elements of $\mathbf{M}_k$ minus the smallest element of $[\mathbf{M}_1, \mathbf{M}_2]$. For any $p_1 \times p_1$ matrix $\mathbf{M}$, denote $\overline{\mathrm{diag}}(\mathbf{M})$ to be the $p_1 \times p_1$ matrix having the same off-diagonal part of $\mathbf{M}$ but with zero diagonal, and $\mathrm{vdg}(\mathbf{M})$ to be the vector consisting of the diagonal elements of $\mathbf{M}$. We have

$$\begin{aligned} &\max_{\mathbf{P}\in\Pi_{p_1}} \mathrm{tr}\left(\mathbf{M}_1\mathbf{P}\mathbf{M}_2\mathbf{P}^\top\right) \\ &\Leftrightarrow \min_{\mathbf{P}\in\Pi_{p_1}} \left\|\mathbf{M}_1 - \mathbf{P}\mathbf{M}_2\mathbf{P}^\top\right\|_F^2 \\ &\Leftrightarrow \min_{\mathbf{P}\in\Pi_{p_1}} \left\|\mathbf{M}_1^+ - \mathbf{P}\mathbf{M}_2^+\mathbf{P}^\top\right\|_F^2 \\ &\Leftrightarrow \min_{\mathbf{P}\in\Pi_{p_1}} \left\{\left\|\overline{\mathrm{diag}}(\mathbf{M}_1^+) - \mathbf{P}\overline{\mathrm{diag}}(\mathbf{M}_2^+)\mathbf{P}^\top\right\|_F^2 + \left\|\mathrm{vdg}(\mathbf{M}_1^+) - \mathbf{P}\,\mathrm{vdg}(\mathbf{M}_2^+)\right\|_F^2\right\} \\ &\Leftrightarrow \max_{\mathbf{P}\in\Pi_{p_1}} \left\{\mathrm{tr}\left(\mathbf{P}^\top\overline{\mathrm{diag}}(\mathbf{M}_1^+)\mathbf{P}\overline{\mathrm{diag}}(\mathbf{M}_2^+)\right) + \mathrm{tr}\left(\mathbf{P}^\top\,\mathrm{vdg}(\mathbf{M}_1^+)[\mathrm{vdg}(\mathbf{M}_2^+)]^\top\right)\right\}, \end{aligned} \tag{23}$$

where the last objective function is the formula (4) of [32] for the graph matching problem. Graph matching is known to be NP-hard for the optimal solution. We use the doubly stochastic projected fixed-point (DSPFP) algorithm of [32] to obtain an efficient approximation of $\mathbf{P}_*$, which has time complexity only $O(p_1^3)$ per iteration and space complexity $O(p_1^2)$. For ultra-large p_1, one may further apply the approximation procedure of [37] that employs a clustering method before DSPFP.

3.3. Estimation

Often in practice, the data matrices $\{\mathbf{Y}_k\}_{k=1}^2$ are high-dimensional and are the only observable data in decomposition (1). The literature of (1) regularly assumes high-dimensional $\{\mathbf{Y}_k\}_{k=1}^2$ to be "low-rank plus noise". Indeed, big data matrices are often approximately low-rank in many real-world applications [50],

so their low-rank approximations provide feasible or more efficient computation and meanwhile preserve the major information [25]. Moreover, the low-rank plus noise structure can circumvent the curse of dimensionality [56, 27] in recovering the common-source and distinctive-source matrices $\{\mathbf{C}_k, \mathbf{D}_k\}_{k=1}^2$ from which our defined common-pattern and distinctive-pattern matrices are derived. Following the D-CCA paper [45], we consider the low-rank plus noise structure:

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{B}_{f_k}\mathbf{F}_k + \mathbf{E}_k, \tag{24}$$

$$\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \mathbf{B}_{f_k}\boldsymbol{f}_k + \boldsymbol{e}_k, \tag{25}$$

where $\mathbf{B}_{f_k} \in \mathbb{R}^{p_k \times r_k}$ is a real deterministic matrix, the columns of $\mathbf{F}_k$ and $\mathbf{E}_k$ are respectively the n i.i.d. copies of mean-zero random vectors $\boldsymbol{f}_k$ and $\boldsymbol{e}_k$, the columns of $\mathbf{F} = [\mathbf{F}_1; \mathbf{F}_2]$ are also statistically independent, and the vector $\boldsymbol{f}_k \in \mathbb{R}^{r_k}$ contains r_k latent factors such that $\mathrm{cov}(\boldsymbol{f}_k) = \mathbf{I}_{r_k \times r_k}$, $\mathrm{cov}(\boldsymbol{f}_k, \boldsymbol{e}_k) = \mathbf{0}_{r_k \times p_k}$, and $\mathrm{span}(\boldsymbol{f}_k^\top)$ is a fixed subspace in $(\mathcal{L}_0^2, \mathrm{cov})$ that is independent of $\{n, p_1, p_2\}$. Hence, r_1, r_2 and r_{12} are fixed numbers. We can choose $\boldsymbol{f}_k^\top$ to be the augmented canonical variables $\boldsymbol{z}_k^\top$. The covariance matrix $\mathrm{cov}(\boldsymbol{y}_k) = \mathbf{B}_{f_k}\mathbf{B}_{f_k}^\top + \mathrm{cov}(\boldsymbol{e}_k)$ is assumed to be a spiked covariance matrix for which the largest r_k eigenvalues are significantly larger than the rest, namely, signals are distinguishably stronger than noises.

Before recovering our common-pattern and distinctive-pattern matrices, we introduce the D-CCA's estimators of $\mathbf{X}_k$ and $\mathbf{C}_k$. For simplicity, we write all estimators with true matrix ranks $\{r_1, r_2, r_{12}\}$. In practice, as implemented in D-CCA, ranks $\{r_k\}_{k=1}^2$ and r_{12} can be well selected by the edge distribution (ED) method of [38] and the minimum description length information-theoretic criterion (MDL-IC) of [48], respectively; see Appendix A2. The estimator of $\mathbf{X}_k$ is defined by using the soft-thresholding method of [53] as

$$\widehat{\mathbf{X}}_k = \mathbf{U}_{k1}\,\mathrm{diag}(\hat{\sigma}_1^S(\mathbf{Y}_k), \ldots, \hat{\sigma}_{r_k}^S(\mathbf{Y}_k))\mathbf{U}_{k2}^\top, \tag{26}$$

where $\mathbf{U}_{k1}\,\mathrm{diag}(\sigma_1(\mathbf{Y}_k), \ldots, \sigma_{r_k}(\mathbf{Y}_k))\mathbf{U}_{k2}^\top$ forms the top-r_k SVD of $\mathbf{Y}_k$, and the soft-thresholded singular value $\hat{\sigma}_\ell^S(\mathbf{Y}_k) = \sqrt{\max\{\sigma_\ell^2(\mathbf{Y}_k) - \tau_k p_k, 0\}}$ with $\tau_k = \sum_{\ell=r_k+1}^{p_k} \sigma_\ell^2(\mathbf{Y}_k)/(np_k - nr_k - p_k r_k)$. Then from $\widehat{\mathbf{X}}_k$, define the estimator of $\boldsymbol{\Sigma}_k$ by $\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n}\widehat{\mathbf{X}}_k\widehat{\mathbf{X}}_k^\top$, and denote its SVD by $\widehat{\boldsymbol{\Sigma}}_k = \widehat{\mathbf{V}}_k\widehat{\boldsymbol{\Lambda}}_k\widehat{\mathbf{V}}_k^\top$, where $\widehat{\mathbf{V}}_k \in \mathbb{R}^{p_k \times r_k}$ has orthonormal columns and $\widehat{\boldsymbol{\Lambda}}_k = \mathrm{diag}(\sigma_1(\widehat{\boldsymbol{\Sigma}}_k), \ldots, \sigma_{r_k}(\widehat{\boldsymbol{\Sigma}}_k))$. Following Section 2.1, let $\widehat{\mathbf{Z}}_k^* = (\widehat{\boldsymbol{\Lambda}}_k^\dagger)^{1/2}\widehat{\mathbf{V}}_k^\top\widehat{\mathbf{X}}_k$ and $\widehat{\boldsymbol{\Theta}} = \frac{1}{n}\widehat{\mathbf{Z}}_1^*(\widehat{\mathbf{Z}}_2^*)^\top$, and write the latter's full SVD by $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{U}}_{\theta 1}\widehat{\boldsymbol{\Lambda}}_\theta\widehat{\mathbf{U}}_{\theta 2}^\top$ with $\widehat{\boldsymbol{\Lambda}}_\theta = \mathrm{diag}(\sigma_1(\widehat{\boldsymbol{\Theta}}), \ldots, \sigma_{\hat{r}_\theta}(\widehat{\boldsymbol{\Theta}}), \mathbf{0}_{(r_1-\hat{r}_\theta)\times(r_2-\hat{r}_\theta)})$ and $\hat{r}_\theta = \mathrm{rank}(\widehat{\boldsymbol{\Theta}})$. Define the estimated sample matrix of $\boldsymbol{z}_k$ by $\widehat{\mathbf{Z}}_k = \widehat{\mathbf{U}}_{\theta k}^\top\widehat{\mathbf{Z}}_k^*$. Let $\widehat{\mathbf{A}}_C = \mathrm{diag}(\hat{a}_1, \ldots, \hat{a}_{r_{12}})$, where $\hat{a}_\ell = \frac{1}{2}\big[1 - (\frac{1-\sigma_\ell(\widehat{\boldsymbol{\Theta}})}{1+\sigma_\ell(\widehat{\boldsymbol{\Theta}})})^{1/2}\big]$ for $\ell \le \hat{r}_\theta$, and otherwise $\hat{a}_\ell = 0$. The estimators of $\mathbf{C}_k$ and $\mathbf{D}_k$ are defined by

$$\widehat{\mathbf{C}}_k = \frac{1}{n}\widehat{\mathbf{X}}_k(\widehat{\mathbf{Z}}_k^{[1:r_{12},:]})^\top\widehat{\mathbf{A}}_C\sum_{j=1}^2\widehat{\mathbf{Z}}_j^{[1:r_{12},:]} = \widehat{\mathbf{B}}_k\widehat{\mathbf{C}}_0 \quad \text{and} \quad \widehat{\mathbf{D}}_k = \widehat{\mathbf{X}}_k - \widehat{\mathbf{C}}_k.$$

where $\widehat{\mathbf{B}}_k = \frac{1}{n}\widehat{\mathbf{X}}_k(\widehat{\mathbf{Z}}_k^{[1:r_{12},:]})^\top = \widehat{\mathbf{V}}_k\widehat{\boldsymbol{\Lambda}}_k^{1/2}\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}$ similar to $\mathbf{B}_k = \mathbf{V}_k\boldsymbol{\Lambda}_k^{1/2}\mathbf{U}_{\theta k}^{[:,1:r_{12}]}$, and $\widehat{\mathbf{C}}_0 = \widehat{\mathbf{A}}_C\sum_{j=1}^2\widehat{\mathbf{Z}}_j^{[1:r_{12},:]}$ is the estimated sample matrix of $(c_1, \ldots, c_{r_{12}})^\top$.

We now derive the estimators of our common-pattern and distinctive-pattern matrices. Let $\widehat{\mathbf{B}}_{2A} = [\widehat{\mathbf{B}}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$, $\widehat{\mathbf{Q}}_k \in \mathbb{R}^{p_k \times r_{12}}$ be the left singular matrix of $\widehat{\mathbf{B}}_k$, $\widehat{\mathbf{Q}}_{2A} = [\widehat{\mathbf{Q}}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$, and $\widehat{\mathbf{\Theta}}_B = \widehat{\mathbf{Q}}_1^\top \mathbf{P} \widehat{\mathbf{Q}}_{2A}$. Recall that we assume the permutation matrix $\mathbf{P}$ is prespecified. If the row matching of $\mathbf{B}_1$ and $\mathbf{B}_{2A}$ is necessary, one may choose $\mathbf{P}$ to be the matrix $\mathbf{P}_*$ in the NP-hard problem (22), approximated by the DSPFP method with data samples. Note that $\mathbf{P}_*$, as a permutation matrix, is either obtained exactly or approximated with at least two wrong entries. To ease theoretical analysis without such misspecification, we assume that $\mathbf{P}$ is well determined. Write the full SVD of $\widehat{\mathbf{\Theta}}_B$ by $\widehat{\mathbf{\Theta}}_B = \widehat{\mathbf{U}}_{B_1} \widehat{\mathbf{\Lambda}}_B \widehat{\mathbf{U}}_{B_2}^\top$, where $\widehat{\mathbf{\Lambda}}_B$ has nonincreasing diagonal elements, and define $\widehat{\mathbf{V}}_{B_1} = \widehat{\mathbf{Q}}_1 \widehat{\mathbf{U}}_{B_1}$ and $\widehat{\mathbf{V}}_{B_2} = \mathbf{P}\widehat{\mathbf{Q}}_{2A} \widehat{\mathbf{U}}_{B_2}$. It follows from (6) that the diagonal elements of $\widehat{\mathbf{\Lambda}}_B$ and the columns of $\{\widehat{\mathbf{V}}_{B_k}\}_{k=1}^2$ are respectively the cosines of principal angles and the principal vectors of $\mathrm{colsp}(\widehat{\mathbf{B}}_1)$ and $\mathrm{colsp}(\mathbf{P}\widehat{\mathbf{B}}_{2A})$. Substituting them for their true counterparts in (15) yields our estimator $\hat{\boldsymbol{c}}_{B\ell}$ for $\boldsymbol{c}_{B\ell}$. Then from (19), we define the estimator of $\mathbf{C}$ by

$$\widehat{\mathbf{C}} = \frac{1}{2}[\hat{c}_{B\ell}]_{\ell=1}^{r_{12}} \Big(\widehat{\mathbf{V}}_{B_1}^\top \widehat{\mathbf{B}}_1 [\mathrm{tr}(\widehat{\mathbf{\Sigma}}_1)]^{-1/2} + \widehat{\mathbf{V}}_{B_2}^\top \mathbf{P} \widehat{\mathbf{B}}_{2A} [\mathrm{tr}(\widehat{\mathbf{\Sigma}}_2)]^{-1/2} \Big) \widehat{\mathbf{C}}_0, \tag{27}$$

where $[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2} = [\mathrm{tr}(\frac{1}{n}\widehat{\mathbf{X}}_k \widehat{\mathbf{X}}_k^\top)]^{1/2} = \frac{1}{\sqrt{n}} \|\widehat{\mathbf{X}}_k\|_F$ estimates $[\mathrm{tr}(\mathbf{\Sigma}_k)]^{1/2}$. The estimator of the scaled version $\mathbf{C}^{(k)}$ is defined by

$$\widehat{\mathbf{C}}^{(k)} = [\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2} \widehat{\mathbf{C}}.$$

Given $\{r_1, r_2, r_{12}, \mathbf{P}\}$, the computational complexity of obtaining $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{C}}^{(k)}$ is $O(np_1^2 \wedge n^2 p_1)$ majorly due to the SVD of $\{\mathbf{Y}_k\}_{k=1}^2$. We define the estimators $\widehat{\mathbf{H}}_k = \widehat{\mathbf{C}}_k^\diamond - \widehat{\mathbf{C}}^{(k)}$ and $\widehat{\mathbf{\Delta}}_k = \widehat{\mathbf{H}}_k + \widehat{\mathbf{D}}_k^\diamond$ for $\mathbf{H}_k$ and $\mathbf{\Delta}_k$, respectively.

The estimation approach for the CDPA decomposition is summarized in Algorithm 2.

Algorithm 2 CDPA estimation

Input: Observed datasets $\mathbf{Y}_k \in \mathbb{R}^{p_k \times n}$, $k = 1, 2$

1: Select ranks $r_k = \mathrm{rank}(\mathbf{\Sigma}_k)$ and $r_{12} = \mathrm{rank}(\mathbf{\Sigma}_{12})$, respectively, by the ED method and the MDL-IC method (Appendix A2).

2: Obtain the denoised data $\widehat{\mathbf{X}}_k$ by the soft thresholding in (26).

3: Obtain coefficient matrix estimates $\{\widehat{\mathbf{B}}_k\}_{k=1}^2$ and the sample matrix $\widehat{\mathbf{C}}_0$ of common latent factors by the sample D-CCA (Section 3.3).

4: If necessary, zero-pad and/or row-match $\{\widehat{\mathbf{B}}_k\}_{k=1}^2$ by the graph-matching based approach (Section 3.2).

5: Compute the common-pattern matrix estimate $\widehat{\mathbf{C}}$ by (27).

6: Obtain the scaled common-pattern matrix estimate $\widehat{\mathbf{C}}^{(k)} = [\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2}\widehat{\mathbf{C}}$ with $\widehat{\mathbf{\Sigma}}_k = \widehat{\mathbf{X}}_k \widehat{\mathbf{X}}_k^\top / n$, and the distinctive-pattern matrix estimate $\widehat{\mathbf{\Delta}}_k = \widehat{\mathbf{X}} - \widehat{\mathbf{C}}^{(k)}$.

Output: Common-pattern matrix estimate $\widehat{\mathbf{C}}$, scaled common-pattern matrix estimates $\{\widehat{\mathbf{C}}^{(k)}\}_{k=1}^2$, distinctive-pattern matrix estimates $\{\widehat{\mathbf{\Delta}}_k\}_{k=1}^2$

The following assumption given in [53, 45], which guarantees the consistency of $\{\widehat{\mathbf{X}}_k\}_{k=1}^2$, is also used to derive our asymptotic results. Readers are referred to [53, 45] for detailed discussions on this assumption.

Assumption 1. *We assume the following conditions for model given in* (24) *and* (25).

(I) *Let* $\lambda_{k,1} > \cdots > \lambda_{k,r_k} > \lambda_{k,r_k+1} \geq \cdots \geq \lambda_{k,p_k} > 0$ *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* $\ell \leq r_k$, $\lambda_{k,1}/\lambda_{k,r_k}$ *is bounded from above and below, and* $\sqrt{p_k}(\log n)^{1/\gamma_{k2}} = o(\lambda_{k,r_k})$ *with* γ_{k2} *given in (V).*

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

(IV) *The matrix* $\mathbf{B}_{f_k}^\top \mathbf{B}_{f_k}$ *is a diagonal matrix. For all* $i \leq p_k$ *and* $\ell \leq r_k$, $|\mathbf{B}_{f_k}^{[i,\ell]}| \leq \kappa_B \sqrt{\lambda_{k,\ell}/p_k}$ *with a constant* $\kappa_B > 1$.

(V) *Denote* $\boldsymbol{e}_k = (e_{k1}, \ldots, e_{kp_k})^\top$ *and* $\boldsymbol{f}_k = (f_{k1}, \ldots, f_{kr_k})^\top$. *Assume that* $\|\mathrm{cov}(\boldsymbol{e}_k)\|_\infty < s_0$ *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* $t > 0$, $P(|e_{ki}| > t) \leq \exp(-(t/b_{k1})^{\gamma_{k1}})$ *and* $P(|f_{k\ell}| > t) \leq \exp(-(t/b_{k2})^{\gamma_{k2}})$.

Theorem 2. *Suppose that Assumption 1 and* $r_{12} \geq 1$ *hold. Assume that any distinct values in* $\{\cos\theta_{B\ell}\}_{\ell=1}^{r_{12}} \cup \{0, -\infty\}$ *are separated by at least a positive constant. Define*

$$\delta_\theta = \left(\frac{1}{\sqrt{n}} + \sum_{k=1}^{2} \sqrt{\frac{\log p_k}{n\,\mathrm{SNR}_k}} \right) \wedge 1,$$

where $\mathrm{SNR}_k = \frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_k)\}}{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{e}_k)\}}$ *is the signal-to-noise ratio of* $\boldsymbol{y}_k$. *For* $k = 1, 2$, *we have that*

$$\frac{\|\widehat{\mathbf{C}} - \mathbf{C}\|_\star^2}{\frac{1}{2}(\|\mathbf{X}_1^S\|_\star^2 + \|\mathbf{X}_2^S\|_\star^2)} \vee \frac{\|\widehat{\mathbf{C}}^{(k)} - \mathbf{C}^{(k)}\|_\star^2}{\|\mathbf{X}_k\|_\star^2} = O_P(\delta_\theta),$$

and

$$\left| \mathrm{tr}\left(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top\right) - \mathrm{tr}\{\mathrm{cov}(\boldsymbol{c})\} \right| = O_P(\delta_\theta^{1/2}),$$

where $\|\cdot\|_\star$ *denotes either the Frobenius norm or the spectral norm, and* $\mathbf{X}_k^S = [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\mathbf{X}_k$.

Remark 3. *From Theorem 3 and Corollary 1 of [45], we have* $\frac{\|\widehat{\mathbf{M}}_k - \mathbf{M}_k\|_\star^2}{\|\mathbf{X}_k\|_\star^2} = O_P(\delta_\theta)$ *for* $\mathbf{M}_k \in \{\mathbf{X}_k, \mathbf{C}_k, \mathbf{D}_k\}$. *Additionally by our Theorem 2 and the triangle inequality of norms, we also have this error bound for* $\mathbf{M}_k \in \{\mathbf{H}_k, \boldsymbol{\Delta}_k\}$. *Note that the scaled squared error in the Frobenius norm indicates the scaled loss in matrix variation (sum of squares).*

Theorem 3. *Let* $\widehat{\mathbf{P}}_* = \arg\max_{\mathbf{P}\in\Pi_{p_1}} \operatorname{tr}(\widehat{\mathbf{Q}}_1^\top \mathbf{P}\widehat{\mathbf{Q}}_{2A}(\widehat{\mathbf{Q}}_1^\top \mathbf{P}\widehat{\mathbf{Q}}_{2A})^\top)$. *Suppose that Assumption 1 and* $r_{12} \geq 1$ *hold. Then, we have*

$$\left|\operatorname{tr}\left(\mathbf{Q}_1^\top \widehat{\mathbf{P}}_* \mathbf{Q}_{2A}(\mathbf{Q}_1^\top \widehat{\mathbf{P}}_* \mathbf{Q}_{2A})^\top\right) - \operatorname{tr}\left(\mathbf{Q}_1^\top \mathbf{P}_* \mathbf{Q}_{2A}(\mathbf{Q}_1^\top \mathbf{P}_* \mathbf{Q}_{2A})^\top\right)\right| = O_P(\delta_\theta).$$

For the row matching problem of $\mathbf{B}_1$ and $\mathbf{B}_{2A}$, Theorem 3 provides an asymptotically vanishing bound on the change in the objective function value of (22) when the optimal solution $\mathbf{P}_*$ is replaced by $\widehat{\mathbf{P}}_*$.

4. Simulation studies

In this section, we evaluate the finite-sample performance of the proposed CDPA estimation via simulations, comparing with the six D-CCA-type methods mentioned in Section 1.

4.1. Simulation setups

We consider the following two simulation setups for signals $\{\boldsymbol{x}_k\}_{k=1}^2$.
Setup 1: Let variable dimensions $p_1 = p_2$, ranks $r_1 = r_2 = 5$, and eigenvalues $\lambda_\ell(\boldsymbol{\Sigma}_k) = 500 - 100(\ell - 1)$ for $\ell \leq 5$. The signals are $\boldsymbol{x}_k = \mathbf{V}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{z}_k$ for $k = 1, 2$, where canonical variables $[\boldsymbol{z}_1; \boldsymbol{z}_2]$ follow a multivariate Gaussian distribution with $\boldsymbol{z}_k \sim \mathcal{N}(\mathbf{0}_{r_k\times 1}, \mathbf{I}_{r_k\times r_k})$ and $\operatorname{cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \operatorname{diag}\big\{\cos(\theta \wedge 30°), \cos(\theta \wedge 60°), \cos\theta, \cos(\theta + 15°), \cos((\theta + 30°) \wedge 90°)\big\}$. Let $\mathbf{Q}_k = \mathbf{V}_k^{[:,1:r_{12}]}$, $\mathbf{P} = \mathbf{I}_{(p_1\vee p_2)\times(p_1\vee p_2)}$, and $(\mathbf{Q}_1^0)^\top \mathbf{Q}_2^0 = \operatorname{cov}(\boldsymbol{z}_1, \boldsymbol{z}_2)^{[1:r_{12},1:r_{12}]}$ of which the diagonal contains the cosines of principal angles of $\operatorname{colsp}(\mathbf{B}_1^0)$ and $\operatorname{colsp}(\mathbf{B}_2^0)$, where $\mathbf{M}_k^0 = [\mathbf{M}_k; \mathbf{0}_{(p_k - p_1\wedge p_2)\times r_{12}}]$ with $\mathbf{M} \in \{\mathbf{Q}, \mathbf{B}\}$. Matrices $\{\mathbf{V}_k\}_{k=1}^2$ are randomly generated under the above constraints and are fixed for all simulation replications.
Setup 2: We vary p_1 but fix $p_2 = 900$. The other settings are the same as in Setup 1. This setup aims to evaluate the performance of considered methods when $p_1 \neq p_2$.

We generate noises $\{e_{ki}\}_{k\leq 2, i\leq p_k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2)$ independent of signals $\{\boldsymbol{x}_k\}_{k=1}^2$. Simulations are conducted with sample size $n = 300$, variable dimension p_1 ranging from 100 to 1500, angle θ from 0° to 75°, noise variance σ_e^2 from 0.25 to 9, and 1000 replications under each setting. The proportion of average variance of $\boldsymbol{x}_1^S$ and $\boldsymbol{x}_2^S$ explained by $\boldsymbol{c}$, that is, $\operatorname{tr}\{\operatorname{cov}(\boldsymbol{c})\}$, has values 0.890, 0.479, 0.213, 0.126, 0.092 and 0.088 corresponding to θ from 0° to 75° by a step 15°. This pattern of the explained proportion of variance persists across all chosen values of p_1.

4.2. Finite-sample performance of CDPA estimators

We numerically evaluate the finite-sample performance of proposed CDPA estimators by comparing to the asymptotic results given in Section 3.3. Since the

signal-to-noise ratio $\mathrm{SNR}_k = 1500/(p_k\sigma_e^2)$ in the above simulation setups, for simplicity we examine the trend of estimation errors with respect to (p_k, σ_e^2) instead of (p_k, SNR_k) in the theorems. We use the true $\{r_k\}_{k=1}^2$, r_{12}, and $\mathbf{P}$ in our matrix estimation here to exclude the error induced by their misspecification. The ranks $\{r_k\}_{k=1}^2$ and r_{12} can be well selected by the ED and MDL-IC methods, respectively, as shown in [45]. The selection of $\mathbf{P}$ by the DSPFP-based row-matching method in Section 3.2 is evaluated later in this subsection.

We first investigate the performance of our common-pattern matrix estimator $\widehat{\mathbf{C}}$ defined in (27). The first two rows of Figures 2 and 3 summarize the scaled squared errors of $\widehat{\mathbf{C}}$ as studied in Theorem 2 and also its relative squared errors under Setup 1 with $\theta \in \{15°, 75°\}$. The squared errors in the Frobenius norm represent the scaled or relative losses in matrix variation (sum of squares). The average estimation errors increase as the dimension p_1 or the noise variance σ_e^2 grows, and are even well controlled under 0.1 for many cases with large $p_1 \geq 900$ and large $\sigma_e^2 \geq 4$ (or $\mathrm{SNR}_k \leq 0.42$). These results are consistent with the influence of p_1 and σ_e^2 $(= 1500/(p_k \mathrm{SNR}_k)$, here) on the convergence rates given in Theorem 2. Similar numerical results are observed for the scaled version $\widehat{\mathbf{C}}^{(k)} = [\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2}\widehat{\mathbf{C}}$ and the distinctive-pattern matrix estimator $\widehat{\mathbf{\Delta}}_k = \widehat{\mathbf{X}}_k - \widehat{\mathbf{C}}^{(k)}$ for $k \in \{1, 2\}$, and hence are omitted for brevity.

As a similarity indicator of signals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the common-pattern explained proportion of signal variance, $\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c})\}$, is estimated by $\mathrm{tr}(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top) = \frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2$. The third rows of Figures 2 and 3 plot the average absolute error and the average relative error of this estimator for Setup 1 with $\theta \in \{15°, 75°\}$. Same with Theorem 2, the row shows that the average estimation errors grow with increasing p_1 or σ_e^2 and have a larger magnitude than those squared errors of $\widehat{\mathbf{C}}$ as shown in the first two rows of the figure. The errors are controlled below 0.1 even for some cases with large $p_1 \geq 900$ or $\sigma_e^2 \geq 4$.

For the row-matching approach of coefficient matrices $\{\mathbf{B}_k\}_{k=1}^2$ described in Section 3.2, its theoretical performance stated in Theorem 3 is numerically investigated with the intractable $\mathbf{P}_*$ and $\widehat{\mathbf{P}}_*$ being replaced by their DSPFP approximations denoted as $\mathbf{P}_a$ and $\widehat{\mathbf{P}}_a$. The fourth rows of Figures 2 and 3 display the average absolute and relative errors of $\mathrm{tr}\{(\mathbf{Q}_1^0)^\top\widehat{\mathbf{P}}_a\mathbf{Q}_2^0[(\mathbf{Q}_1^0)^\top\widehat{\mathbf{P}}_a\mathbf{Q}_2^0]^\top\}$ for Setup 1 with $\theta \in \{15°, 75°\}$. Although its absolute error seems to have larger values than that of its oracle version (with $\widehat{\mathbf{P}}_*$) expected in Theorem 3, its relative error is controlled under or around 0.1 even for some cases with large $p_1 \geq 900$ or $\sigma_e^2 \geq 4$, and moreover, the two types of errors both follow the influence of p_1 and σ_e^2 $(= 1500/(p_k \mathrm{SNR}_k)$, here) on the convergence rate shown in the theorem.

The above result patterns also generally hold for settings with more different values of θ (or equivalently $\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c})\}$) and for those under Setup 2 where $p_1 \neq p_2$, which are provided in Appendix A3.

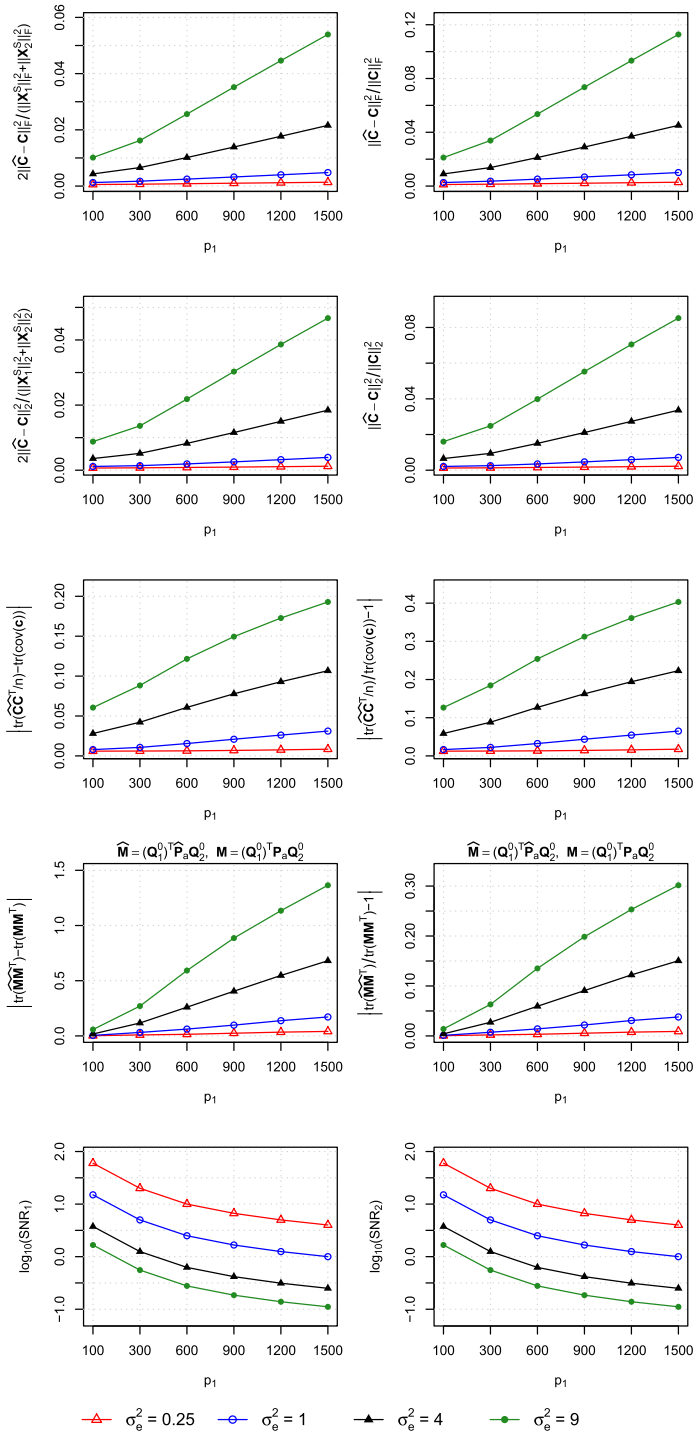


FIG 2. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 1 with* $\theta = 15°$.

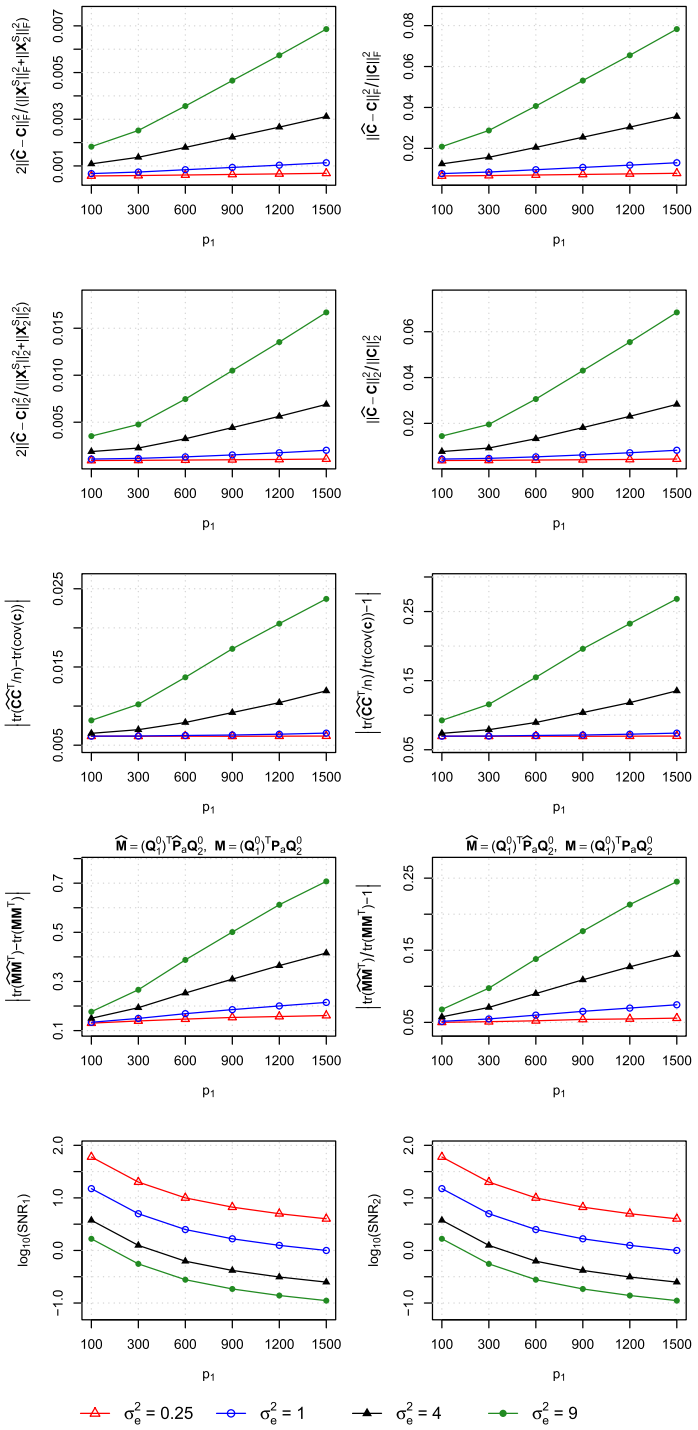


FIG 3. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 1 with* $\theta = 75°$.

TABLE 1

Averages (and standard deviations) of the estimates for the first principal angle θ_{B1} and its cosine $\cos\theta_{B1}$ of $\mathrm{colsp}(\mathbf{B}_1^0)$ and $\mathrm{colsp}(\mathbf{B}_2^0)$ obtained from existing methods over 1000 simulation replications with $(\theta, p_1, \sigma_e^2) = (75°, 300, 1)$.

Method	Setup 1	Setup 2
D-CCA	30.6°(0.374°)/0.860(0.003)	30.8°(0.339°)/0.859(0.003)
OnPLS	31.8°(0.931°)/0.850(0.009)	32.0°(0.881°)/0.848(0.008)
COBE	NA	NA
JIVE	32.0°(0.998°)/0.848(0.009)	32.9°(1.210°)/0.840(0.012)
AJIVE	NA	NA
DISCO-SCA	31.1°(0.573°)/0.856(0.005)	31.3°(0.545°)/0.855(0.005)

Note: NA means the result is not available due to zero common-source matrix estimates.

4.3. Performance of related methods

We now investigate the numerical performance of the six competing methods, including D-CCA, OnPLS, COBE, JIVE, AJIVE, and DISCO-SCA. Unlike our CDPA, all the six D-CCA-type methods are developed without taking into account the common and distinctive patterns between the two coefficient matrices $\{\mathbf{B}_k\}_{k=1}^2$ of their common latent factors.[2]

The simulation here aims to corroborate the existence of both common and distinctive patterns in their coefficient matrices $\{\mathbf{B}_k\}_{k=1}^2$. The existence can be shown if $\mathrm{colsp}(\mathbf{B}_1^0)$ and $\mathrm{colsp}(\mathbf{B}_2^0)$ are neither overlapping nor orthogonal, that is, their first principal angle $\theta_{B1} \notin \{0°, 90°\}$ or equivalently their first canonical correlation $\cos\theta_{B1} \notin \{1, 0\}$. Since the six D-CCA-type methods have different definitions of $\{\boldsymbol{c}_k\}_{k=1}^2$ for decomposition (2) due to their different constraints, they may have different $\{\mathbf{B}_k^0\}_{k=1}^2$ and thus different values of θ_{B1} under our simulation setups. The ground-truth θ_{B1} is $\theta \wedge 30°$ for D-CCA in our simulation, but may not be easy to theoretically determine for the other five methods and is thus estimated by simulated data.

Table 1 summarizes the first principal angle θ_{B1} and its cosine $\cos\theta_{B1}$ of $\mathrm{colsp}(\mathbf{B}_1^0)$ and $\mathrm{colsp}(\mathbf{B}_2^0)$ estimated by the six methods under the two simulation setups with $(\theta, p_1, \sigma_e^2) = (75°, 300, 1)$. We see that the COBE and AJIVE give zero common-source matrix estimates and thus fail to discover any common pattern of the two correlated signal datasets. The average estimates of θ_{B1} and $\cos\theta_{B1}$ from the other four methods are all close to 30° and 0.866, respectively, with very small standard deviations. Therefore, there is significant statistical evidence that their $\theta_{B1} \notin \{0°, 90°\}$ and $\cos\theta_{B1} \notin \{1, 0\}$. This indicates the non-negligible existence of both the common and the distinctive patterns between their coefficient matrices, but these patterns are unfortunately not considered by these D-CCA-type methods.

[2] We implement the OnPLS with the post-processing step in footnote 1 to obtain the coefficient matrices $\{\mathbf{B}_k\}_{k=1}^2$ of its common latent factors.

5. Real data analysis

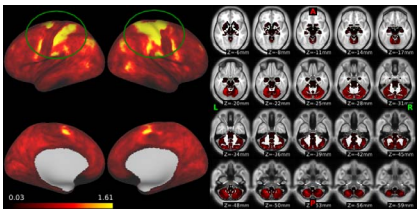
We apply our CDPA to two real-world data examples, respectively, from the HCP and TCGA, comparing with the six D-CCA-type methods mentioned in Section 1. We focus on the comparison with the state-of-the-art D-CCA in this section, and present the results of the other five methods in Appendix A4.

5.1. Application to HCP motor-task functional MRI data

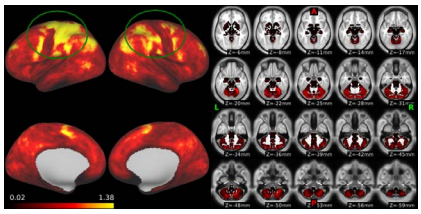
We consider the HCP motor-task functional MRI data obtained from 1080 healthy young adults [2]. All participants were asked by visual cues to perform five motor tasks during the image scanning, including tapping left and right fingers, squeezing left and right toes, and moving tongue. From the acquired brain images, for every participant and each task, the HCP computed a z-statistic map of the task's contrast against the fixation baseline at 91,282 grayordinates including 59,412 cortical surface vertices and 31,870 subcortical gray matter voxels. The z-statistic maps of all participants for each individual task constitute a 91,282×1080 data matrix. We focus on the left-hand and right-hand tasks, and apply the proposed CDPA to discover their common pattern on the brain, with comparison to the D-CCA method.

Each of the two observed data matrices is row-centered by subtracting the average within each row. Since all z-statistic maps of the two motor tasks are obtained from the same measurement and at the same set of grayordinates, there is no need to choose the signs or match the rows of the two data matrices. We consider the variance maps of $\{\boldsymbol{x}_L, \boldsymbol{x}_R, \boldsymbol{c}_L, \boldsymbol{c}_R, \boldsymbol{c}\}$ on the brain, which are estimated by the sample variances computed from the sample matrix estimates $\{\widehat{\mathbf{X}}_L, \widehat{\mathbf{X}}_R, \widehat{\mathbf{C}}_L, \widehat{\mathbf{C}}_R, \widehat{\mathbf{C}}\}$ obtained by D-CCA and CDPA. Here, the subscripts L and R denote the left-hand and right-hand tasks. The ranks $\{r_L, r_R\}$ and r_{12} are all selected as two by the ED and MDL-IC methods, respectively. The proportions of corresponding signal variances explained by common-source vectors $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ are $\frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c}_L)\}}{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_L)\}} \approx \frac{\|\widehat{\mathbf{C}}_L\|_F^2}{\|\widehat{\mathbf{X}}_L\|_F^2} = 0.113$ and $\frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c}_R)\}}{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{x}_R)\}} \approx \frac{\|\widehat{\mathbf{C}}_R\|_F^2}{\|\widehat{\mathbf{X}}_R\|_F^2} = 0.111$. The common-pattern explained proportion of signal variance is $\mathrm{tr}\{\mathrm{cov}(\boldsymbol{c})\} \approx \frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2 = 0.077$.

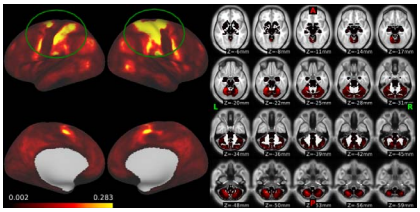
Figure 4 presents the estimated variance maps of D-CCA and CDPA. For all the five maps, the estimated variances of cortical surface vertices overall dominate those of subcortical voxels. We hence focus on the part of each variance map for the cortical surface. From the estimated signal variance maps $\widehat{\mathrm{var}}(\boldsymbol{x}_L)$ and $\widehat{\mathrm{var}}(\boldsymbol{x}_R)$ shown in Figure 4 (a) and (b), we see that the right half brain is more active, with larger variances, on the cortical surface for the left-hand task, while the pattern is almost opposite for the right-hand task. In particular, the contralateral pattern is clearly seen on the somatomotor cortex annotated by green circles, a brain region known to be linked with hand tasks [4]. A similar contralateral pattern is also observed for D-CCA's $\widehat{\mathrm{var}}(\boldsymbol{c}_L)$ and $\widehat{\mathrm{var}}(\boldsymbol{c}_R)$ in Figure 4 (c) and (d). This indicates that the $\boldsymbol{c}_k$ vector of D-CCA retains



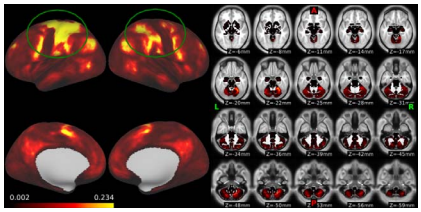
(a) $\widehat{\mathrm{var}}(\boldsymbol{x}_L)$ of D-CCA



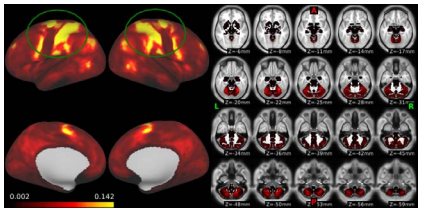
(b) $\widehat{\mathrm{var}}(\boldsymbol{x}_R)$ of D-CCA



(c) $\widehat{\mathrm{var}}(\boldsymbol{c}_L)$ of D-CCA



(d) $\widehat{\mathrm{var}}(\boldsymbol{c}_R)$ of D-CCA



(e) $\widehat{\mathrm{var}}(\boldsymbol{c}) \cdot [\mathrm{tr}^{\frac{1}{2}}\{\widehat{\mathrm{cov}}(\boldsymbol{x}_L)\} + \mathrm{tr}^{\frac{1}{2}}\{\widehat{\mathrm{cov}}(\boldsymbol{x}_R)\}]^2/4$ of CDPA

FIG 4. *The variance maps estimated by the D-CCA and CDPA methods for HCP motor-task functional MRI data. The notations* $\widehat{\mathrm{var}}$ *and* $\widehat{\mathrm{cov}}$ *denote the sample variance vector and sample covariance matrix obtained from the corresponding recovered sample matrix. 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.*

some distinctive pattern of $\boldsymbol{x}_k$ for $k \in \{L, R\}$. It is not surprising because $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ have different coefficient matrices of the common latent factors, which are r_{12} columns in the coefficient matrices of canonical variables for $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$, respectively, as shown in equation (14). In contrast, our CDPA's common-pattern vector $\boldsymbol{c}$ in Figure 4 (e) has an estimated variance map that is nearly symmetric on the two hemispheres, and thus is more reasonable than D-CCA's common-source vectors $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ to represent the common pattern of the left-hand and right-hand tasks on the brain.

5.2. Application to TCGA breast cancer genomic datasets

With the aim to discover new breast cancer subtypes, we apply the proposed CDPA to two TCGA breast cancer genomic datasets [26], and compare the results with the D-CCA. We consider the DNA methylation data and mRNA expression data obtained from a common set of 703 tumor samples. Following the preprocessing procedure of [29], we select the top 1100 variable probes for the DNA methylation dataset and the top 896 variably expressed genes for the mRNA expression dataset. The tumor samples are categorized by the classic PAM50 model [40] into four intrinsic subtypes, including 124 Basal-like, 58 HER2-enriched, 348 Luminal A, and 173 Luminal B tumors.

The two data matrices of interest have sizes 1100×703 and 896×703, and are row-centered before analysis. The ranks $(r_{\text{DNA}}, r_{\text{mRNA}}, r_{12})$ are selected by the ED and MDL-IC methods as $(3, 2, 2)$. From the D-CCA, the proportions of signal variances explained by common-source vectors $\boldsymbol{c}_{\text{DNA}}$ and $\boldsymbol{c}_{\text{mRNA}}$ are $\frac{\text{tr}\{\text{cov}(\boldsymbol{c}_{\text{DNA}})\}}{\text{tr}\{\text{cov}(\boldsymbol{x}_{\text{DNA}})\}} \approx \frac{\|\widehat{\mathbf{C}}_{\text{DNA}}\|_F^2}{\|\widehat{\mathbf{X}}_{\text{DNA}}\|_F^2} = 0.210$ and $\frac{\text{tr}\{\text{cov}(\boldsymbol{c}_{\text{mRNA}})\}}{\text{tr}\{\text{cov}(\boldsymbol{x}_{\text{mRNA}})\}} \approx \frac{\|\widehat{\mathbf{C}}_{\text{mRNA}}\|_F^2}{\|\widehat{\mathbf{X}}_{\text{mRNA}}\|_F^2} = 0.422$, indicating different influences of the common latent factors on the two signal datasets. Thus, by ignoring these different common-source influences, their $\boldsymbol{c}_{\text{DNA}}$ and $\boldsymbol{c}_{\text{mRNA}}$ are not appropriate to be viewed as the common pattern of $\boldsymbol{x}_{\text{DNA}}$ and $\boldsymbol{x}_{\text{mRNA}}$.

Since only 126 (11.5%) DNA methylation probes can be mapped to the genes of the considered mRNA expression data, for simplicity we match the rows of the two data matrices by using the graph-matching based approach described in Section 3.2 before implementing CDPA. The CDPA method shows that the common-pattern explained proportion of signal variance $\text{tr}\{\text{cov}(\boldsymbol{c})\} \approx \frac{1}{n}\|\widehat{\mathbf{C}}\|_F^2$ is 0.161 (95% CI = [0.154, 0.185]) for $\boldsymbol{x}_{\text{DNA}}$ and $\boldsymbol{x}_{\text{mRNA}}$, but is only 0.049 (95% CI = [0.046, 0.057]) for $\boldsymbol{x}_{\text{DNA}}$ and $-\boldsymbol{x}_{\text{mRNA}}$, where each 95% CI is computed by 5000 bootstrapping samples. We hence focus on the common and distinctive patterns extracted from $\{\boldsymbol{x}_{\text{DNA}}, \boldsymbol{x}_{\text{mRNA}}\}$ rather than $\{\boldsymbol{x}_{\text{DNA}}, -\boldsymbol{x}_{\text{mRNA}}\}$.

We explore new cancer subtypes by conducting clustering analysis on each observed or recovered matrix from the CDPA and D-CCA methods. We use the Ward's hierarchical clustering method [54] with the Euclidean distance, and simply specify the number of clusters to be four, which is the same number of the PAM50 intrinsic subtypes.

Table 2 compares the differences in survival curves of identified clusters or given subtypes using two most popular methods, the log-rank test [34] and the Peto-Peto's Wilcoxon test [42], where the latter test is more sensitive to early survival differences. Our CDPA's $\widehat{\boldsymbol{\Delta}}_{\text{mRNA}}$-identified clusters and the PAM50 intrinsic subtypes both have very significantly distinct survival behaviors with the two smallest p-values ≤ 0.009 in both tests, while the other matrices generate much less pronounced clusters, in particular, the matrices $\{\widehat{\mathbf{C}}_k, \widehat{\mathbf{D}}_k\}_{k\in\{\text{DNA,mRNA}\}}$ of D-CCA all have large p-values ≥ 0.290. By comparing the p-values of $\widehat{\mathbf{C}}$, $\widehat{\mathbf{X}}_k$ and $\widehat{\boldsymbol{\Delta}}_k$ for each k, the improved discriminative power of distinctive-pattern matrix estimate $\widehat{\boldsymbol{\Delta}}_k$ can be attributed to removing the less sensitive common-pattern matrix estimate $\widehat{\mathbf{C}}$ from the denoised data matrix $\widehat{\mathbf{X}}_k$. The adjusted

TABLE 2

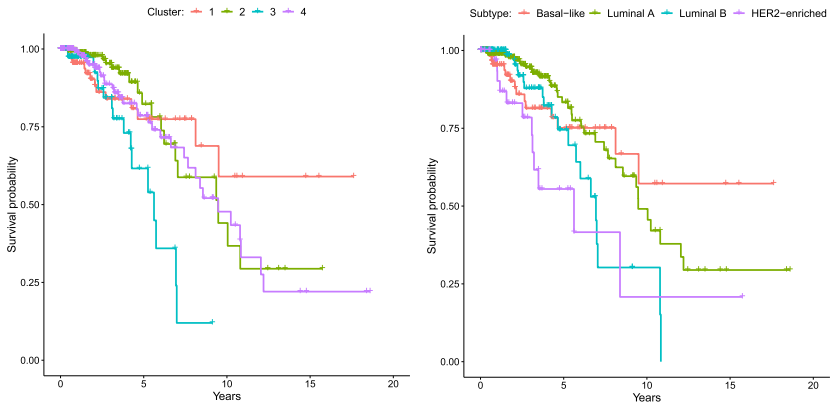
Log-rank test and Peto-Peto's Wilcoxon test for survival curve differences among the clusters identified from each observed or recovered matrix from the CDPA and D-CCA methods for TCGA breast cancer datasets.

Data	Log-rank/Peto's p-values	Data	Log-rank/Peto's p-values	Data	Log-rank/Peto's p-values
$\mathbf{Y}_{\text{DNA}}$	0.175/0.230	$\mathbf{Y}_{\text{mRNA}}$	0.251/0.299	$[\mathbf{Y}^N_{\text{DNA}}; \mathbf{Y}^N_{\text{mRNA}}]$	0.245/0.129
$\widehat{\mathbf{X}}_{\text{DNA}}$	**0.077/0.112**	$\widehat{\mathbf{X}}_{\text{mRNA}}$	**0.063/0.061**	$[\widehat{\mathbf{X}}^N_{\text{DNA}}; \widehat{\mathbf{X}}^N_{\text{mRNA}}]$	0.565/0.619
$\widehat{\mathbf{C}}_{\text{DNA}}$	0.820/0.979	$\widehat{\mathbf{C}}_{\text{mRNA}}$	0.619/0.704	$[\widehat{\mathbf{C}}^N_{\text{DNA}}; \widehat{\mathbf{C}}^N_{\text{mRNA}}]$	0.752/0.751
$\widehat{\mathbf{D}}_{\text{DNA}}$	0.515/0.417	$\widehat{\mathbf{D}}_{\text{mRNA}}$	0.290/0.354	$[\widehat{\mathbf{D}}^N_{\text{DNA}}; \widehat{\mathbf{D}}^N_{\text{mRNA}}]$	0.149/0.223
$\widehat{\mathbf{H}}_{\text{DNA}}$	0.430/0.502	$\widehat{\mathbf{H}}_{\text{mRNA}}$	0.330/0.409	$[\widehat{\mathbf{H}}^N_{\text{DNA}}; \widehat{\mathbf{H}}^N_{\text{mRNA}}]$	0.337/0.369
$\widehat{\mathbf{\Delta}}_{\text{DNA}}$	**0.058/0.075**	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$	**0.004/0.009**	$[\widehat{\mathbf{\Delta}}^N_{\text{DNA}}; \widehat{\mathbf{\Delta}}^N_{\text{mRNA}}]$	0.218/0.208
$\widehat{\mathbf{C}}$	**0.106/0.163**	PAM50	**0.003/0.001**		

Note: Denote $\mathbf{M}^N = \mathbf{M}/\|\mathbf{M}\|_F$ for any matrix $\mathbf{M}$.

Rand index [22] between our $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-identified clusters and PAM50 subtypes is 0.343 (95% CI = [0.335, 0.352]), indicating a poor agreement. It is evident that, built on top of D-CCA, our CDPA can benefit data mining with additional pattern matrices $\left\{\widehat{\mathbf{C}}, \{\widehat{\mathbf{\Delta}}_k, \widehat{\mathbf{H}}_k\}_{k\in\{\text{DNA,mRNA}\}}\right\}$.

Let $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-$i$ denote the i-th cluster identified from $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$. Figure 5 displays the Kaplan-Meier survival curves of $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-identified clusters and PAM50 subtypes. With the worst survival curve among the four identified clusters, $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 behaves similar to the HER2-enriched subtype, but is notably different with all other identified clusters and intrinsic subtypes. This is further confirmed in Table 3 by the minimum p-value of corresponding log-rank test and Peto-Peto's Wilcoxon test. Also seen in the table, the other three $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-identified clusters have no significant survival differences with large p-values ≥ 0.320. Moreover, the matching matrix in Table 4 shows that most of $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-1 and $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-2 samples belong to the Basal-like and Luminal A subtypes, respectively. Hence, the other three $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-identified clusters are less of interest



(a) CDPA's $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-identified clusters (b) PAM50 subtypes

FIG 5. *Kaplan-Meier survival curves of TCGA breast cancer clusters and subtypes.*

TABLE 3
Log-rank test and Peto-Peto's Wilcoxon test for survival curve differences among CDPA's $\widehat{\mathbf{\Delta}}_{mRNA}$-identified clusters and PAM50 subtypes for TCGA breast cancer data.

Comparison	Log-rank/Peto's p-values	Comparison	Log-rank/Peto's p-values
$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-1 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-2	0.895/0.550	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-1 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3	0.022/0.070
$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-1 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-4	0.491/0.816	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-2 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3	3.34e-4/5.35e-4
$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-2 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-4	0.375/0.320	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 vs. $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-4	0.006/0.013
$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 vs. Basal-like	0.041/0.121	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 vs. Luminal A	5.89e-5/1.26e-4
$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 vs. Luminal B	0.069/0.070	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 vs. HER2-enriched	0.585/0.361

TABLE 4
Matching matrix and clinical features of CDPA's $\widehat{\mathbf{\Delta}}_{mRNA}$-identified clusters and PAM50 subtypes for TCGA breast cancer data analysis.

PAM50	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-1	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-2	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3	$\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-4	Total	ER+/−	PR+/−	HER2+/−
Basal-like	122	0	0	2	124	6%/81%	6%/79%	7%/54%
Luminal A	0	194	31	123	348	89%/1%	82%/8%	9%/53%
Luminal B	0	13	77	83	173	87%/2%	72%/17%	16%/46%
HER2-enriched	8	8	**10**	32	58	**33%/52%**	**17%/71%**	**62%/16%**
Total	130	215	118	240	703			
ER+/−	6%/80%	91%/3%	**80%/4%**	79%/10%				
PR+/−	5%/79%	84%/10%	**64%/20%**	68%/20%				
HER2+/−	8%/52%	10%/57%	**19%/42%**	21%/41%				

Notes: The columns of the matching matrix are well reordered such that its diagonal sum is maximized. Receptor status for estrogen (ER), progesterone (PR) and human epidermal growth factor 2 (HER2) includes positive (+), negative (−), and N/A or equivocal.

to be new subtypes, and we focus on $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 which has the poorest survival, and further compare it with the HER2-enriched subtype. From Table 4, we see that the $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 cluster (118 samples) and the HER2-enriched subtype (58 samples) share only 10 samples and have substantially distinct clinical features in terms of the three important receptors' status. In particular, the $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 cluster primarily includes those samples that are ER+ and/or PR+, whereas the HER2-enriched subtype contains those that are HER2+ and/or PR−. To conclude, the $\widehat{\mathbf{\Delta}}_{\text{mRNA}}$-3 cluster, with a low survival rate, is remarkably different from the four PAM50 subtypes and appears to be an important new breast cancer subtype worth further investigation.

6. Discussion

In this paper, we propose a new decomposition method, called CDPA, to extract the common and distinctive patterns of two correlated datasets by incorporating the conventionally ignored common and distinctive patterns between the two coefficient matrices of common latent factors. We also develop a graph-matching based approach to match the unpaired rows between the coefficient matrices. Consistent CDPA matrix estimation is established under high-dimensional settings and is supported by simulations. Our simulation studies and two real-data examples show that CDPA can better delineate the common and distinctive patterns between datasets than D-CCA-type methods, thereby benefiting data mining applications.

There are two possible extensions of the CDPA. The first is to extend it to

three or more datasets. One may construct a multi-set CDPA method by first developing a multi-set D-CCA from the generalized CCA [24]. The next challenge is how to appropriately match the rows of the multiple coefficient matrices of the resulting common latent factors. The second extension is to incorporate the nonlinear patterns between the two datasets. The CDPA only considers the linear patterns extracted from the inner product spaces $(\mathcal{L}_0^2, \mathrm{cov})$ and $(\mathbb{R}^{p_1 \vee p_2}, \cdot)$. A nonlinear version of our row-matching approach and a nonlinear D-CCA may be expected in this extension, where the latter is possibly developed from the kernel CCA [14] or the deep CCA [1].

Appendix A1: Theoretical proofs

A1.1. Proof of Theorem 1

For $k = 1, 2$, denote $\boldsymbol{z}_k^{[1:r_{12}]}$ and $\tilde{\boldsymbol{z}}_k^{[1:r_{12}]}$ to be the vectors containing two different sets of the first r_{12} canonical variables associated with $\boldsymbol{x}_k$. By the first paragraph of page 5 in the supplement of [45], there exists an orthogonal matrix $\mathbf{O}_{zk}$ such that $\tilde{\boldsymbol{z}}_k^{[1:r_{12}]} = \mathbf{O}_{zk}\boldsymbol{z}_k^{[1:r_{12}]}$. Let $\mathbf{B}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{[1:r_{12}]})$ and $\widetilde{\mathbf{B}}_k = \mathrm{cov}(\boldsymbol{x}_k, \tilde{\boldsymbol{z}}_k^{[1:r_{12}]})$. We have $\widetilde{\mathbf{B}}_k = \mathrm{cov}(\boldsymbol{x}_k, \boldsymbol{z}_k^{[1:r_{12}]})\mathbf{O}_{zk}^\top = \mathbf{B}_k\mathbf{O}_{zk}^\top$. Thus, $\mathrm{colsp}(\widetilde{\mathbf{B}}_k) = \mathrm{colsp}(\mathbf{B}_k)$. Define $\widetilde{\mathbf{B}}_{2A} = [\widetilde{\mathbf{B}}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$. We still have $\mathrm{colsp}(\mathbf{P}\widetilde{\mathbf{B}}_{2A}) = \mathrm{colsp}(\mathbf{P}\mathbf{B}_{2A})$. For $\ell = 1, \ldots, r_{12}$, recall that $\mathbf{V}_{B_1}^{[:,\ell]}$ and $\mathbf{V}_{B_2}^{[:,\ell]}$ are the ℓ-th pair of principal vectors of $\mathrm{colsp}(\mathbf{B}_1)$ and $\mathrm{colsp}(\mathbf{P}\mathbf{B}_{2A})$. Let $\{\widetilde{\mathbf{V}}_{B_k}\}_{k=1}^2$ be the matrices whose columns $\{\widetilde{\mathbf{V}}_{B_1}^{[:,\ell]}, \widetilde{\mathbf{V}}_{B_2}^{[:,\ell]}\}_{\ell=1}^{r_{12}}$ are another set of principal vectors of $\mathrm{colsp}(\mathbf{B}_1)$ and $\mathrm{colsp}(\mathbf{P}\mathbf{B}_{2A})$ with $\theta(\widetilde{\mathbf{V}}_{B_1}^{[:,\ell]}, \widetilde{\mathbf{V}}_{B_2}^{[:,\ell]}) = \theta_{B\ell}$. There exist orthogonal matrices $\{\mathbf{O}_{V_k}\}_{k=1}^2$ such that $\widetilde{\mathbf{V}}_{B_k} = \mathbf{V}_{B_k}\mathbf{O}_{V_k}$. Let $\boldsymbol{\Lambda}_B = \mathrm{diag}(\cos\theta_{B1}, \ldots, \cos\theta_{Br_{12}})$. Note that $\boldsymbol{\Lambda}_B = \widetilde{\mathbf{V}}_{B_1}^\top\widetilde{\mathbf{V}}_{B_2} = \mathbf{O}_{V_1}^\top\mathbf{V}_{B_1}^\top\mathbf{V}_{B_2}\mathbf{O}_{V_2} = \mathbf{O}_{V_1}^\top\boldsymbol{\Lambda}_B\mathbf{O}_{V_2}$. Then, $\mathbf{O}_{V_k} = \mathrm{diag}(\mathbf{M}_{k,1}, \ldots, \mathbf{M}_{k,m}, \mathbf{M}_{k,m+1})$, where $\mathbf{M}_{k,\ell}, \ell \le m$ is an orthogonal matrix with column dimension equal to the repetition number of the ℓ-th largest distinct nonzero singular value of $\boldsymbol{\Lambda}_B$, and $\mathbf{M}_{k,m+1}$ might be an empty matrix. By $\mathbf{O}_{V_1}\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_B\mathbf{O}_{V_2}$, we obtain $\mathbf{M}_{1,\ell} = \mathbf{M}_{2,\ell}$ for all $\ell \le m$. Define $r_\lambda = \mathrm{rank}(\boldsymbol{\Lambda}_B)$,

$$\tilde{\boldsymbol{c}}_{B\ell} = \frac{1}{2}\left(1 - \sqrt{\frac{1-\cos\theta_{B\ell}}{1+\cos\theta_{B\ell}}}\right)\left(\widetilde{\mathbf{V}}_{B_1}^{[:,\ell]} + \widetilde{\mathbf{V}}_{B_2}^{[:,\ell]}\right),$$

and $\mathbf{A}_B = \mathrm{diag}(a_{B1}, \ldots, a_{Br_{12}})$ with $a_{B\ell} = \frac{1}{2}\left[1 - \left(\frac{1-\cos\theta_{B\ell}}{1+\cos\theta_{B\ell}}\right)^{1/2}\right]$ for $\ell \le r_{12}$. Note that

$$\begin{aligned}
[\tilde{\boldsymbol{c}}_{B\ell}]_{\ell=1}^{r_{12}}\widetilde{\mathbf{V}}_{B_k}^\top &= [\tilde{\boldsymbol{c}}_{B\ell}]_{\ell=1}^{r_\lambda}(\widetilde{\mathbf{V}}_{B_k}^{[:,1:r_\lambda]})^\top \\
&= (\widetilde{\mathbf{V}}_{B_1}^{[:,1:r_\lambda]} + \widetilde{\mathbf{V}}_{B_2}^{[:,1:r_\lambda]})\mathbf{A}_B^{[1:r_\lambda,1:r_\lambda]}(\widetilde{\mathbf{V}}_{B_k}^{[:,1:r_\lambda]})^\top \\
&= (\mathbf{V}_{B_1}^{[:,1:r_\lambda]} + \mathbf{V}_{B_2}^{[:,1:r_\lambda]})\,\mathrm{diag}(\mathbf{M}_{1,1}, \ldots, \mathbf{M}_{1,m})\mathbf{A}_B^{[1:r_\lambda,1:r_\lambda]} \\
&\quad \cdot [\mathrm{diag}(\mathbf{M}_{1,1}, \ldots, \mathbf{M}_{1,m})]^\top(\mathbf{V}_{B_k}^{[:,1:r_\lambda]})^\top
\end{aligned}$$

$$
\begin{aligned}
&= (\mathbf{V}_{B_1}^{[:,1:r_\lambda]} + \mathbf{V}_{B_2}^{[:,1:r_\lambda]})\mathbf{A}_B^{[1:r_\lambda,1:r_\lambda]}(\mathbf{V}_{B_k}^{[:,1:r_\lambda]})^\top \\
&= [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_\lambda}(\mathbf{V}_{B_k}^{[:,1:r_\lambda]})^\top \\
&= [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}\mathbf{V}_{B_k}^\top.
\end{aligned}
$$

Hence, $[\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}\mathbf{V}_{B_k}^\top$ is unique for $k = 1, 2$. By Theorem 2 in [45], we have that $\boldsymbol{c}_k$ in (14) is unique for $k = 1, 2$. Then by $\mathbf{B}_1([c_\ell]_{\ell=1}^{r_{12}})^\top = \boldsymbol{c}_1$ and $\mathbf{P}\mathbf{B}_{2A}([c_\ell]_{\ell=1}^{r_{12}})^\top = \mathbf{P}(\boldsymbol{c}_2^\top, \mathbf{0}_{1\times(p_1-p_2)})^\top$, we have that both $\mathbf{B}_1([c_\ell]_{\ell=1}^{r_{12}})^\top$ and $\mathbf{P}\mathbf{B}_{2A}([c_\ell]_{\ell=1}^{r_{12}})^\top$ are unique. Then by the definition in (17), we obtain the uniqueness of $\boldsymbol{c}_k^*$ for $k = 1, 2$. Hence, $\boldsymbol{c} = \frac{1}{2}\sum_{k=1}^2[\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\boldsymbol{c}_k^*$ is unique.

A1.2. Proof of Theorem 2

Let $\tilde{r}_k = \mathrm{rank}(\widehat{\mathbf{X}}_k)$. From (S.17) in [45], we have $\tilde{r}_k = r_k$ with probability tending to 1 as $n \to \infty$. Due to Lemma S.1 in [45], we simply assume $\tilde{r}_k = r_k$ in the rest of the proof. Thus, $\widehat{\boldsymbol{\Lambda}}_k$ is rank-r_k, and then $\widehat{\mathbf{B}}_k = \widehat{\mathbf{V}}_k\widehat{\boldsymbol{\Lambda}}_k^{1/2}\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]}$ is rank-r_{12}.

From (S.7) of [45], we have $\lambda_1(\boldsymbol{\Sigma}_k) \asymp \lambda_{r_k}(\boldsymbol{\Sigma}_k)$. By Weyl's inequality [19, Theorem 3.3.16(a)] and Assumption 1 (I) and (V), $\kappa_1 \le \lambda_{k,p_k} = \lambda_{k,(r_k+1)+(p_k-r_k)-1} - \lambda_{r_k+1}(\mathrm{cov}(\boldsymbol{x}_k)) \le \lambda_{p_k-r_k}(\mathrm{cov}(\boldsymbol{e}_k)) \le \lambda_1(\mathrm{cov}(\boldsymbol{e}_k)) = \|\mathrm{cov}(\boldsymbol{e}_k)\|_2 \le \|\mathrm{cov}(\boldsymbol{e}_k)\|_\infty \le s_0$. Thus,

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

Let $\widetilde{\mathbf{Q}}_k \in \mathbb{R}^{p_k\times r_{12}}$ be the left singular matrix of $\mathbf{B}_k$. Note that $\|\mathbf{B}_k\|_2 \le \|\mathbf{V}_k\boldsymbol{\Lambda}_k^{1/2}\|_2\|\mathbf{U}_{\theta k}^{[:,1:r_{12}]}\|_2 = \lambda_1^{1/2}(\boldsymbol{\Sigma}_k)$. By (S.31) in [45], we have

$$
\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = O_P(\lambda_1^{1/2}(\boldsymbol{\Sigma}_k)\delta_\theta). \tag{A1}
$$

Thus, $\|\widehat{\mathbf{B}}_k\|_2 \le \|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 + \|\mathbf{B}_k\|_2 = O_P(\lambda_1^{1/2}(\boldsymbol{\Sigma}_k))$. By Lemma 1 of [28] and then Theorem 3 of [57], there exists an orthogonal matrix $\mathbf{O}_k$ such that

$$
\begin{aligned}
\|\widehat{\mathbf{Q}}_k - \widetilde{\mathbf{Q}}_k\mathbf{O}_k\|_F &\le \|\widehat{\mathbf{Q}}_k\mathbf{O}_k^\top - \widetilde{\mathbf{Q}}_k\|_F\|\mathbf{O}_k\|_2 \\
&\lesssim_P \lambda_1^{1/2}(\boldsymbol{\Sigma}_k)\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2/\lambda_1(\boldsymbol{\Sigma}_k) \lesssim_P \delta_\theta.
\end{aligned} \tag{A2}
$$

Here and in the following text, we write $A \lesssim_P B$ if and only if $A = O_P(B)$. Note that for any real matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we have

$$
\|\widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_2 - \mathbf{M}_1\mathbf{M}_2\|_2 \le \begin{cases} \|\widehat{\mathbf{M}}_1\|_2\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2 + \|\mathbf{M}_2\|_2\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2, \\ \|\mathbf{M}_1\|_2\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2 + \|\widehat{\mathbf{M}}_2\|_2\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2, \end{cases} \tag{A3}
$$

and

$$
\|\widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_2 - \mathbf{M}_1\mathbf{M}_2\|_F \le \begin{cases} \|\widehat{\mathbf{M}}_1\|_2\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_F + \|\mathbf{M}_2\|_2\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_F, \\ \|\mathbf{M}_1\|_2\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_F + \|\widehat{\mathbf{M}}_2\|_2\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_F. \end{cases} \tag{A4}
$$

Let $\mathbf{Q}_k = \widetilde{\mathbf{Q}}_k \mathbf{O}_k$ and $\mathbf{Q}_{2A} = [\mathbf{Q}_2; \mathbf{0}_{(p_1-p_2)\times r_{12}}]$. Note that the columns of $\mathbf{Q}_k$ form an orthonormal basis of $\mathrm{colsp}(\mathbf{B}_k)$, and those of $\mathbf{P}\mathbf{Q}_{2A}$ also form an orthonormal basis of $\mathrm{colsp}(\mathbf{P}\mathbf{B}_{2A})$. Let $\mathbf{\Theta}_B = \mathbf{Q}_1^\top \mathbf{P} \mathbf{Q}_{2A}$. Then by (A4) and (A2), we have

$$\begin{aligned}\|\widehat{\mathbf{\Theta}}_B - \mathbf{\Theta}_B\|_F &\le \|\widehat{\mathbf{Q}}_1^\top\|_2 \|\mathbf{P}\widehat{\mathbf{Q}}_{2A} - \mathbf{P}\mathbf{Q}_{2A}\|_F + \|\mathbf{P}\mathbf{Q}_{2A}\|_2 \|\widehat{\mathbf{Q}}_1^\top - \mathbf{Q}_1^\top\|_F \\ &\le \|\widehat{\mathbf{Q}}_1^\top\|_2 \|\mathbf{P}\|_2 \|\widehat{\mathbf{Q}}_{2A} - \mathbf{Q}_{2A}\|_F + \|\mathbf{P}\|_2 \|\mathbf{Q}_{2A}\|_2 \|\widehat{\mathbf{Q}}_1^\top - \mathbf{Q}_1^\top\|_F \\ &= \|\widehat{\mathbf{Q}}_2 - \mathbf{Q}_2\|_F + \|\widehat{\mathbf{Q}}_1 - \mathbf{Q}_1\|_F \\ &\lesssim_P \delta_\theta,\end{aligned}$$

and

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

By Weyl's inequality (see Theorem 3.3.16(c) in [19]),

$$\max_{1\le \ell \le r_{12}} |\sigma_\ell(\widehat{\mathbf{\Theta}}_B) - \sigma_\ell(\mathbf{\Theta}_B)| \le \|\widehat{\mathbf{\Theta}}_B - \mathbf{\Theta}_B\|_2 \le \|\widehat{\mathbf{\Theta}}_B - \mathbf{\Theta}_B\|_F \lesssim_P \delta_\theta. \tag{A5}$$

Denote $\{\widetilde{\mathbf{U}}_{B_k}\}_{k=1}^2$ to be one pair of orthogonal matrices such that $\mathbf{\Theta}_B = \widetilde{\mathbf{U}}_{B_1} \mathbf{\Lambda}_B \widetilde{\mathbf{U}}_{B_2}^\top$. Let $\sigma_{B,1} > \cdots > \sigma_{B,r_B}$ be the distinct singular values of $\mathbf{\Theta}_B$, and define $\sigma_{B,r_{12}+1}^2 = -\infty$. By Lemma 1 of [28] and then Theorem 2 of [57], there exists a matrix $\mathbf{O}_{B_k} = \mathrm{diag}(\mathbf{O}_{B_k,1}, \ldots, \mathbf{O}_{B_k,r_B})$, where $\mathbf{O}_{B_k,\ell}$ is an orthogonal matrix with column dimension equal to the repetition number of $\sigma_{B,\ell}$, such that

$$\begin{aligned}\|\widehat{\mathbf{U}}_{B_k} - \widetilde{\mathbf{U}}_{B_k}\mathbf{O}_{B_k}\|_F &\le \|\widehat{\mathbf{U}}_{B_k}\mathbf{O}_{B_k}^\top - \widetilde{\mathbf{U}}_{B_k}\|_F \|\mathbf{O}_{B_k}\|_2 \\ &\lesssim_P \min\left\{\delta_\theta \Big/ \min_{1\le \ell \le r_B} \{\sigma_{B,\ell}^2 - \sigma_{B,\ell+1}^2\}, 1\right\} \lesssim_P \delta_\theta.\end{aligned} \tag{A6}$$

We define $\widetilde{\mathbf{O}}_{B_2} = \mathrm{diag}(\mathbf{O}_{B_1,1}, \ldots, \mathbf{O}_{B_1,r_B-1}, \mathbf{O}_{B_1,r_B})$ if $\sigma_{B,r_B} \ne 0$, and otherwise let $\widetilde{\mathbf{O}}_{B_2} = \mathrm{diag}(\mathbf{O}_{B_1,1}, \ldots, \mathbf{O}_{B_1,r_B-1}, \mathbf{O}_{B_2,r_B})$. Let $\mathbf{U}_{B_1} = \widetilde{\mathbf{U}}_{B_1}\mathbf{O}_{B_1}$ and $\mathbf{U}_{B_2} = \widetilde{\mathbf{U}}_{B_2}\widetilde{\mathbf{O}}_{B_2}$. We have $\mathbf{U}_{B_1}\mathbf{\Lambda}_B\mathbf{U}_{B_2} = \widetilde{\mathbf{U}}_{B_1}\mathbf{O}_{B_1}\mathbf{\Lambda}_B\widetilde{\mathbf{O}}_{B_2}^\top\widetilde{\mathbf{U}}_{B_2}^\top = \widetilde{\mathbf{U}}_{B_1}\mathbf{\Lambda}_B\widetilde{\mathbf{U}}_{B_2}^\top = \mathbf{\Theta}_B$. Define $\mathbf{U}_{B_2}^\star = \widetilde{\mathbf{U}}_{B_2}\mathbf{O}_{B_2}$ and $r_{\theta_B} = \mathrm{rank}(\mathbf{\Theta}_B)$. Then,

$$\mathbf{U}_{B_2}^{[:,(r_{\theta_B}+1):r_{12}]} = \mathbf{U}_{B_2}^{\star[:,(r_{\theta_B}+1):r_{12}]} \quad \text{if} \quad r_{\theta_B} < r_{12}, \tag{A7}$$

$$\|\widehat{\mathbf{U}}_{B_1} - \mathbf{U}_{B_1}\|_F \lesssim_P \delta_\theta, \tag{A8}$$

and

$$\|\widehat{\mathbf{U}}_{B_2} - \mathbf{U}_{B_2}^\star\|_F \lesssim_P \delta_\theta. \tag{A9}$$

By (A3), (A5) and the above two inequalities,

$$\begin{aligned}&\left\|\widehat{\mathbf{U}}_{B_1}\widehat{\mathbf{\Lambda}}_B\widehat{\mathbf{U}}_{B_2}^\top - \mathbf{U}_{B_1}\mathbf{\Lambda}_B\mathbf{U}_{B_2}^{\star\top}\right\|_2 \\ &\le \|\widehat{\mathbf{U}}_{B_1}\widehat{\mathbf{\Lambda}}_B - \mathbf{U}_{B_1}\mathbf{\Lambda}_B\|_2 \|\widehat{\mathbf{U}}_{B_2}^\top\|_2 + \|\mathbf{U}_{B_1}\mathbf{\Lambda}_B\|_2 \|\widehat{\mathbf{U}}_{B_2}^\top - \mathbf{U}_{B_2}^{\star\top}\|_2\end{aligned}$$

$$\leq \|\widehat{\mathbf{U}}_{B_1} - \mathbf{U}_{B_1}\|_2\|\mathbf{\Lambda}_B\|_2 + \|\widehat{\mathbf{U}}_{B_1}\|_2\|\widehat{\mathbf{\Lambda}}_B - \mathbf{\Lambda}_B\|_2 + \|\mathbf{\Lambda}_B\|_2\|\widehat{\mathbf{U}}_{B_2}^\top - \mathbf{U}_{B_2}^{\star\top}\|_2$$
$$\lesssim_P \delta_\theta.$$

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

$$\|\mathbf{U}_{B_1}\mathbf{\Lambda}_B(\mathbf{U}_{B_2} - \mathbf{U}_{B_2}^\star)^\top\|_2 \lesssim_P \delta_\theta.$$

It follows that

$$\begin{aligned}\|\mathbf{U}_{B_2}^{[:,1:r_{\theta_B}]} - \mathbf{U}_{B_2}^{\star[:,1:r_{\theta_B}]}\|_F &\leq \sqrt{r_{12}}\|\mathbf{U}_{B_2}^{[:,1:r_{\theta_B}]} - \mathbf{U}_{B_2}^{\star[:,1:r_{\theta_B}]}\|_2 \\ &\leq \sqrt{r_{12}}\left\|\mathbf{\Lambda}_B^\dagger\right\|_2\left\|\mathbf{U}_{B_1}^\top\right\|_2\left\|\mathbf{U}_{B_1}\mathbf{\Lambda}_B(\mathbf{U}_{B_2} - \mathbf{U}_{B_2}^\star)^\top\right\|_2 \\ &\lesssim_P \delta_\theta. \end{aligned} \tag{A10}$$

Combining (A10), (A7) and (A9) yields

$$\|\widehat{\mathbf{U}}_{B_2} - \mathbf{U}_{B_2}\|_F \lesssim_P \delta_\theta. \tag{A11}$$

By (6), we have that the ℓ-th columns of $\mathbf{V}_{B_1} := \mathbf{Q}_1\mathbf{U}_{B_1}$ and $\mathbf{V}_{B_2} := \mathbf{P}\mathbf{Q}_{2A}\mathbf{U}_{B_2}$ are the ℓ-th pair of principal vectors of $\mathrm{colsp}(\mathbf{B}_1)$ and $\mathrm{colsp}(\mathbf{P}\mathbf{B}_{2A})$. By (A4), (A2) and (A8), we have

$$\begin{aligned}\|\widehat{\mathbf{V}}_{B_1} - \mathbf{V}_{B_1}\|_F &= \|\mathbf{Q}_1\mathbf{U}_{B_1} - \widehat{\mathbf{Q}}_1\widehat{\mathbf{U}}_{B_1}\|_2 \\ &\leq \|\widehat{\mathbf{U}}_{B_1}\|_2\|\widehat{\mathbf{Q}}_1 - \mathbf{Q}_1\|_F + \|\mathbf{Q}_1\|_2\|\widehat{\mathbf{U}}_{B_1} - \mathbf{U}_{B_1}\|_F \\ &\lesssim_P \delta_\theta. \end{aligned} \tag{A12}$$

Similarly, by (A11) we obtain

$$\|\widehat{\mathbf{V}}_{B_2} - \mathbf{V}_{B_2}\|_F \lesssim_P \delta_\theta. \tag{A13}$$

Then, together with (A4) and (A1), we have

$$\begin{aligned}\|\widehat{\mathbf{V}}_{B_1}^\top\widehat{\mathbf{B}}_1 - \mathbf{V}_{B_1}^\top\mathbf{B}_1\|_F &\leq \|\widehat{\mathbf{B}}_1\|_2\|\widehat{\mathbf{V}}_{B_1}^\top - \mathbf{V}_{B_1}^\top\|_F + \|\mathbf{V}_{B_1}^\top\|_2\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_F \\ &\lesssim_P \lambda_1^{1/2}(\mathbf{\Sigma}_1)\delta_\theta, \end{aligned} \tag{A14}$$

and similarly,

$$\|\widehat{\mathbf{V}}_{B_2}^\top\mathbf{P}\widehat{\mathbf{B}}_{2A} - \mathbf{V}_{B_2}^\top\mathbf{P}\mathbf{B}_{2A}\|_F \lesssim_P \lambda_1^{1/2}(\mathbf{\Sigma}_2)\delta_\theta. \tag{A15}$$

By the results given in (S.16), (S.17) and (S.7) of [45], we have $|\lambda_\ell(\widehat{\mathbf{\Sigma}}_k) - \lambda_\ell(\mathbf{\Sigma}_k)| \lesssim_P \lambda_1(\mathbf{\Sigma}_k)/\sqrt{n}$ for all $\ell \leq r_k$, $[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2} = [\sum_{\ell=1}^{r_k}\lambda_\ell(\widehat{\mathbf{\Sigma}}_k)]^{1/2} \geq [r_k(1 - o_P(1))\lambda_{r_k}(\mathbf{\Sigma}_k)]^{1/2}$, and $\lambda_1(\mathbf{\Sigma}_k) \asymp \lambda_{r_k}(\mathbf{\Sigma}_k)$. Then by the mean value theorem, we obtain

$$\begin{aligned}&\left|[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{1/2} - [\mathrm{tr}(\mathbf{\Sigma}_k)]^{1/2}\right| \\ &\leq \frac{1}{2}\left|\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k) - \mathrm{tr}(\mathbf{\Sigma}_k)\right| \cdot \max\left\{[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_k)]^{-1/2}, [\mathrm{tr}(\mathbf{\Sigma}_k)]^{-1/2}\right\}\end{aligned}$$

$$\lesssim_P \lambda_1^{1/2}(\boldsymbol{\Sigma}_k)/\sqrt{n}. \tag{A16}$$

Hence,

$$\begin{aligned}
&\big|[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_k)]^{-1/2} - [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{-1/2}\big| \\
&= \big|[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_k)]^{1/2} - [\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{1/2}\big| \Big/ \big([\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_k)]^{1/2}[\mathrm{tr}(\boldsymbol{\Sigma}_k)]^{1/2}\big) \\
&\lesssim_P \lambda_1^{-1/2}(\boldsymbol{\Sigma}_k)/\sqrt{n}.
\end{aligned} \tag{A17}$$

By (A4), (A14) and (A17),

$$\begin{aligned}
&\Big\|\widehat{\mathbf{V}}_{B_1}^\top \widehat{\mathbf{B}}_1[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_1)]^{-1/2} - \mathbf{V}_{B_1}^\top \mathbf{B}_1[\mathrm{tr}(\boldsymbol{\Sigma}_1)]^{-1/2}\Big\|_F \\
&\lesssim_P \lambda_1^{-1/2}(\boldsymbol{\Sigma}_1)(\lambda_1^{1/2}(\boldsymbol{\Sigma}_1)\delta_\theta) + \lambda_1^{1/2}(\boldsymbol{\Sigma}_1)\lambda_1^{-1/2}(\boldsymbol{\Sigma}_1)/\sqrt{n} \\
&\lesssim_P \delta_\theta.
\end{aligned}$$

Similarly, by (A15), $\big\|\widehat{\mathbf{V}}_{B_2}^\top \mathbf{P}\widehat{\mathbf{B}}_{2A}[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_2)]^{-1/2} - \mathbf{V}_{B_2}^\top \mathbf{P}\mathbf{B}_{2A}[\mathrm{tr}(\boldsymbol{\Sigma}_2)]^{-1/2}\big\|_F \lesssim_P \delta_\theta$. Thus,

$$\begin{aligned}
&\Big\|\big(\widehat{\mathbf{V}}_{B_1}^\top \widehat{\mathbf{B}}_1[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_1)]^{-1/2} + \widehat{\mathbf{V}}_{B_2}^\top \mathbf{P}\widehat{\mathbf{B}}_{2A}[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_2)]^{-1/2}\big) \\
&\quad - \big(\mathbf{V}_{B_1}^\top \mathbf{B}_1[\mathrm{tr}(\boldsymbol{\Sigma}_1)]^{-1/2} + \mathbf{V}_{B_2}^\top \mathbf{P}\mathbf{B}_{2A}[\mathrm{tr}(\boldsymbol{\Sigma}_2)]^{-1/2}\big)\Big\|_F \lesssim_P \delta_\theta.
\end{aligned} \tag{A18}$$

Define $\mathbf{C}_B = [\boldsymbol{c}_{B\ell}]_{\ell=1}^{r_{12}}$ and $\mathbf{A}_B = \mathrm{diag}(a_{B1}, \ldots, a_{Br_{12}})$ with $a_{B\ell} = \frac{1}{2}\Big[1 - \big(\frac{1-\boldsymbol{\Lambda}_B^{[\ell,\ell]}}{1+\boldsymbol{\Lambda}_B^{[\ell,\ell]}}\big)^{1/2}\Big]$. We have $\mathbf{C}_B = (\mathbf{V}_{B_1} + \mathbf{V}_{B_2})\mathbf{A}_B$. By the same technique used to derive (S.32) in [45], we have $\|\widehat{\mathbf{A}}_B - \mathbf{A}_B\|_F \lesssim_P \delta_\theta^{1/2}$. From (A12) and (A13), $\|(\widehat{\mathbf{V}}_{B_1} + \widehat{\mathbf{V}}_{B_2}) - (\mathbf{V}_{B_1} + \mathbf{V}_{B_2})\|_F \lesssim_P \delta_\theta$. Then by (A4),

$$\|\widehat{\mathbf{C}}_B - \mathbf{C}_B\|_F \lesssim_P \delta_\theta^{1/2} + \delta_\theta \lesssim_P \delta_\theta^{1/2}. \tag{A19}$$

From (S.23) in [45], $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F \lesssim \delta_\theta$. Using the same proof technique for (A8) and (A11), we have $\|\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]} - \mathbf{U}_{\theta k}^{[:,1:r_{12}]}\|_F \lesssim_P \delta_\theta$. Then following the same proof lines for (S.28) in [45], we can obtain

$$\Big\|(\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]})^\top \widehat{\boldsymbol{\Lambda}}_k^{-1/2}\widehat{\mathbf{V}}_k^\top - (\mathbf{U}_{\theta k}^{[:,1:r_{12}]})^\top \boldsymbol{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\Big\|_F \lesssim_P \lambda_1^{-1/2}(\boldsymbol{\Sigma}_k)\delta_\theta.$$

From the results given in (S.9), (S.13), (S.15) and (S.32) of [45], we have that $\max\{\|\widehat{\mathbf{X}}_k\|_F, \|\mathbf{X}_k\|_F\} \lesssim_P \sqrt{n\lambda_1(\boldsymbol{\Sigma}_k)}$, $\|\widehat{\mathbf{X}}_k - \mathbf{X}_k\|_F \lesssim_P \min\{\sqrt{\lambda_1(\boldsymbol{\Sigma}_k)/n} + \sqrt{p_k \log p_k}, \sqrt{n\lambda_1(\boldsymbol{\Sigma}_k)}\}$, and $\|\widehat{\mathbf{A}}_C - \mathbf{A}_C\|_F \lesssim_P \delta_\theta^{1/2}$, where $\mathbf{A}_C = \mathrm{diag}(a_1, \ldots, a_{r_{12}})$ with $a_\ell = \frac{1}{2}\big[1 - \big(\frac{1-\sigma_\ell(\boldsymbol{\Theta})}{1+\sigma_\ell(\boldsymbol{\Theta})}\big)^{1/2}\big]$. Let $\mathbf{Z}_k = \mathbf{U}_{\theta k}^\top \boldsymbol{\Lambda}_k^{-1/2}\mathbf{V}_k^\top \mathbf{X}_k$ and $\mathbf{C}_0 = \mathbf{A}_C \sum_{j=1}^2 \mathbf{Z}_j^{[1:r_{12},:]}$, which are the sample matrices of $\boldsymbol{z}_k$ and $(c_1, \ldots, c_{r_{12}})^\top$, respectively. Then by (A4),

$$\Big\|\widehat{\mathbf{Z}}_k^{[1:r_{12},:]} - \mathbf{Z}_k^{[1:r_{12},:]}\Big\|_F$$

$$
\begin{aligned}
&= \left\| (\widehat{\mathbf{U}}_{\theta k}^{[:,1:r_{12}]})^\top \widehat{\mathbf{\Lambda}}_k^{-1/2} \widehat{\mathbf{V}}_k^\top \widehat{\mathbf{X}}_k - (\mathbf{U}_{\theta k}^{[:,1:r_{12}]})^\top \mathbf{\Lambda}_k^{-1/2} \mathbf{V}_k^\top \mathbf{X}_k \right\|_F \\
&\lesssim_P \lambda_1^{-1/2}(\mathbf{\Sigma}_k) \delta_\theta \sqrt{n \lambda_1(\mathbf{\Sigma}_k)} \\
&\quad + \min\{\sqrt{\lambda_1(\mathbf{\Sigma}_k)/n} + \sqrt{p_k \log p_k}, \sqrt{n\lambda_1(\mathbf{\Sigma}_k)}\} / \sqrt{\lambda_1(\mathbf{\Sigma}_k)} \\
&\lesssim_P \delta_\theta \sqrt{n},
\end{aligned}
$$

and thus,

$$
\|\widehat{\mathbf{C}}_0 - \mathbf{C}_0\|_F \lesssim_P \delta_\theta^{1/2} \sqrt{n}. \tag{A20}
$$

From (A18), (A19), (A20) and (A4), we obtain

$$
\|\widehat{\mathbf{C}} - \mathbf{C}\|_2 \le \|\widehat{\mathbf{C}} - \mathbf{C}\|_F = O_P(\delta_\theta^{1/2} \sqrt{n}). \tag{A21}
$$

Combining (A16) and (A21) yields

$$
\|\widehat{\mathbf{C}}^{(k)} - \mathbf{C}^{(k)}\|_2 \le \|\widehat{\mathbf{C}}^{(k)} - \mathbf{C}^{(k)}\|_F \lesssim_P \delta_\theta^{1/2} \sqrt{n} \cdot \lambda_1^{1/2}(\mathbf{\Sigma}_k).
$$

By (S.14) in [45], there exists a constant $\kappa_3 \in (0, 1]$ such that $\|\mathbf{X}_k\|_F \ge \|\mathbf{X}_k\|_2 \ge [\kappa_3 + o_P(1)] \sqrt{n\lambda_1(\mathbf{\Sigma}_k)}$. Hence,

$$
\frac{\|\widehat{\mathbf{C}} - \mathbf{C}\|_\star^2}{\frac{1}{2}(\|\mathbf{X}_1^S\|_\star^2 + \|\mathbf{X}_2^S\|_\star^2)} = O_P(\delta_\theta),
$$

and

$$
\frac{\|\widehat{\mathbf{C}}^{(k)} - \mathbf{C}^{(k)}\|_\star^2}{\|\mathbf{X}_k\|_\star^2} = O_P(\delta_\theta).
$$

Let $\boldsymbol{c}_0 = (c_1, \ldots, c_{r_{12}})^\top$ and $\boldsymbol{z}_c = [\boldsymbol{z}_1^{[1:r_{12}]}; \boldsymbol{z}_2^{[1:r_{12}]}]$. Define $\mathbf{Z}_c = [\mathbf{Z}_1^{[1:r_{12},:]}; \mathbf{Z}_2^{[1:r_{12},:]}]$, which is the sample matrix of $\boldsymbol{z}_c$. We have $\boldsymbol{c}_0 = \mathbf{A}_C[\mathbf{I}_{r_{12}\times r_{12}}, \mathbf{I}_{r_{12}\times r_{12}}]\boldsymbol{z}_c$ and $\mathbf{C}_0 = \mathbf{A}_C[\mathbf{I}_{r_{12}\times r_{12}}, \mathbf{I}_{r_{12}\times r_{12}}]\mathbf{Z}_c$. From the central limit theorem,

$$
\left\| \frac{1}{n} \mathbf{Z}_c \mathbf{Z}_c^\top - \mathrm{cov}(\boldsymbol{z}_c) \right\|_F \le 2r_{12} \left\| \frac{1}{n} \mathbf{Z}_c \mathbf{Z}_c^\top - \mathrm{cov}(\boldsymbol{z}_c) \right\|_{\max} \lesssim_P n^{-1/2}.
$$

Hence,

$$
\begin{aligned}
\left\| \frac{1}{n} \mathbf{C}_0 \mathbf{C}_0^\top - \mathrm{cov}(\boldsymbol{c}_0) \right\|_F &\le \left\| \frac{1}{n} \mathbf{Z}_c \mathbf{Z}_c^\top - \mathrm{cov}(\boldsymbol{z}_c) \right\|_F \left\| \mathbf{A}_C[\mathbf{I}_{r_{12}\times r_{12}}, \mathbf{I}_{r_{12}\times r_{12}}] \right\|_F^2 \\
&\lesssim_P n^{-1/2}.
\end{aligned}
$$

Then,

$$
\left\| \frac{1}{n} \mathbf{C}\mathbf{C}^\top - \mathrm{cov}(\boldsymbol{c}) \right\|_F \le \left\| \frac{1}{n} \mathbf{C}_0 \mathbf{C}_0^\top - \mathrm{cov}(\boldsymbol{c}_0) \right\|_F \|\mathbf{C}_B \mathbf{S}\|_F^2 \lesssim_P n^{-1/2}.
$$

By (A21), we have

$$
\left\| \frac{1}{n} \widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top - \frac{1}{n} \mathbf{C}\mathbf{C}^\top \right\|_F \le \frac{1}{n} \|\widehat{\mathbf{C}} - \mathbf{C}\|_F (\|\widehat{\mathbf{C}}\|_F + \|\mathbf{C}\|_F) \lesssim_P \delta_\theta^{1/2}.
$$

Combining the above two inequalities yields

$$\left\|\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top - \mathrm{cov}(\boldsymbol{c})\right\|_F \lesssim_P \delta_\theta^{1/2}.$$

By Weyl's inequality (see Theorem 3.3.16(c) in [19]),

$$\begin{aligned}&\max_{\ell\le r_{12}}\left|\lambda_\ell(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top) - \lambda_\ell(\mathrm{cov}(\boldsymbol{c}))\right| \\ &\le \left\|\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top - \mathrm{cov}(\boldsymbol{c})\right\|_2 \le \left\|\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top - \mathrm{cov}(\boldsymbol{c})\right\|_F \\ &\lesssim_P \delta_\theta^{1/2}.\end{aligned}$$

Then,

$$\left|\mathrm{tr}(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top) - \mathrm{tr}(\mathrm{cov}(\boldsymbol{c}))\right| \le \sum_{\ell=1}^{r_{12}}\left|\lambda_\ell(\frac{1}{n}\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top) - \lambda_\ell(\mathrm{cov}(\boldsymbol{c}))\right| \lesssim_P \delta_\theta^{1/2}.$$

The proof is complete.

A1.3. Proof of Theorem 3

By (A2), there exists a matrix $\mathbf{Q}_k$, whose columns form an orthonormal basis of $\mathrm{colsp}(\mathbf{B}_k)$, such that $\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_F = O_p(\delta_\theta)$. Note that $\mathrm{tr}(\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A}(\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A})^\top) = \|\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A}\|_F^2$. Then by (A4), for any $\mathbf{P}\in\Pi_{p_1}$, we have

$$\begin{aligned}&\left|\|\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}\widehat{\mathbf{Q}}_{2A}\|_F^2\right| \\ &\le \left|\|\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A}\|_F - \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}\widehat{\mathbf{Q}}_{2A}\|_F\right|(\|\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A}\|_F + \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}\widehat{\mathbf{Q}}_{2A}\|_F) \\ &\le \|\mathbf{Q}_1^\top\mathbf{P}\mathbf{Q}_{2A} - \widehat{\mathbf{Q}}_1^\top\mathbf{P}\widehat{\mathbf{Q}}_{2A}\|_F(\|\mathbf{Q}_1^\top\mathbf{P}\|_F\|\mathbf{Q}_{2A}\|_F + \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}\|_F\|\widehat{\mathbf{Q}}_{2A}\|_F) \\ &\le (\|\widehat{\mathbf{Q}}_1^\top\|_2\|\mathbf{P}\mathbf{Q}_{2A} - \mathbf{P}\widehat{\mathbf{Q}}_{2A}\|_F + \|\mathbf{P}\mathbf{Q}_{2A}\|_2\|\mathbf{Q}_1^\top - \widehat{\mathbf{Q}}_1^\top\|_F)2r_k \\ &= O_P(\delta_\theta).\end{aligned}$$

Hence, $\left|\|\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}_*\widehat{\mathbf{Q}}_{2A}\|_F^2\right| = O_P(\delta_\theta) = \left|\|\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\widehat{\mathbf{P}}_*\widehat{\mathbf{Q}}_{2A}\|_F^2\right|$. Note that $\|\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}\|_F^2 \le \|\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}\|_F^2$ and $\|\widehat{\mathbf{Q}}_1^\top\mathbf{P}_*\widehat{\mathbf{Q}}_{2A}\|_F^2 \le \|\widehat{\mathbf{Q}}_1^\top\widehat{\mathbf{P}}_*\widehat{\mathbf{Q}}_{2A}\|_F^2$. We have

$$\begin{aligned}0 &\le \|\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}\|_F^2 - \|\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}\|_F^2 \\ &= (\|\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}_*\widehat{\mathbf{Q}}_{2A}\|_F^2) + (\|\widehat{\mathbf{Q}}_1^\top\widehat{\mathbf{P}}_*\widehat{\mathbf{Q}}_{2A}\|_F^2 - \|\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}\|_F^2) \\ &\quad + (\|\widehat{\mathbf{Q}}_1^\top\mathbf{P}_*\widehat{\mathbf{Q}}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\widehat{\mathbf{P}}_*\widehat{\mathbf{Q}}_{2A}\|_F^2) \\ &\le \left|\|\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}\|_F^2 - \|\widehat{\mathbf{Q}}_1^\top\mathbf{P}_*\widehat{\mathbf{Q}}_{2A}\|_F^2\right| + \left|\|\widehat{\mathbf{Q}}_1^\top\widehat{\mathbf{P}}_*\widehat{\mathbf{Q}}_{2A}\|_F^2 - \|\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}\|_F^2\right| \\ &= O_P(\delta_\theta).\end{aligned}$$

Hence, $\left|\mathrm{tr}\left(\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A}(\mathbf{Q}_1^\top\widehat{\mathbf{P}}_*\mathbf{Q}_{2A})^\top\right) - \mathrm{tr}\left(\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A}(\mathbf{Q}_1^\top\mathbf{P}_*\mathbf{Q}_{2A})^\top\right)\right| = O_P(\delta_\theta)$.

Appendix A2: Selection of matrix ranks

Following [45], we select $r_k = \text{rank}(\mathbf{\Sigma}_k)$ for $k = 1, 2$ and $r_{12} = \text{rank}(\mathbf{\Sigma}_{12})$ by the ED method of [38] and the MDL-IC method of [48], respectively. Specifically, the ED method estimates r_k by

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

where $\hat{\lambda}_{k,\ell}$ is the ℓ-th eigenvalue of $\mathbf{Y}_k\mathbf{Y}_k^\top/n$, $T_k = \left|\{i : \hat{\lambda}_{k,i} \geq \frac{1}{m_k}\sum_{\ell=1}^{m_k}\hat{\lambda}_{k,\ell}\}\right| \wedge \frac{m_k}{10}$ with $m_k = n \wedge p_k$, and δ is calibrated as in Section IV of [38]. If there exit two variables from different denoised datsets have a significant nonzero correlation detected by the normal approximation test of [10], then we conclude $r_{12} > 0$. Otherwise, the CDPA method is unnecessary due to no correlation between the two signal datasets. The MDL-IC method estimates nonzero r_{12} by

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

where s_ℓ is the ℓ-th largest singular value of $(\mathbf{U}_{12}^{[:,1:\hat{r}_1]})^\top\mathbf{U}_{22}^{[:,1:\hat{r}_2]}$, and the i-th column of $\mathbf{U}_{k2} \in \mathbb{R}^{n\times n}$ is the right-singular vector of $\mathbf{Y}_k$ corresponding to its i-th largest singular value.

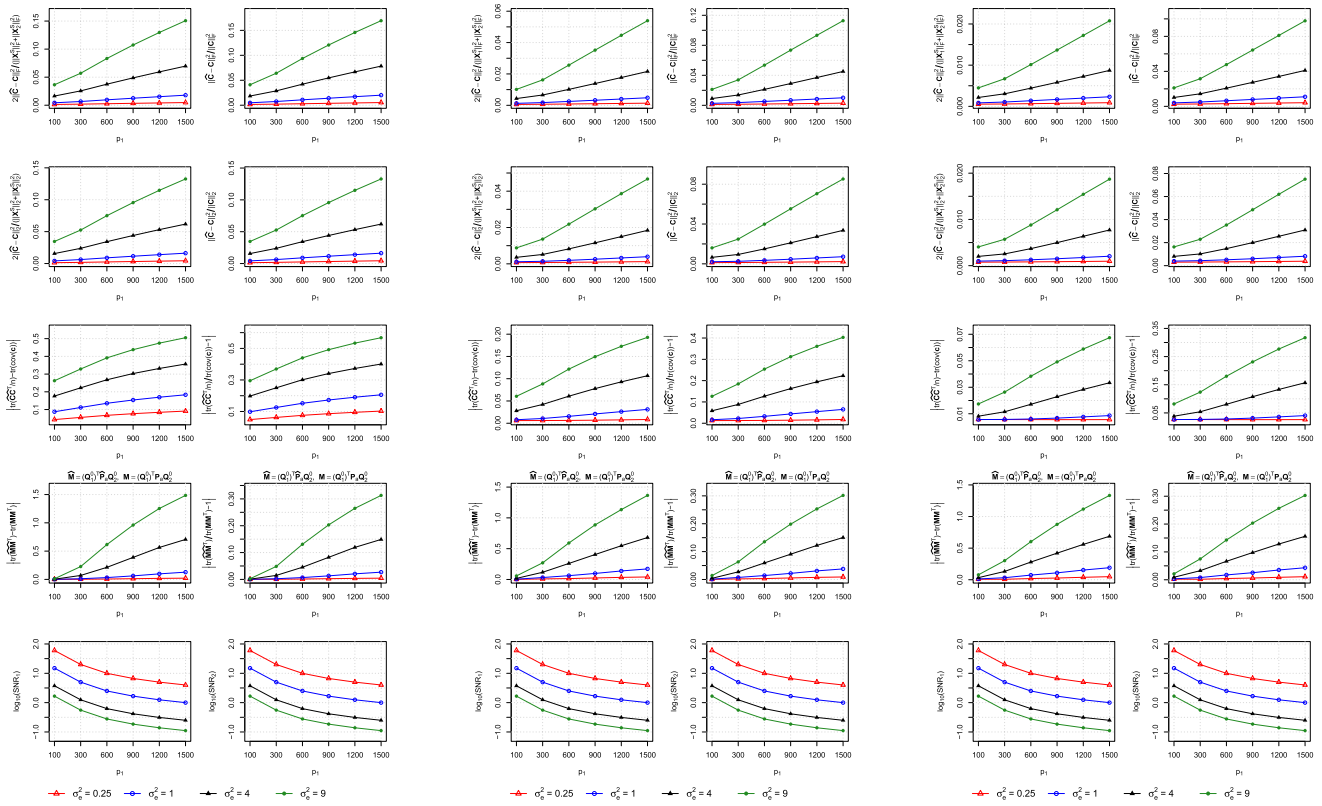
Appendix A3: Additional simulation results

Figures A1–A4 display the simulation results for the CDPA estimators under Setups 1 and 2. The result analysis given in Section 4 generally holds here.

Appendix A4: Additional real-data results

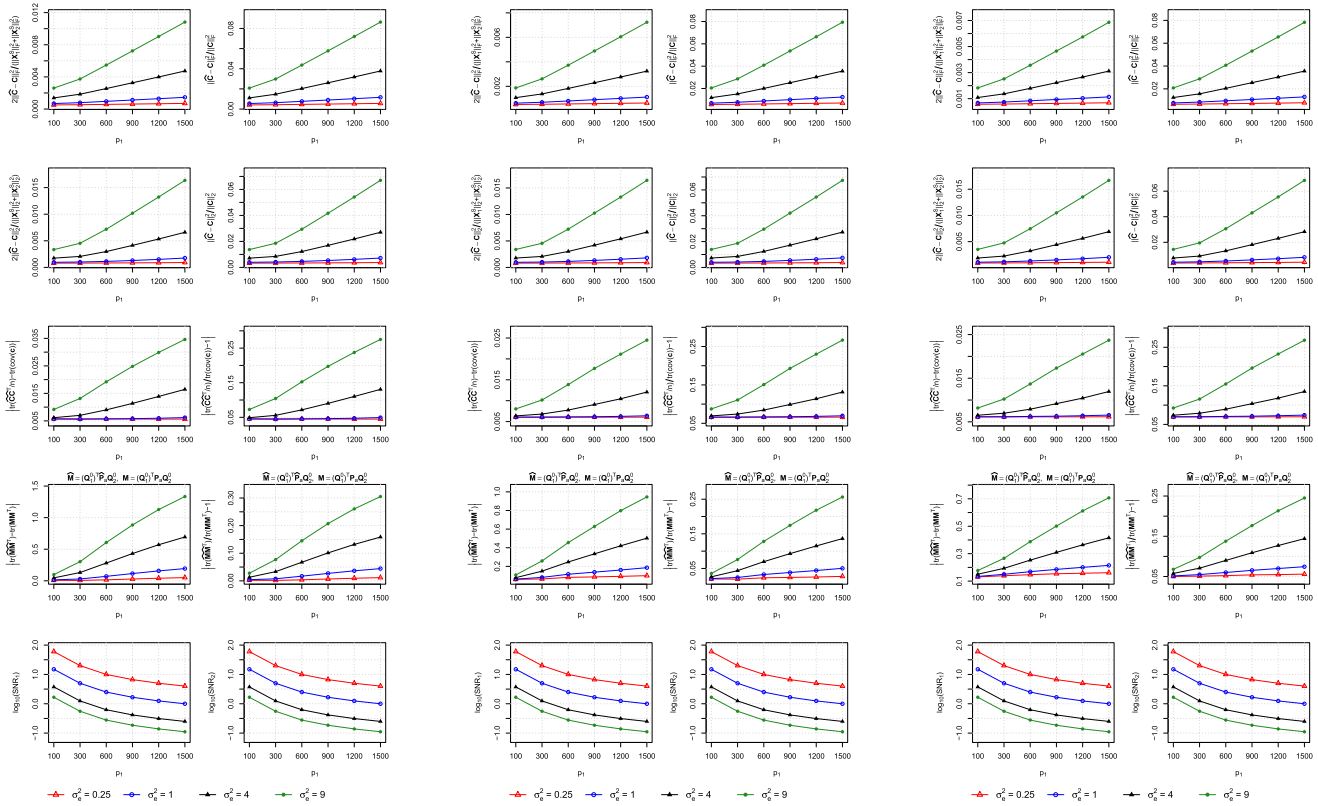
A4.1. Additional results of HCP motor-task functional MRI data

We also apply the five D-CCA-type methods (OnPLS, DISCO-SCA, COBE, JIVE, and AJIVE) to analyze the HCP motor-task functional MRI data. The result of OnPLS is not available because this method exceeds the 62GB memory limit of our computing node due to the SVD computation of the large 91,282×91,282 matrix $\mathbf{Y}_L\mathbf{Y}_R^\top$ in its algorithm. The COBE method fails to generate nonzero common-source matrix estimates. Figure A5 shows the maps of $\widehat{\text{var}}(\boldsymbol{c}_L)$ and $\widehat{\text{var}}(\boldsymbol{c}_R)$ obtained from the DISCO-SCA, JIVE and AJIVE methods. Similar to those shown in Figure 4 (c) and (d) for D-CCA, the common-source vectors $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ of the three methods have estimated variance maps that are asymmetric on the two hemispheres, and thus are less plausible than the common-pattern vector $\boldsymbol{c}$ of CDPA to represent the common pattern of the left-hand and right-hand tasks on the brain.



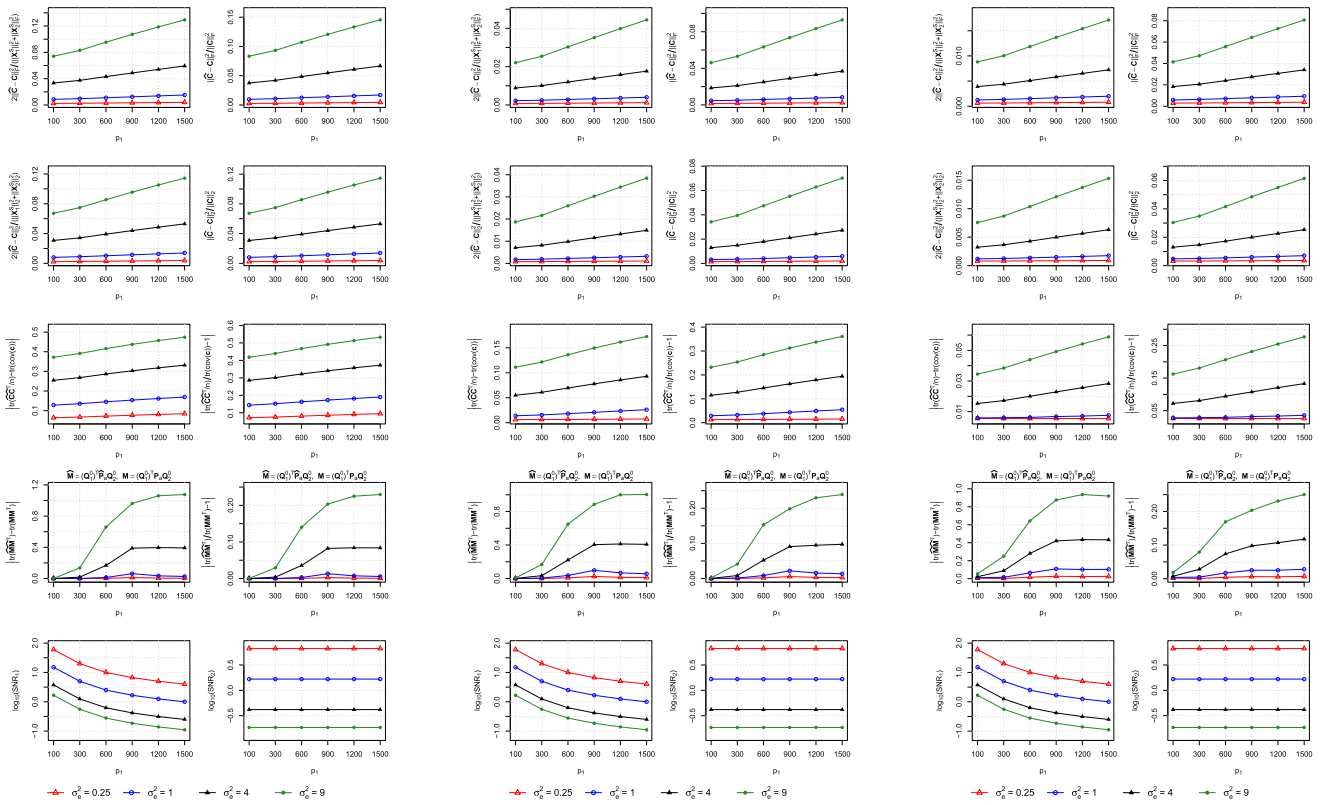
(a) Setup 1 with $\theta = 0^\circ$ (b) Setup 1 with $\theta = 15^\circ$ (c) Setup 1 with $\theta = 30^\circ$

FIG A1. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 1 with* $\theta \in \{0^\circ, 15^\circ, 30^\circ\}$.



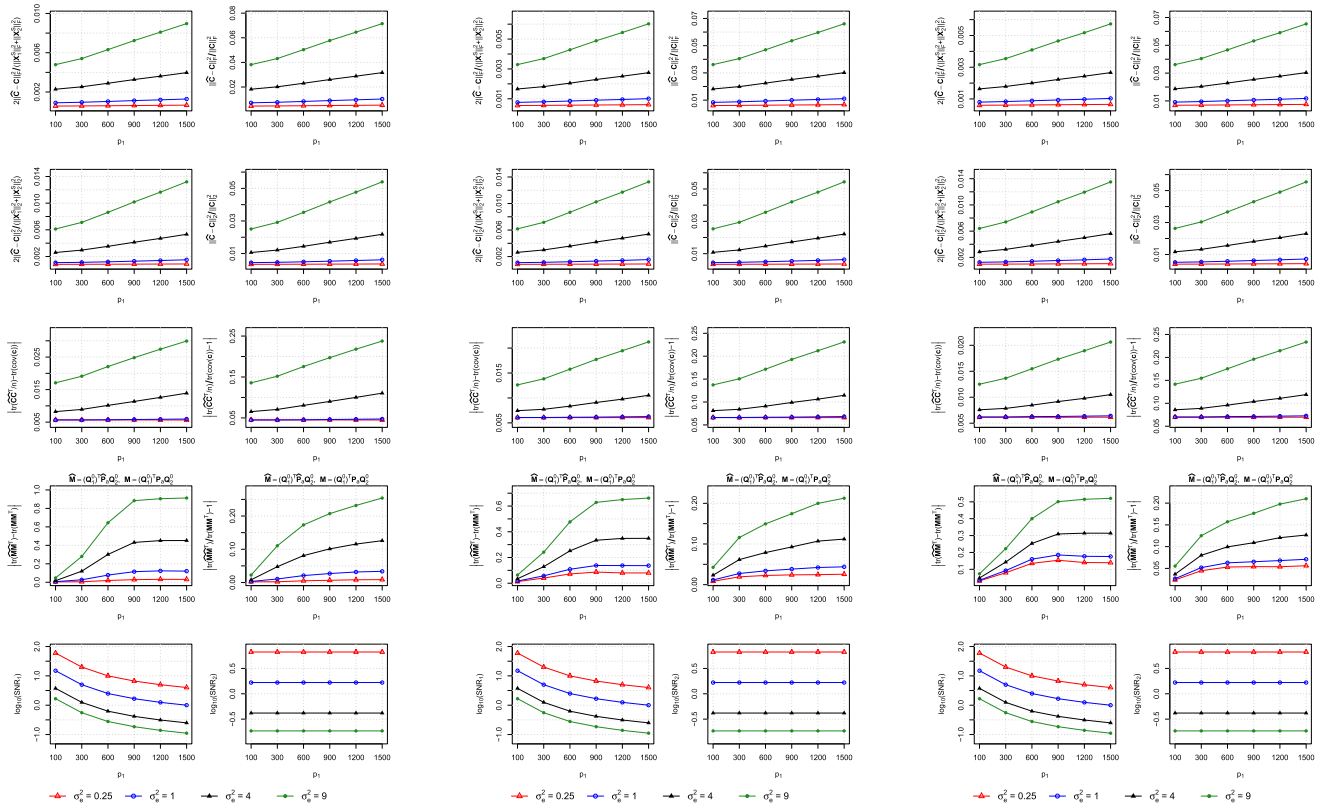
(a) Setup 1 with $\theta = 45°$ (b) Setup 1 with $\theta = 60°$ (c) Setup 1 with $\theta = 75°$

FIG A2. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 1 with* $\theta \in \{45°, 60°, 75°\}$.



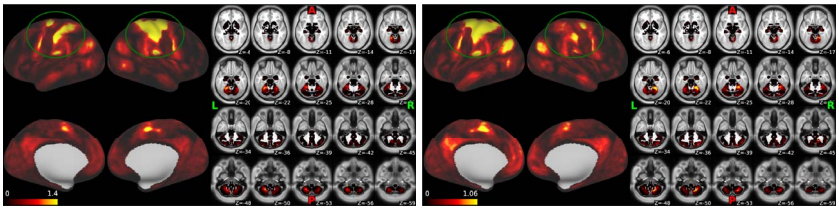
(a) Setup 2 with $\theta = 0^\circ$ (b) Setup 2 with $\theta = 15^\circ$ (c) Setup 2 with $\theta = 30^\circ$

FIG A3. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 2 with* $\theta \in \{0^\circ, 15^\circ, 30^\circ\}$.

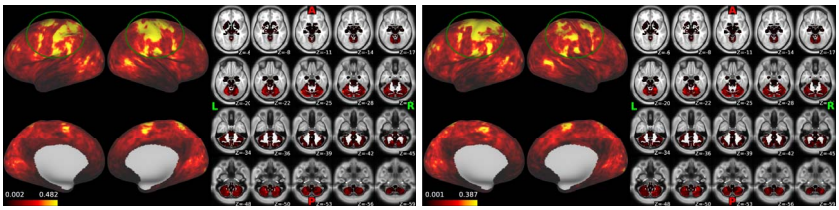


(a) Setup 2 with $\theta = 45°$ (b) Setup 2 with $\theta = 60°$ (c) Setup 2 with $\theta = 75°$

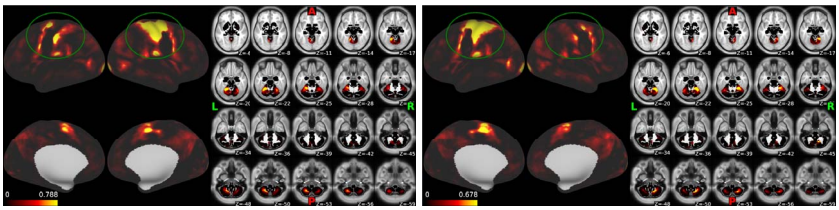
FIG A4. *Average errors of CDPA estimates over 1000 replications and the signal-to-noise ratios for Setup 2 with* $\theta \in \{45°, 60°, 75°\}$.



(a) $\widehat{\mathrm{var}}(\boldsymbol{c}_L)$ of DISCO-SCA (b) $\widehat{\mathrm{var}}(\boldsymbol{c}_R)$ of DISCO-SCA



(c) $\widehat{\mathrm{var}}(\boldsymbol{c}_L)$ of JIVE (d) $\widehat{\mathrm{var}}(\boldsymbol{c}_R)$ of JIVE



(e) $\widehat{\mathrm{var}}(\boldsymbol{c}_L)$ of AJIVE (f) $\widehat{\mathrm{var}}(\boldsymbol{c}_R)$ of AJIVE

FIG A5. *The variance maps estimated by the DISCO-SCA, JIVE, and AJIVE methods for HCP motor-task functional MRI data. The notation* $\widehat{\mathrm{var}}$ *denotes the sample variance vector obtained from the corresponding recovered sample matrix. 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.*

A4.2. Additional results of TCGA breast cancer genomic datasets

We also apply the same clustering method used in Section 5.2 to each recovered matrix from the five D-CCA-type methods: OnPLS, COBE, JIVE, AJIVE, and DISCO-SCA. Table A1 reports the p-values of the log-rank test and the Peto-Peto's Wilcoxon test for the survival differences among the clusters from each of these matrices. All the five methods have the p-values above 0.05 and thus fail to discover breast cancer subtypes with significant survival differences.

TABLE A1

Log-rank test and Peto-Peto's Wilcoxon test for survival curve differences among the clusters identified from each matrix of the five D-CCA-type methods for TCGA breast cancer datasets.

Data	Log-rank/Peto's p-values for competing methods				
	OnPLS	COBE	JIVE	AJIVE	DISCO-SCA
$\widehat{\mathbf{X}}_{\text{DNA}}$	0.340/0.568	0.093/0.137	0.585/0.389	0.125/0.139	0.774/0.866
$\widehat{\mathbf{X}}_{\text{mRNA}}$	0.060/0.078	0.189/0.107	0.577/0.589	0.266/0.192	0.175/0.116
$[\widehat{\mathbf{X}}^N_{\text{DNA}};\widehat{\mathbf{X}}^N_{\text{mRNA}}]$	0.461/0.506	0.325/0.319	0.207/0.225	0.296/0.330	0.452/0.517
$\widehat{\mathbf{C}}_{\text{DNA}}$	0.846/0.957	NA	0.133/0.156	0.213/0.193	0.147/0.204
$\widehat{\mathbf{C}}_{\text{mRNA}}$	0.060/0.078	NA	0.133/0.156	0.083/0.116	0.205/0.097
$[\widehat{\mathbf{C}}^N_{\text{DNA}};\widehat{\mathbf{C}}^N_{\text{mRNA}}]$	0.493/0.707	NA	0.133/0.156	0.321/0.240	0.217/0.104
$\widehat{\mathbf{D}}_{\text{DNA}}$	0.618/0.559	0.093/0.137	0.137/0.086	0.282/0.205	0.791/0.657
$\widehat{\mathbf{D}}_{\text{mRNA}}$	NA	0.189/0.107	0.074/0.076	0.439/0.141	0.846/0.842
$[\widehat{\mathbf{D}}^N_{\text{DNA}};\widehat{\mathbf{D}}^N_{\text{mRNA}}]$	NA	0.325/0.319	0.089/0.062	0.155/0.187	0.614/0.594

Notes: Denote $\mathbf{M}^N = \mathbf{M}/\|\mathbf{M}\|_F$ for any matrix $\mathbf{M}$. NA means that the result is not available due to a zero matrix estimate.

References

[1] ANDREW, G., ARORA, R., BILMES, J. and LIVESCU, K. (2013). Deep canonical correlation analysis. In *International Conference on Machine Learning* 1247–1255.

[2] BARCH, D. M., BURGESS, G. C., HARMS, M. P., PETERSEN, S. E., SCHLAGGAR, B. L., CORBETTA, M., GLASSER, M. F., CURTISS, S., DIXIT, S., FELDT, C. et al. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80** 169–189.

[3] BJÖRCK, A. and GOLUB, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **27** 579–594. MR0348991

[4] BUCKNER, R. L., KRIENEN, F. M., CASTELLANOS, A., DIAZ, J. C. and THOMAS YEO, B. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology* **106** 2322-2345.

[5] CAMPBELL, J. D., YAU, C., BOWLBY, R., LIU, Y., BRENNAN, K., FAN, H., TAYLOR, A. M., WANG, C., WALTER, V., AKBANI, R. et al. (2018). Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell reports* **23** 194–212.

[6] CARROLL, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In *Proc. Am. Psychol. Ass.* 227–228.

[7] CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050

[8] CRAWFORD, K. L., NEU, S. C. and TOGA, A. W. (2016). The image and data archive at the laboratory of neuro imaging. *Neuroimage* **124** 1080–1083.

[9] DEZA, M. M. and DEZA, E. (2014). Distances on Numbers, Polynomials, and Matrices. In *Encyclopedia of Distances* 227–244. Springer. MR3243690

[10] DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association* **112** 1211–1220. MR3735371
[11] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall. MR1270903
[12] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B.* **75** 603–680. MR3091653
[13] Feng, Q., Jiang, M., Hannig, J. and Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis* **166** 241–265. MR3799646
[14] Fukumizu, K., Bach, F. R. and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research* **8** 361–383. MR2320675
[15] Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika* **40** 33–51. MR0405725
[16] Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes problems* **30**. Oxford University Press. MR2051013
[17] Harman, H. H. (1976). *Modern Factor Analysis*, Third, revised ed. U of Chicago Press. MR0400546
[18] Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V. et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173** 291–304.
[19] Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge. MR1288752
[20] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.
[21] Huang, H. (2017). Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research* **18** 1–21. MR3655310
[22] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.
[23] Jensen, M. A., Ferretti, V., Grossman, R. L. and Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130** 453–459.
[24] Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–451. MR0341750
[25] Kishore Kumar, N. and Schneider, J. (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra* **65** 2212–2244. MR3740692
[26] Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

[27] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768

[28] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37** 4254–4278. MR2572459

[29] Lock, E. and Dunson, D. (2013). Bayesian consensus clustering. *Bioinformatics* **29** 2610–16.

[30] Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics* **7** 523–542. MR3086429

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

[32] Lu, Y., Huang, K. and Liu, C.-L. (2016). A fast projected fixed-point algorithm for large graph matching. *Pattern Recognition* **60** 971–982.

[33] Mai, Q. and Zhang, X. (2019). An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics* **75** 734–744. MR4012080

[34] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50** 163–170.

[35] Moakher, M. and Batchelor, P. G. (2006). Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields* 285–298. Springer. MR2210524

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

[37] Olivetti, E., Sharmin, N. and Avesani, P. (2016). Alignment of tractograms as graph matching. *Frontiers in Neuroscience* **10** 554.

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

[39] Papadias, C. B. (2000). Globally convergent blind source separation based on a multiuser kurtosis maximization criterion. *IEEE Transactions on Signal Processing* **48** 3508–3519. MR1848829

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

[41] Parra, L. and Sajda, P. (2003). Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research* **4** 1261–1269. MR2103629

[42] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A* **135** 185–198.

[43] Saeed, U., Compagnone, J., Aviv, R. I., Strafella, A. P.,

Black, S. E., Lang, A. E. and Masellis, M. (2017). Imaging biomarkers in Parkinson's disease and Parkinsonian syndromes: current and emerging concepts. *Translational Neurodegeneration* **6** 8.

[44] Schouteden, M., Van Deun, K., Pattyn, S. and Van Mechelen, I. (2013). SCA with rotation to distinguish common and distinctive information in linked data. *Behavior Research Methods* **45** 822–833.

[45] Shu, H., Wang, X. and Zhu, H. (2020). D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *J. Am. Stat. Assoc.* **115** 292–306. MR4078464

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

[47] Smilde, A. K., Westerhuis, J. A. and de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics* **17** 323–337.

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

[49] Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika* **76** 257. MR2788885

[50] Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science* **1** 144–160. MR3949704

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

[52] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H. et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* **80** 62–79.

[53] Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics* **45** 1342–1374. MR3662457

[54] Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58** 236–244. MR0148188

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

[56] Yin, Y.-Q., Bai, Z.-D. and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theory Rel.* **78** 509–521. MR0950344

[57] Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006

[58] Zhou, G., Cichocki, A., Zhang, Y. and Mandic, D. P. (2016). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **27** 2426–2439. MR3571617