



Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(529): 292–306. doi:10.1080/01621459.2018.1543599.

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

Hai Shu^a, Xiao Wang^b, Hongtu Zhu^{a,c,*}

^aDepartment of Biostatistics, The University of Texas MD Anderson Cancer Center

^bDepartment of Statistics, Purdue University

^cDepartment of Biostatistics, The University of North Carolina at Chapel Hill

Abstract

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

Keywords

approximate factor model; canonical variable; common structure; distinctive structure; soft thresholding

1 Introduction

Many large biomedical studies have collected high-dimensional genetic and/or imaging data and associated data (e.g., clinical data) from increasingly large cohorts to delineate the complex genetic and environmental contributors to many diseases, such as cancer and Alzheimer's disease. For example, The Cancer Genome Atlas (TCGA; Koboldt et al., 2012) project collected human tumor specimens and derived different types of large-scale genomic data such as mRNA expression and DNA methylation to enhance the understanding of

* Address for correspondence and reprints: Hongtu Zhu, Ph.D., htzhu@email.unc.edu; Phone No: 919-929-9010.

cancer biology and therapy. The Human Connectome Project (Van Essen et al., 2013) acquired imaging datasets from multiple modalities (HARDI, R-fMRI, T-fMRI, MEG) across a large cohort to build a “network map” (connectome) of the anatomical and functional connectivity within the healthy human brain. These cross-platform datasets share some common information, but individually contain distinctive patterns. Disentangling the underlying common and distinctive patterns is critically important for facilitating the integrative and discriminative analysis of these cross-platform datasets (van der Kloet et al., 2016; Smilde et al., 2017).

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

$$Y_k = C_k + D_k + E_k \quad \text{for } k = 1, 2, \quad (1)$$

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

Most decomposition methods for model (1) are based on the Euclidean space (\mathbb{R}^n, \cdot) endowed with the dot product. Such methods include JIVE (Lock et al., 2013), angle-based JIVE (AJIVE; Feng et al., 2018), OnPLS (Trygg, 2002; Löfstedt and Trygg, 2011), COBE (Zhou et al., 2016), and DISCO-SCA (Schouteden et al., 2014). A common characteristic among all these methods is to enforce the row-space orthogonality between the common and distinctive matrices within each dataset, that is, $C_k D_k^T = 0$ for $k=1, 2$. With the exception of OnPLS, these methods impose additional orthogonality across the datasets, that is, $C_k D_\ell^T = 0$ for all k and ℓ . A potential issue associated with these methods is that they inadequately consider the more desired orthogonality between the distinctive matrices D_1 and D_2 , which guarantees that no common structure is retained therein. Specifically, the first four methods do not impose any orthogonality constraint between D_1 and D_2 . Although DISCO-SCA and a modified JIVE (O’Connell and Lock (2016); denoted as R.JIVE) have considered the row-space orthogonality between the distinctive matrices, it may be incompatible with their orthogonal condition that $C_k D_\ell^T = 0$ for all $k, \ell = 1, 2$ even as $p_1 = p_2 = 1$.

Rather than the conventionally used Euclidean space (\mathbb{R}^n, \cdot) , the aim of this paper is to develop a new decomposition method for model (1) based on the inner product space $(\mathcal{L}_0^2, \text{cov})$, which is the vector space composed of all zero-mean and finite-variance real-valued

random variables and endowed with the covariance operator as the inner product. Specifically, model (1) is a sample-matrix version of the prototype given by

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

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

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

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

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

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

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

Motivated by the toy example above, we introduce a novel method, decomposition-based canonical correlation analysis (D-CCA), which generalizes the classical canonical correlation analysis (CCA; Hotelling, 1936) by further separating common vectors $\{c_k\}_{k=1}^2$ and distinctive vectors $\{d_k\}_{k=1}^2$ between signal vectors $\{x_k\}_{k=1}^2$ subject to constraints (3)-(5). In contrast, classical CCA only seeks the association between two random vectors by sequentially determining the mutually orthogonal pairs of canonical variables that have maximal correlations between the vector spaces respectively spanned by entries of the two random vectors. Another related but different method, the sparse CCA (Chen et al., 2013; Gao et al., 2015, 2017), focuses on the sparse linear combinations of original variables for representing canonical variables with improved interpretability, which is neither required nor pursued by our D-CCA.

The “low-rank plus noise” model $y_k = x_k + e_k$ for each single k can be naturally formulated by a factor model as $y_k = Bf_k + e_k$, where the latent factor f_k is an orthonormal basis of $\text{span}(x_k^T)$ with B being the coefficient matrix. In factor model analysis (Bai and Ng, 2008), $x_k = Bf_k$ is called the “common component”, and e_k the “idiosyncratic error”. These two terms should not be confused with our considered common vectors $\{c_k\}_{k=1}^2$ and distinctive vectors $\{d_k\}_{k=1}^2$ that are solely based on signals $\{x_k\}_{k=1}^2$ excluding noises $\{e_k\}_{k=1}^2$. For general dynamic factor models (Forni et al., 2000), Hallin and Liška (2011) proposed a joint decomposition method, which divides each dataset into strongly common, weakly common, weakly idiosyncratic, and strongly idiosyncratic components (also see Forni et al. (2017) and Barigozzi et al. (2018)). Applying their method to our considered scenarios with no temporal dependence, and additionally assuming no correlations between signals $\{x_k\}_{k=1}^2$, and noises $\{e_k\}_{k=1}^2$, then for each y_k , x_k is the sum of strongly common and weakly common components, e_k is the strongly idiosyncratic component, and no weakly idiosyncratic component exists. One may treat their strongly common and weakly common components as the common vector c_k and the distinctive vector d_k , respectively, but the desired orthogonality (4) is still not imposed. Especially when $\text{span}(x_1^T) \cap \text{span}(x_2^T) = \{0\}$, x_k is entirely a weakly common component, and thus the orthogonality (4) fails for the toy example shown in Figure 1. See Remark S.1 in the supplementary material for more detailed discussions.

Our major contributions of this paper are as follows. The proposed D-CCA method appropriately decomposes each paired canonical variables of signal vectors x_1 and x_2 into a common variable and two orthogonal distinctive variables, and then collects all of them to form the common vector c_k and the distinctive vector d_k for each x_k . The common matrix C_k and the distinctive matrix D_k are defined with columns as n realizations of c_k and d_k , respectively. Three challenging issues that arise in estimating the low-rank matrices defined by D-CCA are high dimensionality, the corruption of signal random vectors by unobserved noises, and the unknown signal covariance and cross-covariance matrices that are needed in CCA. To address these issues, we study the considered “low-rank plus noise” model under the framework of approximate factor models (Wang and Fan, 2017), and develop a novel

estimation approach by integrating the S-POET method for spiked covariance matrix estimation (Wang and Fan, 2017) and the construction of principal vectors (Björck and Golub, 1973). Under some mild conditions, we systematically investigate the consistency and convergence rates of the proposed matrix estimators under a high-dimensional setting with $\min(p_1, p_2) > \kappa_0 n$ for a positive constant κ_0 .

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

Here, we introduce some notation. For a real matrix $M = (M_{ij})_{1 \leq i \leq p, 1 \leq j \leq n}$, the ℓ th largest singular value and the ℓ th largest eigenvalue (if $p = n$) are respectively denoted by $\sigma_\ell(M)$ and $\lambda_\ell(M)$, the spectral norm $\|M\|_2 = \sigma_1(M)$, the Frobenius norm $\|M\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^n M_{ij}^2}$, and the matrix \mathcal{L}^∞ norm $\|M\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^n |M_{ij}|$. We use $M^{[s:t, u:v]}$, $M^{[s:t, :]}$ and $M^{[:, u:v]}$ to represent the submatrices $(M_{ij})_{s \leq i \leq t, u \leq j \leq v}$, $(M_{ij})_{s \leq i \leq t, 1 \leq j \leq n}$ and $(M_{ij})_{1 \leq i \leq p, u \leq j \leq v}$ of the $p \times n$ matrix M , respectively. Denote the Moore-Penrose pseudoinverse of matrix M by M^\dagger . Define $0_{p \times n}$ to be the $p \times n$ zero matrix and $I_{p \times p}$ to be the $p \times p$ identity matrix. Denote $\text{diag}(M_1, \dots, M_m)$ to be a block diagonal matrix with M_1, \dots, M_m as its main diagonal blocks. For signal vectors \mathbf{x}_k 's, denote $\Sigma_k = \text{cov}(\mathbf{x}_k)$, $\Sigma_{12} = \text{cov}(\mathbf{x}_1, \mathbf{x}_2)$, $r_k = \text{rank}(\Sigma_k)$, $r_{\min} = \min(r_1, r_2)$, $r_{\max} = \max(r_1, r_2)$ and $r_{12} = \text{rank}(\Sigma_{12})$. For a subspace B of a vector space A , denote its orthogonal complement in A by $A \setminus B$. We write $a \propto b$ if a is proportional to b , i.e., $a = \kappa b$ for some constant κ . Throughout the paper, our asymptotic arguments are by default under $n \rightarrow \infty$. We reserve $\{c, c_j\}$, $\{c_k\}$ and $\{C_k\}$ for the common variables, common vectors and common matrices, respectively, and use other notation for constants, e.g., κ_0 .

2 The D-CCA Method

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

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

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

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k + \mathbf{B}_k \mathbf{F}_k + \mathbf{E}_k, \quad (7)$$

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

2.1 Definition of common and distinctive matrices

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

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

$$\mathbf{x}_k = \sum_{\ell=1}^{r_k} \beta_{k\ell} z_{k\ell} = \mathbf{c}_k + \mathbf{d}_k := \sum_{\ell=1}^{L_{12}} \beta_{k\ell}^{(C)} c_\ell + \sum_{\ell=1}^{L_k} \beta_{k\ell}^{(D)} d_{k\ell}, \quad (8)$$

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

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

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

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

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

where $\text{span}(\mathbf{x}_k^{\top}) \setminus \text{span}(\{z_{km}\}_{m=1}^{\ell-1}) := \text{span}(\mathbf{x}_k^{\top})$. Variables $\{z_{k\ell}\}_{k=1}^2$ are called the ℓ th pair of canonical variables, and their correlation is the ℓ th canonical correlation of \mathbf{x}_1 and \mathbf{x}_2 . Augment $\{z_{k\ell}\}_{k=1}^2$ with any $(r_k - r_{12})$ standardized variables to be $\mathbf{z}_k = (z_{k1}, \dots, z_{krk})^{\top}$ such that \mathbf{z}_k^{\top} is an orthonormal basis of $\text{span}(\mathbf{x}_k^{\top})$. A detailed procedure to obtain a solution of $\{z_k\}_{k=1}^2$ will be presented later after Theorem 2. An important property of these augmented canonical variables is the bi-orthogonality shown in the following theorem.

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

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

where Λ_{12} is a $r_{12} \times r_{12}$ nonsingular diagonal matrix.

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

$$c_{\ell} \propto \arg \max_{w \in (\mathcal{L}_{\ell}^2, \text{cov})} \{ \text{corr}^2(z_{1\ell}, w) + \text{corr}^2(z_{2\ell}, w) \} \tag{11}$$

with the constraints

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

Constraints (12) and (13) are actually the special case of (8) and (9) for two standardized random variables. Constraint (14) indicates that c_{ℓ} explains more variances of $z_{1\ell}$ and $z_{2\ell}$ when their correlation ρ_{ℓ} increases. Although ρ_{ℓ} here referring to the ℓ th canonical correlation of $\{\mathbf{x}_k\}_{k=1}^2$, is always positive for $1 \leq \ell \leq r_{12}$, we include $\rho_{\ell} = 0$ to enable (11) as

a general optimization problem for any two standardized variables with nonnegative correlation. The unique solution of (11) is given by

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

where $\theta_\ell = \arccos(\rho_\ell)$ is the angle between $z_{1\ell}$ and $z_{2\ell}$ in $(\mathcal{L}_0^2, \text{cov})$. More desirable than constraint (14), it easily follows from (15) that $\text{var}(c_\ell)$ is a continuous and strictly monotonic increasing function for $\rho_\ell \in [0, 1]$. We defer the detailed derivation of (15) to the supplementary material (Proposition S.2 and its proof). This solution is geometrically illustrated in Figure 1 with ℓ omitted in the subscriptions. Simply let $d_{k\ell} = z_{k\ell}$ for $r_{12} + 1 - \ell = r_k$. The two-step optimization strategy arrives at the following decomposition of form (8): For $k = 1, 2$,

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

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

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

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

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

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

We now present a procedure to obtain the augmented canonical variables $\{z_1, z_2\}$. For $k = 1, 2$, let a singular value decomposition (SVD) of Σ_k be $\Sigma_k = \mathbf{V}_k \Lambda_k \mathbf{V}_k^T$, where $\Lambda_k = \text{diag}(\sigma_1$

$(\Sigma_k), \dots, \sigma_{r_k}(\Sigma_k)$ and V_k is a $p_k \times r_k$ matrix with orthonormal columns. Let $z_k^* = \Lambda_k^{-1/2} V_k^T \mathbf{x}_k$, then we have $\text{cov}(z_k^*) = I_{r_k \times r_k}$. Define

$$\Theta = \text{cov}(z_1^*, z_2^*) = \Lambda_1^{-1/2} V_1^T \Sigma_{12} V_2 \Lambda_2^{-1/2}.$$

The rank of Θ is also r_{12} . Denote a full SVD of Θ by $\Theta = U_{\theta 1} \Lambda_{\theta} U_{\theta 2}^T$, where $U_{\theta 1}$ and $U_{\theta 2}$ are two orthogonal matrices, and Λ_{θ} is a $r_1 \times r_2$ rectangular diagonal matrix for which the main diagonal is $(\sigma_1(\Theta), \dots, \sigma_{r_{12}}(\Theta), \mathbf{0}_{1 \times (r_{\min} - r_{12})})$. We then define

$$z_k = U_{\theta k}^T z_k^* = \Gamma_k^T \mathbf{x}_k \quad \text{with} \quad \Gamma_k := V_k \Lambda_k^{-1/2} U_{\theta k}, \tag{18}$$

which satisfies $\text{cov}(z_k) = I_{r_k \times r_k}$ and $\text{corr}(z_1, z_2) = \Lambda_{\theta}$. Note that $\sigma_{\ell}(\Theta) = \rho_{\ell}$ for $\ell = 1, \dots, r_{12}$ are the canonical correlations between \mathbf{x}_1 and \mathbf{x}_2 .

Now look back to $c_k = \sum_{\ell=1}^{r_{12}} \beta_{k\ell} c_{\ell}$ that is defined in (16). Plugging (17) and (15) for $\beta_{k\ell}$ and c_{ℓ} in the formula together with $\{z_j\}_{j=1}^2 = 1$, given in (18), we obtain

$$c_k = \text{cov}(\mathbf{x}_k, z_k^{[1:r_{12}]}) A_C \sum_{j=1}^2 z_j^{[1:r_{12}]}, \tag{19}$$

where $A_C = \text{diag}(a_1, \dots, a_{r_{12}})$ and $a_{\ell} = \frac{1}{2} \left[1 - \left(\frac{1 - \sigma_{\ell}(\Theta)}{1 + \sigma_{\ell}(\Theta)} \right)^{1/2} \right]$ for $\ell = 1, \dots, r_{12}$. Replacing random vector $z_k = \Gamma_k^T \mathbf{x}_k$ by its sample matrix $Z_k := \Gamma_k^T X_k$ in the rightmost of (19) yields

$$C_k = \text{cov}(\mathbf{x}_k, z_k^{[1:r_{12}]}) A_C \sum_{j=1}^2 z_j^{[1:r_{12}, :]}. \tag{20}$$

This equation is useful to our design of estimators for C_k and $D_k = X_k - C_k$ in the next subsection.

2.2 Estimation of D-CCA matrices

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

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

$$Y_k = U_{k1} \Lambda_{y_k} U_{k2}^T, \tag{21}$$

where U_{k1} and U_{k2} are two orthogonal matrices and Λ_{y_k} is a rectangular diagonal matrix with the singular values in decreasing order on its main diagonal. The matrix \tilde{X}_k is then obtained via soft-thresholding the singular values of Y_k by

$$\tilde{X}_k = U_{k1}^{[1:r_k]} \text{diag}(\hat{\sigma}_1^S(Y_k), \dots, \hat{\sigma}_{r_k}^S(Y_k)) (U_{k2}^{[1:r_k]})^T, \tag{22}$$

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

We next use \tilde{X}_k to develop estimators for C_k in (20) and $D_k = X_k - C_k$. Define the estimators of Σ_k and Σ_{12} as $\hat{\Sigma}_k = n^{-1} \tilde{X}_k \tilde{X}_k^T$ and $\hat{\Sigma}_{12} = n^{-1} \tilde{X}_1 \tilde{X}_2^T$, respectively. Then, based on $\hat{\Sigma}_k$ and $\hat{\Sigma}_{12}$, we obtain estimators $\hat{V}_k, \hat{\Lambda}_k, \hat{U}_{\theta k} = \text{diag}(\hat{U}_{\theta k}^{[1:\tilde{r}_k, 1:\tilde{r}_k]}, I_{(r_k - \tilde{r}_k) \times (r_k - \tilde{r}_k)})$ and $\hat{\Lambda}_\theta$ in the same way as their true counterparts $V_k, \Lambda_k, U_{\theta k}$ and Λ_θ with a $r_1 \times r_2$ matrix $\hat{\Theta} := (\hat{\Lambda}_1^\dagger)^{1/2} \hat{V}_1^T \hat{\Sigma}_{12} \hat{V}_2 (\hat{\Lambda}_2^\dagger)^{1/2}$. Define $\hat{Z}_k^* = (\hat{\Lambda}_k^\dagger)^{1/2} \hat{V}_k^T \tilde{X}_k$ and $\hat{Z}_k = \hat{U}_{\theta k}^T \hat{Z}_k^*$. We have

$$n^{-1} \hat{Z}_k^* (\hat{Z}_k^*)^T = n^{-1} \hat{Z}_k (\hat{Z}_k)^T = \text{diag}(I_{\tilde{r}_k} \times \tilde{r}_k, 0_{(r_k - \tilde{r}_k) \times (r_k - \tilde{r}_k)}),$$

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

From Theorem 1 in Björck and Golub (1973), it follows that $n^{-1/2} \hat{Z}_1^{[\ell, :]}$ and $n^{-1/2} \hat{Z}_2^{[\ell, :]}$ for ℓ r_{\min} are the principal vectors of the row spaces of \tilde{X}_1 and \tilde{X}_2 , and moreover, $\sigma_\ell(\hat{\Theta}) \leq 1$. Let $\hat{A}_C^{(r)} = \text{diag}(\hat{a}_1, \dots, \hat{a}_r)$ with $\hat{a}_\ell = \frac{1}{2} \left[1 - \left(\frac{1 - \sigma_\ell(\hat{\Theta})}{1 + \sigma_\ell(\hat{\Theta})} \right)^{1/2} \right]$ for $\ell \leq \tilde{r}_{12} := \text{rank}(\hat{\Theta})$ and otherwise $\hat{a}_\ell = 0$. Define estimators of C_k, D_k and X_k by

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

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

and

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

Here, we substitute \hat{X}_k for \tilde{X}_k as the estimator of X_k . The latter can be written as

$$\tilde{X} = \hat{C}_k^{(\tilde{r}_{12})} + \hat{D}_k \tag{26}$$

with

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

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

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

Throughout our estimation construction, the key idea is to develop a good estimator of $\{X_k, Z_k\}_{k=1}^2$. Thus, the S-POET method (Wang and Fan, 2017) may be replaced by any other good approach, but with possibly different assumptions. For example, given the cleaned signal data X_k 's, Chen et al. (2013) and Gao et al. (2015, 2017) showed that sparse CCA algorithms can consistently estimate the canonical coefficient matrix Γ_k for $Z_k = \Gamma_k^T X_k$ by imposing certain sparsity on Γ_k 's and that all eigenvalues of $\text{cov}(x_k)$ are bounded from above and below by positive constants. These two conditions are not assumed for our proposed method. In particular, their bounded eigenvalue condition contradicts our low-rank structure of signal x_k that introduces the spiked covariance matrix $\text{cov}(y_k)$. The sparse CCA algorithms need the cleaned signal data X_k 's available beforehand. Alternatively, they may be directly applicable to the observable data Y_k 's by assuming zero E_k 's, if the bounded eigenvalue condition holds for $\text{cov}(y_k)$. For the TCGA datasets in our real-data application, the scree plots given later in Figure 6 favorably suggest our spiked eigenvalue assumption. Moreover, the approximate factor model with spiked covariance structure has been widely used in various fields such as signal processing (Nadakuditi and Silverstein, 2010) and machine learning (Huang, 2017), and fits the low-rank plus noise structure considered in the six competing methods mentioned in Section 1. Our paper hence focuses on this spiked covariance model and leaves the extension to sparse CCA models for future research.

2.3 Rank selection

In practice, matrix ranks r_1 , r_2 and r_{12} are usually unknown and need to be determined. There is a rich literature on determining r_k , $k \in \{1, 2\}$, which is the number of latent factors for the high-dimensional approximate factor model. Examples of consistent estimators include but are not limited to Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). Several heuristic approaches for selecting r_{12} , the number of nonzero canonical

correlations for the high-dimensional CCA, have been proposed by Song et al. (2016). In this paper, we apply the edge distribution (ED) method of Onatski (2010) to determine r_k for $k = 1, 2$ by

$$\hat{r}_k = \max\{\ell \leq T_k: \hat{\lambda}_{k\ell} - \hat{\lambda}_{k,\ell+1} \geq \delta\}, \quad (28)$$

where $\hat{\lambda}_{k\ell}$ is the ℓ -th eigenvalue of $Y_k Y_k^\top / n$. The upper bound is chosen as

$T_k = \min(\#\{i \mid \hat{\lambda}_{ki} \geq \frac{1}{m_k} \sum_{\ell=1}^{m_k} \hat{\lambda}_{k\ell}\}, m_k/10)$ with $m_k = \min(n, p_k)$ which is recommended by

Ahn and Horenstein (2013), and parameter δ is calibrated as in Section IV of Onatski (2010). It is believed that $r_{12} > 0$ if two variables from different cleaned datasets have a significant nonzero correlation detected by, e.g., the normal approximation test of DiCiccio and Romano (2017). Otherwise, it is unnecessary to conduct the proposed matrix decomposition. We select the nonzero r_{12} by using the minimum description length information-theoretic criterion (MDL-IC) proposed by Song et al. (2016):

$$\hat{r}_{12} = \arg \min_{r \in [1, \min(\hat{r}_1, \hat{r}_2)]} \left\{ n \sum_{l=1}^r \log(1 - s_l^2) + r(r_1 + r_2 - r) \log(n) \right\}, \quad (29)$$

where s_l is the ℓ -th singular value of $(U_{12}^{[1:\hat{r}_1]})^\top U_{22}^{[1:\hat{r}_2]}$ with U_{12} and U_{22} defined in (21).

The ranks r_1 , r_2 , and r_{12} determined by (28) and (29) perform well in our numerical studies.

3 Theoretical Properties of D-CCA Estimators

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

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

- (I) Let $\lambda_{k1} > \dots > \lambda_{k,r_k} > \lambda_{k,r_k+1} \dots > \lambda_{k,p_k} > 0$ be the eigenvalues of $\text{cov}(\mathbf{y}_k)$. There exist positive constants κ_1 , κ_2 and δ_0 such that $\kappa_1 \lambda_{k\ell} \kappa_2$ for $\ell > r_k$ and $\min_{\ell > r_k} (\lambda_{k\ell} - \lambda_{k,\ell+1}) / \lambda_{k\ell} \geq \delta_0$.
- (II) Assume $p_k > \kappa_0 n$ with a constant $\kappa_0 > 0$. When $n \rightarrow \infty$, assume $\lambda_{k,r_k} \rightarrow \infty$, $p_k / (n \lambda_{k,r_k})$ is upper bounded for $\ell > r_k$, $\lambda_{k1} / \lambda_{k,r_k}$ is bounded from above and below, and $\sqrt{p_k} (\log n)^{1/\gamma_{k2}} = o(\lambda_{r_k})$ with γ_{k2} given in (V) below.
- (III) The columns of $Z_k^{(y)} := (\Lambda_k^{(y)})^{-1/2} (\mathbf{V}_k^{(y)})^\top \mathbf{Y}_k$ are i.i.d. copies of $z_k^{(y)} := (\Lambda_k^{(y)})^{-1/2} (\mathbf{V}_k^{(y)})^\top \mathbf{y}_k$, where $\mathbf{V}_k^{(y)} \Lambda_k^{(y)} (\mathbf{V}_k^{(y)})^\top$ is the full SVD of $\text{cov}(\mathbf{y}_k)$ with $\Lambda_k^{(y)} = \text{diag}(\lambda_{k1}, \dots, \lambda_{k,p_k})$. The entries of $z_k^{(y)}, z_{k1}^{(y)}, \dots, z_{k,p_k}^{(y)}$ are independent with $\mathbb{E}(z_{ki}^{(y)}) = 0$, $\text{var}(z_{ki}^{(y)}) = 1$, and the sub-Gaussian norm $\sup_{q \geq 1} q^{-1/2} (\mathbb{E} |z_{ki}^{(y)}|^q)^{1/q} \leq K$ with a constant $K > 0$ for all $i \leq p_k$.

- (IV) The matrix $B_k^\top B_k$ is a diagonal matrix, and $|B_k^{[i, \ell]}| \leq M\sqrt{\lambda_{k\ell} p_k}$ with a constant $M > 0$ holds for all $i \leq p_k$ and $\ell \leq r_k$.
- (V) Denote $e_k = (e_{k1}, \dots, e_{kp_k})^\top$ and $f_k = (f_{k1}, \dots, f_{kr_k})^\top$. Assume $\|\text{cov}(e_k)\|_\infty < s_0$ with a constant $s_0 > 0$. For all $i \leq p_k$ and $\ell \leq r_k$, there exist positive constants γ_{k1} , γ_{k2} , b_{k1} and b_{k2} such that for $t > 0$, $\mathbb{P}(|e_{ki}| > t) \leq \exp(-(t/b_{k1})^{\gamma_{k1}})$ and $\mathbb{P}(|f_{k\ell}| > t) \leq \exp(-(t/b_{k2})^{\gamma_{k2}})$.

Assumption 1 follows assumptions 2.1-2.3 and 4.1-4.2 of Wang and Fan (2017) which guarantee desirable performance of the initial signal estimators \tilde{X}_k 's defined in (22). The diverging leading eigenvalues of $\text{cov}(y_k)$ assumed in conditions (I) and (II), together with the approximate sparsity constraint $\|\text{cov}(e_k)\|_\infty < s_0$ in condition (V), indicate the necessity of sufficiently strong signals for soft-thresholding. Although Wang and Fan (2017) considered $p > n$, it is not difficult to relax it to $p_k > \kappa_{0n}$, as given in our condition (II). A random variable is said to be sub-Gaussian if its sub-Gaussian norm is bounded (Vershynin, 2012). Condition (III) imposes the sub-Gaussianity on all entries of $z_k^{(y)}$ with a uniform bound.

Simply letting $f_k = z_k^*$ can lead to a diagonal matrix $B_k^\top B_k$ that is required by condition (IV). In condition (V), the approximately sparse constraint is imposed on $\text{cov}(e_k)$ rather than E_k . See Wang and Fan (2017) and also Fan et al. (2013) for more detailed discussions of the above assumption.

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

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

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

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

$$\frac{\|\hat{C}_k - C_k\|(\cdot)}{\|C_k\|(\cdot)} = O_P\left(\frac{\Delta}{\alpha_{C_k,(\cdot)}}\right), \frac{\|\hat{D}_k - D_k\|(\cdot)}{\|D_k\|(\cdot)} = O_P\left(\frac{\Delta}{\alpha_{D_k,(\cdot)}}\right), \frac{\|\hat{X}_k - X_k\|(\cdot)}{\|X_k\|(\cdot)} = O_P(\Delta),$$

and the error bound of canonical correlation estimators

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

Provided that matrix ranks r_1, r_2 and r_{12} are correctly selected, Theorem 3 shows the consistency of the proposed matrix estimators in the relative errors that are the norms of

estimation errors divided by the norms of true matrices, with associated convergence rates. The ratios $\alpha_{C_k,(\cdot)}$ and $\alpha_{D_k,(\cdot)}$ in the convergence rates of \widehat{C}_k and \widehat{D}_k can be removed if the relative errors are instead scaled by the norms of the signal matrices.

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

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

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

$$\frac{\|\widehat{X}_k^{(r)} - X_k\|(\cdot)}{\|X_k\|(\cdot)} = O_P(\Delta), \text{ and } \max_{1 \leq \ell \leq r_{\min}} |\sigma_\ell(\widehat{\Theta}) - \sigma_\ell(\Theta)| = O_P(\delta_\Theta).$$

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

4 Simulation Studies

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

- Setup 1: Let $\mathbf{x}_1 \stackrel{d}{=} \mathbf{x}_2$, with $r_1 = 3$, $r_{12} = 1$, and $\lambda(\Sigma_1) = 500 - 200(\ell - 1)$ for $\ell = 3$. Set $z_{k1}, z_{k2}, z_{k3} \stackrel{i.i.d.}{\sim} N(0, 1)$ for each $k = 1, 2$. Randomly generate V_1 with orthonormal columns, which is the same for all replications. Let $\mathbf{x}_k = V_1 \Lambda_1^{1/2} z_k$. Generate $e_{ki}, k \leq 2, i \leq p_k \stackrel{i.i.d.}{\sim} N(0, \sigma_e^2)$ that are independent of $\{\mathbf{x}_k\}_{k=1}^2$. Vary dimension p_1 from 100 to 1,500, the first canonical angle $\theta_1 = \arccos(\rho_1)$ from 0° to 75° with $\rho_1 = \text{corr}(z_{11}, z_{21})$, and the noise variance σ_e^2 from 0.01 to 16.
- Setup 2: Use the same settings for \mathbf{x}_1 and $\{e_k\}_{k=1}^2$ as in Setup 1. For \mathbf{x}_2 , fix $p_2 = 300$, and set $r_2 = 5$ and $\lambda(\Sigma_2) = 500 - 100(\ell - 1)$ for $\ell = 5$. Simulate $\mathbf{x}_2 = V_2 \Lambda_2^{1/2} z_2$

with $z_{21}, \dots, z_{25} \stackrel{i.i.d.}{\sim} N(0, 1)$ and a randomly generated V_2 that is the same for all replications. Let $r_{12} = 1$. Vary p_1 , θ_1 and σ_e^2 according to Setup 1.

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

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

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

Consider Figure 3 of Setup 1 as an example. We have nearly identical plots for the two datasets that are generated from the same distribution. From the first row, where all considered cases have almost the same set of average ratios $\{\alpha_{Ck(\cdot)}, \alpha_{Dk(\cdot)}\}$, all the AREs become bigger as the dimension p_1 increases. For the second row, the increasing canonical angle θ_1 results in a change in the average ratios $\alpha_{Ck,2}$ from 0.997 down to 0.18 and in $\alpha_{c_k,F}$ from 0.74 down to 0.14; $\alpha_{Dk,2}$ is stable around 0.78 for the first 5 values of θ_1 and then increases to 0.87 at $\theta_1 = 75^\circ$; and $\alpha_{Dk,F}$ changes from 0.67 to 0.93. Meanwhile, this leads to increasing AREs of \hat{C}_k and decreasing AREs of \hat{D}_k , but does not affect the AREs of \hat{X}_k . The third row shows that all the AREs increase as the noise variance σ_e^2 becomes bigger. Note that increasing σ_e^2 is equivalent to decreasing the eigenvalues of Σ_k by scaling σ_e^2 to 1. These results agree with the influence of p_1, α and $\lambda_1(\Sigma_k)$ on the convergence rates given in Theorem 3.

For Setup 2, with similar arguments, we find a similar pattern of estimation performance for D-CCA, as shown in the second and third rows and the plots of the first dataset in the first row of Figure 4. For the first row of Figure 4, the considered cases of the second dataset have a fixed dimension p_2 and stable ratios $\{\alpha_{C2(\cdot)}, \alpha_{D2(\cdot)}\}$. The corresponding AREs are

still acceptable and interestingly are not much impacted by the change in the dimension p_1 of the first dataset. From Table 1, we see that the estimated canonical angles and correlations perform well for Setups 1 and 2 even in the presence of strong noise levels.

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

Now consider the GDFM method (Hallin and Liška, 2011). We set the sample temporal cross-covariances to be zero in GDFM estimation for our simulated data and TCGA datasets that have no temporal dependence. GDFM decomposes each data matrix by $\mathbf{Y}_k = \boldsymbol{\chi}_k^* + \boldsymbol{\xi}_k^* = (\boldsymbol{\phi}_k + \boldsymbol{\psi}_k) + (\boldsymbol{\nu}_k + \boldsymbol{\xi}_k^*) = \boldsymbol{\chi}_k + \boldsymbol{\xi}_k$ with each component's name shown in Table 6. By Remark S.1 (in the supplementary material), theoretically for our simulated i.i.d. data with no correlations between signals and noises, the weakly idiosyncratic matrix $\boldsymbol{\nu}_k$ is zero, and the joint common matrix $\boldsymbol{\chi}_k^*$ and the marginal common matrix $\boldsymbol{\chi}_k$ are both equal to the signal matrix \mathbf{X}_k . Moreover, the strongly common matrix $\boldsymbol{\phi}_k$ is zero, when $\text{span}(\mathbf{x}_1^T) \cap \text{span}(\mathbf{x}_2^T) = \{0\}$, i.e., the first canonical correlation ρ_1 between \mathbf{x}_1 and \mathbf{x}_2 is smaller than 1. The above theoretical results are evidenced by our simulations. In Table 2, the relative errors of estimators $\hat{\boldsymbol{\chi}}_k^*$ and $\hat{\boldsymbol{\chi}}_k$ to signal \mathbf{X}_k are as comparably small as those of $\hat{\mathbf{X}}_k$ by our D-CCA and the other five well performed methods. The similarly small norm ratios of $\hat{\boldsymbol{\nu}}_k$ to $\hat{\boldsymbol{\chi}}_k^*$ numerically support $\boldsymbol{\nu}_k = 0$. The squares of these quantities are much smaller, and especially in the Frobenius norm are equivalent to matrix-variation ratios. The strongly common matrix estimate $\hat{\boldsymbol{\phi}}_k$ is zero for the setups, with $\rho_1 = 0.707 < 1$, considered in the table. These numerical evidences are more clearly seen in Figure 5(b) under a similar setup.

5 Analysis of TCGA Breast Cancer Data

In this section, we apply the proposed D-CCA method to analyze genomic datasets produced from TCGA breast cancer tumor samples. We investigate the ability to separate tumor subtypes for matrices obtained from D-CCA in comparison to those obtained from the six competing methods as well as GDFM that are mentioned in Section 1. We consider the mRNA expression data and DNA methylation data for a common set of 660 samples. The two datasets are publicly available at <https://tcga-data.nci.nih.gov/docs/publications> and have been respectively preprocessed by Ciriello et al. (2015) and Koboldt et al. (2012). The 660 samples were classified by Ciriello et al. (2015) into 4 subtypes using the PAM50 model (Parker et al., 2009) based on mRNA expression data. Specifically, the samples consist of 112 basal-like, 55 HER2-enriched, 331 luminal A, and 162 luminal B tumors.

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

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

for matrix $\mathbf{A} = (A_{ij})_{p \times n}$, where $\bar{A}_{i,s(j)}$ is the average of the j -th sample's subtype on the i -th row and $\bar{A}_{i\cdot}$ is the average of the i -th row's elements. The SWISS score represents the variation within the subtypes as a proportion of the total variation. A lower score indicates better subtype separation. For the mRNA expression data, we filtered out the subset consisting of the 1,195 variably expressed genes with marginal SWISS > 0.9 from the original 20,533 genes, and denote this subset as EXP90. The 2,083 variably methylated probes of the DNA methylation data, originally with 21,986 probes, are included in the analysis. We denote the 881 probes with marginal SWISS > 0.9 as METH90b and the remaining 1,202 probes as METH90a. We conducted the analysis for the pair of EXP90 and METH90b as well as the pair of EXP90 and METH90a.

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

From Table 5, for the pair of EXP90 and METH90b datasets, the matrix \widehat{X}_k obtained by all the seven methods gains an improved SWISS score compared to the noisy data matrix Y_k . Other than AJIVE and COBE with $\widehat{C}_k = 0$, a clear pattern of increasing SWISS scores, from \widehat{C}_k to \widehat{X}_k and then to \widehat{D}_k , can be seen for the remaining methods except JIVE. This indicates that an enhanced ability to separate the tumor samples by subtype can be expected when integrating two datasets that can exhibit such a distinction to a moderate extent. Also note that the estimated common matrices of our D-CCA have the lowest SWISS scores. While considering the pair of EXP90 and METH90a datasets, for all the seven methods we find a big gap between the SWISS scores of the two estimated signal matrices, and that the denoised matrix of the METH90a dataset still has nearly no discriminative power with SWISS close to 1. The ability on subtype separation seems more likely to be a distinctive feature of EXP90 dataset comparing to METH90a dataset. The estimated distinctive matrix of EXP90 is thus expected to have a lower SWISS score than its estimated common matrix. However, only D-CCA meets this point, except that AJIVE and COBE yield zero common matrices. The failure of the six competing methods may be caused by their inappropriate decomposition constructions, which are mentioned in Section 1. In particular, from Table 3, we see that a lot of significant nonzero correlations exist among all gene-probe pairs based on the estimated distinctive matrices, respectively, obtained by JIVE, OnPLS and DISCO-SCA.

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

6 Discussion

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

There are many possible further studies beyond the current proposed D-CCA. The first is to generalize the D-CCA for three or more datasets. We may assume that at least two datasets have mutually orthogonal distinctive structures. An immediate idea starts from substituting

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Zhu's work was partially supported by NIH grants MH086633 and MH116527, NSF grants SES-1357666 and DMS-1407655, a grant from the Cancer Prevention Research Institute of Texas, and the endowed Bao-Shan Jing Professorship in Diagnostic Imaging. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

References

- Ahn SC and Horenstein AR (2013), "Eigenvalue ratio test for the number of factors," *Econometrica*, 81, 1203–1227.
- Bai J (2003), "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.
- Bai J and Ng S (2002), "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.
- Bai J and Ng S (2008), "Large dimensional factor analysis," *Foundations and Trends in Econometrics*, 3, 89–163.
- Barigozzi M, Hallin M, and Soccorsi S (2018), "Identification of global and local shocks in international financial markets via general dynamic factor models," *Journal of Financial Econometrics*, DOI: 10.1093/jfinec/nby006.
- Benjamini Y and Hochberg Y (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.
- Björck A and Golub GH (1973), "Numerical methods for computing angles between linear subspaces," *Mathematics of Computation*, 27, 579–594.
- Cabanski CR, Qi Y, Yin X, Bair E, Hayward MC, Fan C, Li J, Wilkerson MD, Marron JS, Perou CM, and Hayes DN (2010), "SWISS MADE: Standardized within class sum of squares to evaluate methodologies and dataset elements," *PLoS ONE*, 5, e9905. [PubMed: 20360852]
- Chamberlain G and Rothschild M (1983), "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, 51, 1281–1304.
- Chen M, Gao C, Ren Z, and Zhou HH (2013), "Sparse CCA via precision adjusted iterative thresholding," arXiv preprint arXiv:1311.6186.
- Ciriello G, Gatzka ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandath C, et al. (2015), "Comprehensive molecular portraits of invasive lobular breast cancer," *Cell*, 163, 506–519. [PubMed: 26451490]

- Comon P and Jutten C (2010), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press.
- DiCiccio CJ and Romano JP (2017), “Robust permutation tests for correlation and regression coefficients,” *Journal of the American Statistical Association*, 112, 1211–1220.
- Fan J, Liao Y, and Mincheva M (2013), “Large covariance estimation by thresholding principal orthogonal complements,” *Journal of the Royal Statistical Society: Series B*, 75, 603–680.
- Feng Q, Jiang M, Hannig J, and Marron J (2018), “Angle-based joint and individual variation explained,” *Journal of Multivariate Analysis*, 166, 241–265.
- Forni M, Hallin M, Lippi M, and Reichlin L (2000), “The generalized dynamic-factor model: Identification and estimation,” *Review of Economics and statistics*, 82, 540–554.
- Forni M, Hallin M, Lippi M, and Zaffaroni P (2017), “Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis,” *Journal of Econometrics*, 199, 74–92.
- Fukumizu K, Bach FR, and Gretton A (2007), “Statistical consistency of kernel canonical correlation analysis,” *Journal of Machine Learning Research*, 8, 361–383.
- Gao C, Ma Z, Ren Z, and Zhou HH (2015), “Minimax estimation in sparse canonical correlation analysis,” *The Annals of Statistics*, 43, 2168–2197.
- Gao C, Ma Z, and Zhou HH (2017), “Sparse CCA: Adaptive estimation and computational barriers,” *The Annals of Statistics*, 45, 2074–2101.
- Hallin M and Liška R (2011), “Dynamic factors in the presence of blocks,” *Journal of Econometrics*, 163, 29–41.
- Hotelling H (1936), “Relations between two sets of variates,” *Biometrika*, 28, 321–377.
- Huang H (2017), “Asymptotic behavior of support vector machine for spiked population model,” *Journal of Machine Learning Research*, 18, 1–21.
- Kettenring JR (1971), “Canonical analysis of several sets of variables,” *Biometrika*, 58, 433–451.
- Koboldt D, Fulton R, McLellan M, Schmidt H, Kalicki-Veizer J, McMichael J, Fulton L, Dooling D, Ding L, et al. (2012), “Comprehensive molecular portraits of human breast tumours,” *Nature*, 490, 61–70. [PubMed: 23000897]
- Kuligowski J, Pérez-Guaita D, Sánchez-Illana À, León-González Z, de la Guardia M, Vento M, Lock EF, and Quintàs G (2015), “Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE),” *Analyst*, 140, 4521–4529. [PubMed: 25988771]
- Lock EF, Hoadley KA, Marron JS, and Nobel AB (2013), “Joint and individual variation explained (JIVE) for integrated analysis of multiple data types,” *Annals of Applied Statistics*, 7, 523–542.
- Löfstedt T and Trygg J (2011), “OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation,” *Journal of Chemometrics*, 25, 441–455.
- Nadakuditi RR and Silverstein JW (2010), “Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples,” *IEEE Journal of Selected Topics in Signal Processing*, 4, 468–480.
- O’Connell MJ and Lock EF (2016), “R.JIVE for exploration of multi-source molecular data,” *Bioinformatics*, 32, 2877–2879. [PubMed: 27273669]
- Onatski A (2010), “Determining the number of factors from empirical distribution of eigenvalues,” *The Review of Economics and Statistics*, 92, 1004–1016.
- Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, and Bernard PS (2009), “Supervised risk predictor of breast cancer based on intrinsic subtypes,” *Journal of Clinical Oncology*, 27, 1160–1167. [PubMed: 19204204]
- Ross SA (1976), “The arbitrage theory of capital asset pricing,” *Journal of Economic Theory*, 13, 341–360.
- Schouteden M, Van Deun K, Wilderjans TF, and Van Mechelen I (2014), “Performing DISCO-SCA to search for distinctive and common information in linked data,” *Behavior Research Methods*, 46, 576–587. [PubMed: 24178130]
- Smilde AK, Mage I, Naes T, Hankemeier T, Lips MA, Kiers HAL, Acar E, and Bro R (2017), “Common and distinct components in data fusion,” *Journal of Chemometrics*, 31, e2900.

- Song Y, Schreier PJ, Ramirez D, and Hasija T (2016), “Canonical correlation analysis of high-dimensional data with very small sample support,” *Signal Processing*, 128, 449–458.
- Stock JH and Watson MW (2002), “Forecasting using principal components from a large number of predictors,” *Journal of the American statistical association*, 97, 1167–1179.
- Székely GJ, Rizzo ML, and Bakirov NK (2007), “Measuring and testing dependence by correlation of distances,” *The Annals of Statistics*, 35, 2769–2794.
- Trygg J (2002), “O2-PLS for qualitative and quantitative analysis in multivariate calibration,” *Journal of Chemometrics*, 16, 283–293.
- van der Kloet F, Sebastian-Leon P, Conesa A, Smilde A, and Westerhuis J (2016), “Separating common from distinctive variation,” *BMC Bioinformatics*, 17, S195.
- Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, and Ugurbil K (2013), “The WU-Minn human connectome project: an overview,” *NeuroImage*, 80, 62–79. [PubMed: 23684880]
- Vershynin R (2012), “Introduction to the non-asymptotic analysis of random matrices,” in *Compressed Sensing*, Cambridge University Press, Cambridge, pp. 210–268.
- Wang W and Fan J (2017), “Asymptotics of empirical eigenstructure for high dimensional spiked covariance,” *The Annals of Statistics*, 45, 1342–1374. [PubMed: 28835726]
- Yu Q, Risk BB, Zhang K, and Marron J (2017), “Jive integration of imaging and behavioral data,” *NeuroImage*, 152, 38–49. [PubMed: 28246033]
- Zhou G, Cichocki A, Zhang Y, and Mandic DP (2016), “Group component analysis for multiblock data: common and individual feature extraction,” *IEEE Transactions on Neural Networks and Learning Systems*, 27, 2426–2439. [PubMed: 26529787]

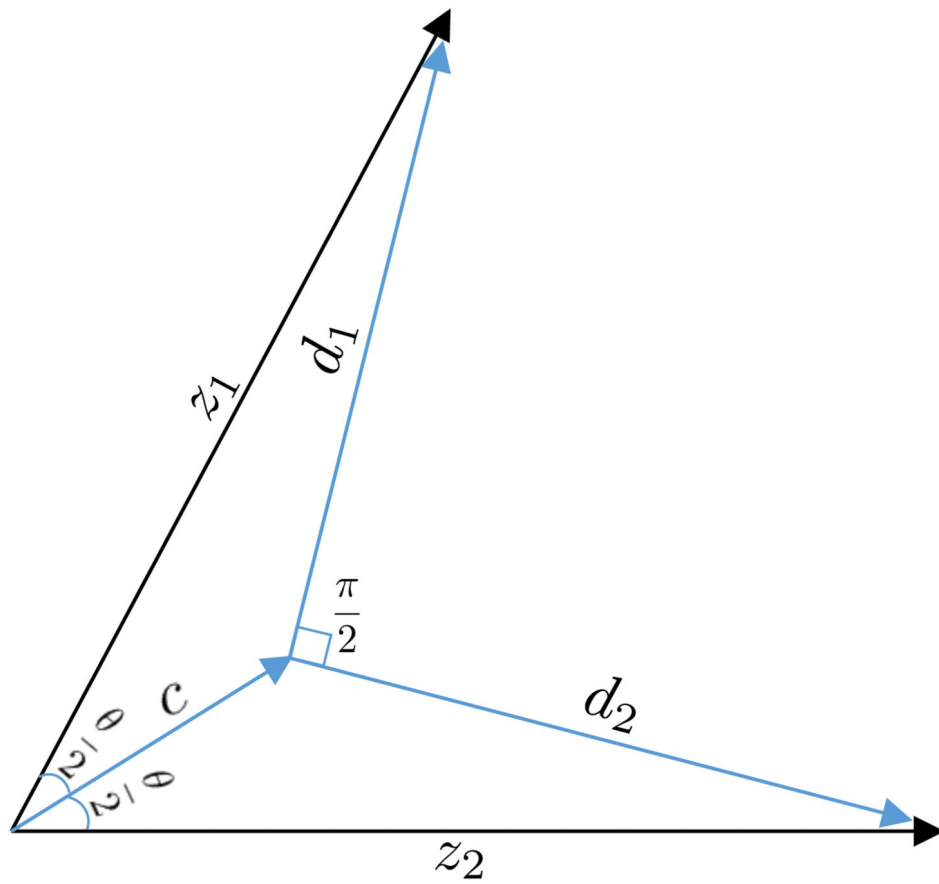


Figure 1:
The geometry of D-CCA for two standardized random variables.

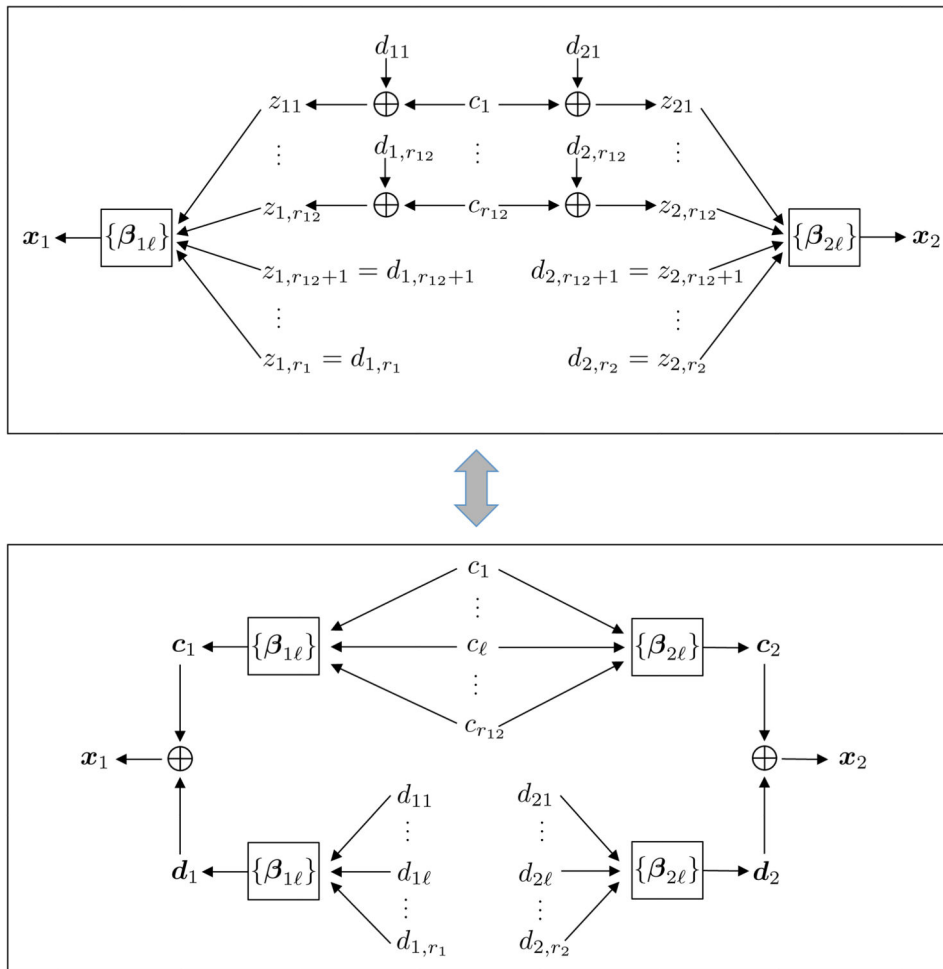


Figure 2:
The decomposition structure of D-CCA.

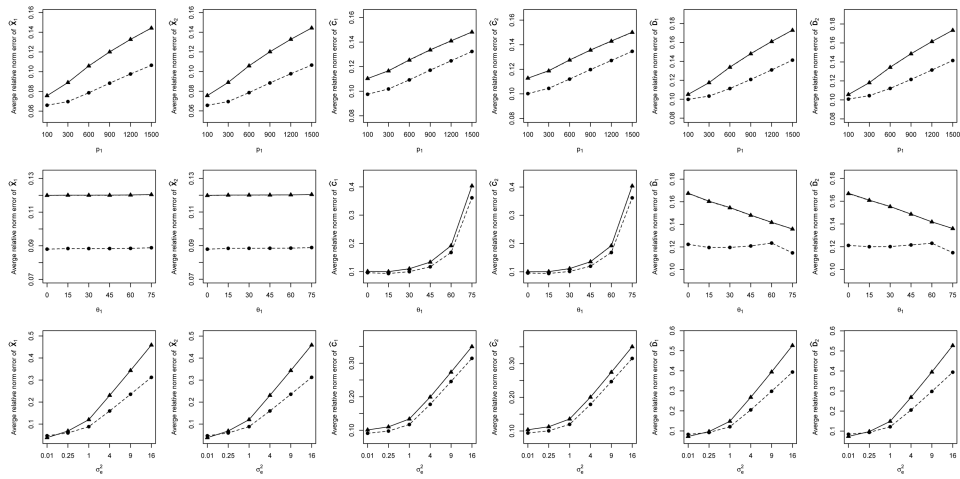


Figure 3: Average relative errors of D-CCA estimates under Setup 1 in spectral norm (○) and Frobenius norm (△) using true r_1 , r_2 and r_{12} , and those in spectral norm (●) and Frobenius norm (▲) using \hat{r}_1 , \hat{r}_2 and \hat{r}_{12} .

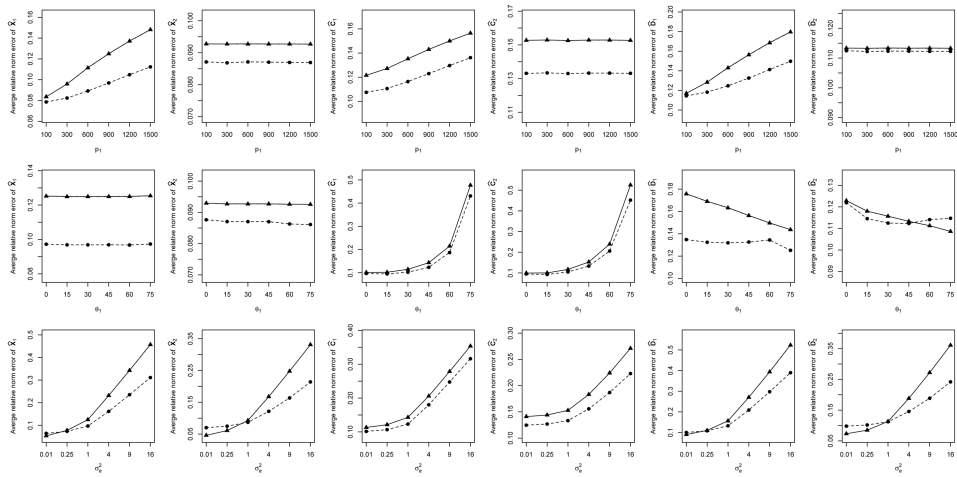


Figure 4: Average relative errors of D-CCA estimates under Setup 2 in spectral norm (○) and Frobenius norm (▲) using true r_1 , r_2 and r_{12} , and those in spectral norm (●) and Frobenius norm (▲) using \hat{r}_1 , \hat{r}_2 and \hat{r}_{12} .

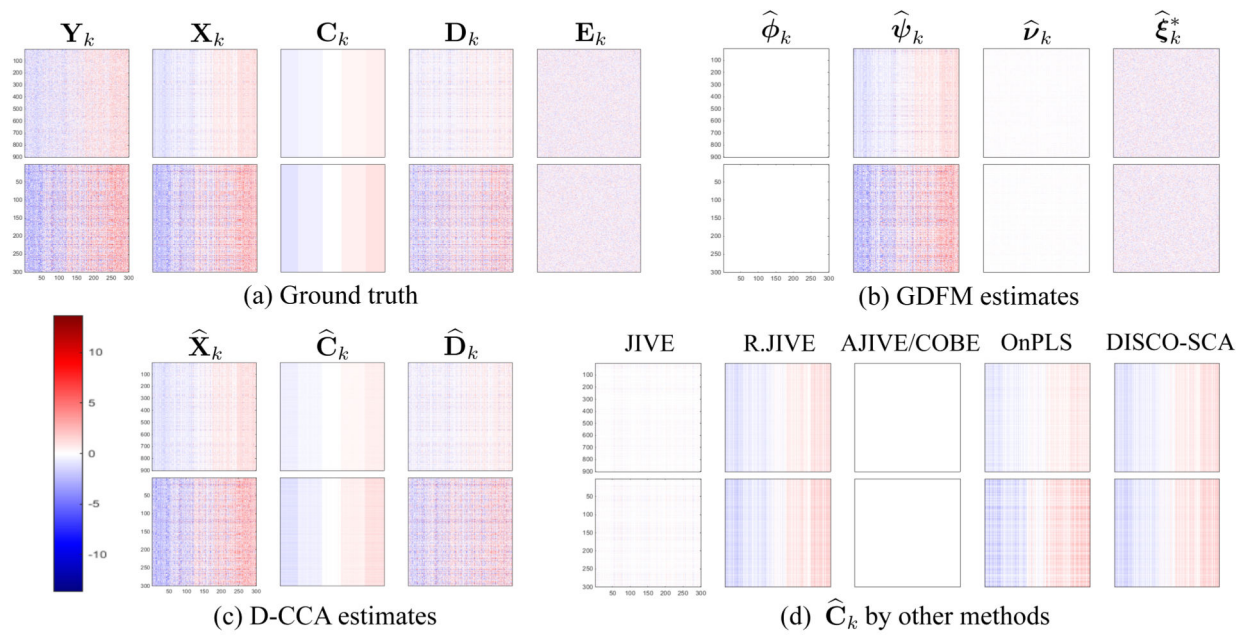


Figure 5:
Color maps for a single replication of Setup 3.

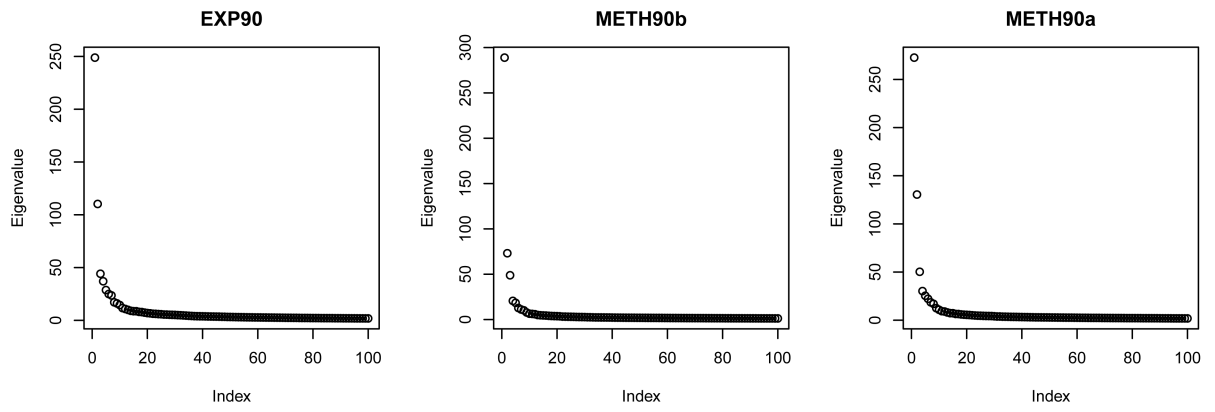


Figure 6:
The scree plot of the sample covariance matrix $\frac{1}{n}Y_k Y_k^T$ for each TCGA dataset.

Table 1:

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

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

Table 2:

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

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

Ratio	Method	Spectral norm	Frobenius norm	Spectral norm	Frobenius norm
$\frac{\ \hat{v}_k\ (\cdot)}{\ \chi_k^*\ (\cdot)}$	GDFM	0.080(0.003)/ 0.080(0.004)	0.099(0.003)/ 0.099(0.003)	0.082(0.003)/ 0.047(0.002)	0.128(0.004)/ 0.044(0.001)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

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

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

Table 4:

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

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

Table 5:

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

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

Table 6:

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

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

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript