Title
Tensor canonical correlation analysis

Journal
Stat, 8(1)

ISSN
0038-9986

Authors
Min, Eun Jeong
Chi, Eric C
Zhou, Hua

Publication Date
2019

DOI
10.1002/sta4.253

Peer reviewed

HHS Public Access

Author manuscript

Stat. Author manuscript; available in PMC 2020 July 10.

Published in final edited form as:

Stat. 2020 ; 8(1): . doi:10.1002/sta4.253.

Tensor canonical correlation analysis

Eun Jeong Min[1], **Eric C. Chi**[2], **Hua Zhou**[3]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, 19104, PA, U.S.A.

[2]Department of Statistics, North Carolina State University, Raleigh, 27695, NC, U.S.A.

[3]Department of Biostatistics, University of California, Los Angeles, Los Angeles, 90095, CA, U.S.A.

Abstract

Canonical correlation analysis (CCA) is a multivariate analysis technique for estimating a linear relationship between two sets of measurements. Modern acquisition technologies, for example, those arising in neuroimaging and remote sensing, produce data in the form of multidimensional arrays or tensors. Classic CCA is not appropriate for dealing with tensor data due to the multidimensional structure and ultrahigh dimensionality of such modern data. In this paper, we present tensor CCA (TCCA) to discover relationships between two tensors while simultaneously preserving multidimensional structure of the tensors and utilizing substantially fewer parameters. Furthermore, we show how to employ a parsimonious covariance structure to gain additional stability and efficiency. We delineate population and sample problems for each model and propose efficient estimation algorithms with global convergence guarantees. Also we describe a probabilistic model for TCCA that enables the generation of synthetic data with desired canonical variates and correlations. Simulation studies illustrate the performance of our methods.

Keywords

block coordinate ascent; CP decomposition; multidimensional array data

1 | INTRODUCTION

Canonical correlation analysis (CCA) is a classic statistical method for identifying associations between two sets of measurements (Hotelling, 1936). Specifically, CCA identifies a pair of coefficient vectors, one for each set of measurements, such that the correlation between the corresponding linear combinations of variables from each set is maximized. By default, CCA applies when each observation consists of a pair of vector covariates. In many modern data analysis problems, however, each observation may consist more generally of a pair of multidimensional arrays or tensors. For example, in imaging

Correspondence: Eric C. Chi, Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A. eric_chi@ncsu.edu.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

genetics, to identify genetic variants that can best capture and explain phenotypic variations in brain function and structure, Stein et al. (2010) studied $p = 448{,}293$ single-nucleotide polymorphisms and $q = 31{,}622$ voxels in brain images, on $n = 740$ individuals. A naive approach to dealing with tensor-valued data would be to reshape tensor covariates into vectors and then apply standard CCA. There are, however, two serious drawbacks to doing so. First, structural information in tensors is discarded through vectorization. Second, the resulting vectors consist of a prohibitively large number of parameters. In the imaging genetics problem (Stein et al., 2010), vectorizing the voxel intensity measurements disregards the spatial correlation among neighbouring voxels. Moreover, applying standard CCA to vectors of single-nucleotide polymorphism covariates and vectorized brain images would require estimating nearly half a million parameters using fewer than a thousand observations.

In light of these issues, there have been extensions of CCA to handle special cases of tensor-valued data (Lee & Choi, 2007; Wang, 2010; Yan, Zheng, Zhou, & Zhao, 2012; Gang, Yong, Yan-Lei, & Jing, 2011; Lu, 2013; Wang, Yan, Sun, Zhao, & Fu, 2016). Although they have exhibited good empirical performance in some applications, there remains no clear population models underlying these sample based heuristics. To address this gap in the literature, we introduce a novel statistical model for tensor CCA (TCCA). We summarize at a high level our formulation and its contributions:

- We propose a TCCA population model that imposes the CANDECOMP/PARAFAC (CP) decomposition (Carroll & Chang, 1970; Harshman, 1970) structure on canonical tensors. This population model enforces model parsimony and enables efficient estimation.
- We propose a refinement of TCCA, which assumes a separable covariance structure (scTCCA). This refinement enables efficient estimation of large covariance matrices of tensor-valued data.
- We derive convenient representations of the covariance between linear combinations of two random tensors under the unstructured and separable covariance structured assumptions.
- We develop efficient estimation algorithms for TCCA and scTCCA, both based on block coordinate ascent, which leverage these efficient representations. Each step of both algorithms solves a substantially lower dimensional CCA problem; thus, both algorithms can be easily implemented using any standard solvers for the CCA problem. Moreover, we prove global convergence guarantees of both estimation algorithms under modest regularity conditions.
- We develop simple modifications to the TCCA and scTCCA estimation algorithms to incorporate recovery of sparse canonical correlation tensors to improve interpretability of the estimated models.
- Finally, we extend the probabilistic interpretation of CCA by Bach and Jordan (2006) to TCCA. This extension leads to a probabilistic model for generating datasets with specified canonical correlation and variates.

The remainder of the paper is organized as follows. In Section 2, we review tensor notation and basic operations used in this paper. In Sections 3 and 4, we propose our two TCCA methods: TCCA and scTCCA. In Section 5, we describe a modification of the estimation algorithms for TCCA and scTCCA for sparse models. In Section 6, we introduce the probabilistic TCCA model. In Section 7, we describe the numerical experiment results. In Section 8, we conclude and highlight directions for future work.

2 | NOTATION AND PRELIMINARIES

We review basic operations on matrices and tensors invoked throughout this paper, adopting the terminology and notation in Kolda and Bader (2009). Throughout the paper, we use lowercase letters to indicate scalars, bold lowercase letters to indicate vectors, bold capital characters to indicate matrices, and bold calligraphic capital characters to indicate tensors. We will also use the shorthand [n] to denote an index set $\{1, \ldots, n\}$.

Tensors can be considered generalizations of scalars, vectors, and matrices. Let $\mathcal{X}$ represent an D-dimensional tensor in $\mathbb{R}^{p1 \times \ldots \times pD}$. The tensor $\mathcal{X}$ has order D, its number of dimensions or modes. For example, vectors are tensors of order one and have one mode. Matrices are tensors of order two and have two modes. We denote an element of $\mathcal{X}$ by $x_{\iota 1 \iota 2, \ldots, \iota D}$, where $\iota_i \in [p_i]$ and $i \in [D]$. Fibres are the generalization of matrix rows and columns to higher order tensors. A fibre is defined by fixing the index of every dimension except one dimension. Mode i fibres are p_i-dimensional vectors extracted from $\mathcal{X}$ by fixing all the indices $(\iota_1, \ldots, \iota_{l-1}, \iota_{i+1}, \ldots, \iota_D)$ except the ith one ι_i. For example, columns of a matrix are Mode 1 fibres, and rows of a matrix are Mode 2 fibres.

It is often useful to reshape a tensor into a matrix. Reordering a tensor into a matrix is referred to as matricization. The mode i matricization of a tensor $\mathcal{X} \in \mathbb{R}^{p1 \times \ldots \times pD}$, denoted $\mathbf{X}_{(i)} \in \mathbb{R}^{p_i \times p_{-i}}$ with $p_{-i} = \Pi_{k=1, k \neq i}^{D} p_k$, arranges the mode i fibres as the columns of the matrix $\mathbf{X}_{(i)}$. In a mode i matricization, the tensor element $x_{\iota_1, \ldots, \iota_D}$ is mapped to the matrix element of $\mathbf{X}_{(i)}$. with index $(\iota_i j)$, where $j = 1 + \sum_{k=1, k \neq i}^{D} (\iota_k - 1) J_k$ with

$$J_k = \begin{cases} 1 & \text{if } k = 1 \text{ or if } k = 2 \text{ and } i = 1. \\ \prod_{k=1, k' \neq i}^{k-1} p_{k'} & \text{otherwise.} \end{cases}$$

Reordering a tensor into a vector is referred to as vectorization. We first describe vectorization of a matrix before describing vectorization of a general tensor. The vectorization of a matrix $\mathbf{X}$ is denoted by vec($\mathbf{X}$) and is the vector obtained by stacking the columns of $\mathbf{X}$ on top of each other. The vectorization of the mode i matricization of a tensor $\mathcal{X}$ in turn is denoted as $\mathbf{x}_{(i)} = \text{vec}(\mathbf{X}_{(i)})$. We then define the vectorization of a tensor $\mathcal{X}$, denoted by vec($\mathcal{X}$), as the vectorization of its Mode 1 matricization, namely, vec($\mathbf{X}_{(1)}$). When unambiguous from context, we will often denote the vectorization of a tensor $\mathcal{X}$ by its corresponding bold lowercase $\mathcal{X}$.

The inner product of two tensors of compatible dimensions $\mathcal{X}, \widetilde{\mathcal{X}} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ is the sum of the product of their entries, namely,

$$\langle \mathcal{X}, \widetilde{\mathcal{X}} \rangle = \sum_{\iota_1 = 1}^{p_1} \cdots \sum_{\iota_D = 1}^{p_D} \mathbf{x}_{\iota_1, \ldots, \iota_D} \tilde{\mathbf{x}}_{\iota_1, \ldots, \iota_D}.$$

The mode i product of a tensor $\widetilde{\mathcal{X}} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ with a matrix $\cup \in \mathbb{R}^{J \times p_d}$ is denoted by $\mathcal{X} \times_i \cup$ and is the tensor of size $p_1 \times \ldots \times p_{i-1} \times J \times p_{i+1} \times \ldots \times p_D$ with elements

$$(\mathcal{X} \times_i \cup)_{\iota_1, \ldots, \iota_{i-1}, j_{\iota_{i+1}}, \ldots, \iota_D} = \sum_{\iota_i = 1}^{p_i} x_{\iota_1 \iota_2, \ldots, \iota_D} U_{j_{\iota_i}}.$$

Finally, we review three kinds of matrix products, as well as one definition of matrix division, that will be used throughout the paper.

- For two matrices $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$ and $\mathbf{B} \in \mathbb{R}^{q_1 \times q_2}$, the Kronecker product is the $p_1 q_1$-by-$p_2 q_2$ matrix,

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1p_2}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p_1}\mathbf{B} & \cdots & a_{p_1 p_2}\mathbf{B} \end{pmatrix}.$$

- For two matrices $\mathbf{A} = \begin{pmatrix} a_1 & \cdots & a_{p_2} \end{pmatrix} \in \mathbb{R}^{p_1 \times p_2}$ and $B = \begin{pmatrix} b_1 & \cdots & b_{p_2} \end{pmatrix} \in \mathbb{R}^{q_1 \times p_2}$ that have the same number of columns p_2, the Khatri-Rao product is the $p_1 q_1$-by-p_2 matrix,

$$A \odot B = \left(a_1 \otimes b_1 a_2 \otimes b_2 \ldots a_{p_2} \otimes b_{p_2}\right).$$

  which is a column-wise Kronecker product of **A** and **B**.

- For two matrices **A** and **B** of the same size, the Hadamard product is the element-wise product $\mathbf{A} * \mathbf{B} = \{a_{ij} b_{ij}\}$. Because the Hadamard product commutes, we use $*_i \mathbf{A}_i$ to denote $\mathbf{A}_1 * \ldots * \mathbf{A}_m = \mathbf{A}_{\boldsymbol{\pi}(1)} * \ldots * \mathbf{A}_{\boldsymbol{\pi}(m)}$ for any permutation $\boldsymbol{\pi}$.
- Finally, for two matrices **A** and **B** of the same size, the Hadamard quotient is the element-wise quotient $\mathbf{A} \oslash \mathbf{B} = \{a_{ij}/b_{ij}\}$.

3 | TENSORCANONICAL CORRELATION ANALYSIS

3.1 | Population TCCA

Let $\mathcal{X} \in \mathbb{R}^{p_1 \times \ldots \times p_{D_x}}$ and $\mathcal{Y} \in \mathbb{R}^{q_1 \times \ldots \times q_{D_y}}$ be two random tensors of order D_x and D_y, respectively. We denote the vectorizations of $\mathcal{X}$ and $\mathcal{Y}$ by **x** and **y**, respectively. Denote by Σ_x and Σ_y the covariances of **x** and **y**, respectively. Denote by $\Sigma_{x,y}$ the covariance between **x**

and **y**. Let $\mathscr{V} \in \mathbb{R}^{p_1 \times \ldots \times p_{D_x}}$, and $\mathscr{W} \in \mathbb{R}^{q_1 \times \ldots \times q_{D_y}}$ be constant tensors, and let $\rho(\mathscr{V}, \mathscr{W})$ denote the correlation between the two linear combinations $\langle \mathscr{X}, \mathscr{V} \rangle$ and $\langle \mathscr{Y}, \mathscr{W} \rangle$ namely,

$$\rho(\mathscr{V}, \mathscr{W}) = \frac{\mathrm{Cov}(\langle \mathscr{X}, \mathscr{V} \rangle, \langle \mathscr{Y}, \mathscr{W} \rangle)}{\sqrt{\mathrm{Var}(\langle \mathscr{X}, \mathscr{V} \rangle)}\sqrt{\mathrm{Var}(\langle \mathscr{Y}, \mathscr{W} \rangle)}} = \frac{\mathbf{v}^\top \Sigma_{\mathbf{x},\mathbf{y}} \mathbf{w}}{\sqrt{\mathbf{v}^\top \Sigma_{\mathbf{x}} \mathbf{v}}\sqrt{\mathbf{w}^\top \Sigma_{\mathbf{y}} \mathbf{w}}}. \quad (1)$$

The pair $(\mathscr{V}, \mathscr{W})$ that maximizes ρ are the canonical tensors, and the optimal is the canonical coefficient. Maximizing the objective in Equation (1) presents two challenges: (a) high dimensionality of optimization variables $\mathscr{V}$ and $\mathscr{W}$ and (b) the estimation of the huge covariance matrices Σ_x and Σ_y and the cross-covariance matrix $\Sigma_{x,y}$. We will address challenge (b) by imposing a separable covariance structure in Section 4.

To address challenge (a), we impose the parsimonious CANDECOMP/PARAFAC (CP), or Kruskal, representation on the canonical tensors. The CP representation generalizes the idea of representing a matrix as the sum of Rank 1 matrices to representing a tensor as the sum of Rank 1 tensors. An order-D tensor $\mathscr{X} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ is Rank 1 if it can be expressed as the outer product of D vectors $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$, …, $\mathbf{a}^{(D)}$, namely, $\mathscr{X} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \ldots \circ \mathbf{a}^{(D)}$, where the binary operator o denotes the vector outer product. Thus, the $(\imath_1, \imath_2, \ldots, \imath_D)$th element of $\mathscr{X}$ is $\mathbf{x}_{\imath_1 \imath_2 \ldots \imath_D} = a^{(1)}_{\imath_1} a^{(2)}_{\imath_2} \ldots a^{(D)}_{\imath_0}$. A rank-$R$ tensor can be written as the sum of R Rank 1 tensors, namely,

$$\mathscr{X} = [A_1, \ldots, A_D] = \sum_{r=1}^{R} a_r^{(1)} \circ \ldots \circ a_r^{(D)},$$

where $A_i = \left(a_1^{(1)} \; a_2^{(i)} \; \ldots \; a_R^{(i)}\right) \in \mathbb{R}^{p_i \times R}$ denotes the mode i factor matrix. We use the Kruskal notation […] to concisely summarize the sum.

Thus, instead of searching over the space of all order-D_x and order-D_y tensor pairs $(\mathscr{V}, \mathscr{W})$, we limit our search to tensors of rank-R_x and rank-R_y,

$$\mathscr{V} = \left[V_1, \ldots, V_{D_x}\right], V_i \in \mathbb{R}^{P \times R_x}, i \in [D_x], \mathscr{W} = \left[W_1, \ldots, W_{D_y}\right], W_j \in \mathbb{R}^{q_j \times R_y}, j \in \left[D_y\right]. \quad (2)$$

As we will see later, this parameterization makes progress towards alleviating the burden of estimating a huge covariance matrix. Note that

$$\langle \mathscr{X}, \mathscr{V} \rangle = \langle \mathscr{X}, \sum_{r=1}^{R_x} \mathbf{w}_r^{(1)} \circ \ldots \circ \mathbf{w}_r^{(D_x)} \rangle = \sum_{r=1}^{R_x} x \times_1 \mathbf{w}_r^{(1)} \times_2 \ldots \times_{D_x} \mathbf{w}_r^{(D_x)}$$

$$\langle \mathcal{Y}, \mathcal{W} \rangle = \langle \mathcal{Y}, \sum_{r=1}^{R_Y} \mathbf{w}_r^{(1)} \circ \ldots \circ \mathbf{w}_r^{(D_y)} \rangle = \sum_{r=1}^{R_y} \mathcal{Y} \times_1 \mathbf{w}_r^{(1)} \times_2 \ldots \times_{D_y} \mathbf{w}_r^{(D_y)}.$$

Thus, we seek to maximize the correlation between a rank-R_x multilinear form in $\mathcal{X}$ and a rank-R_y multilinear form in $\mathcal{Y}$. Multiway information is preserved, and the dimensionality is reduced from an exponential number of parameters $\Pi_{i=1}^{D_x} p_i + \Pi_{j=1}^{D_y} q_j$ to a linear number of parameters $R_x \sum_{i=1}^{D_x} p_i + R_y \sum_{j=1}^{D_y} q_j$. Note that the ranks ($R_x$, R_y) here are not the number of canonical tensor pairs being sought. In this paper, we focus on obtaining only the top canonical tensor pair ($\mathcal{V}, \mathcal{W}$), which have ranks R_x and R_y.

The following representations of Var ($\langle \mathcal{X}, \mathcal{Y} \rangle$), Var ($\langle \mathcal{Y}, \mathcal{W} \rangle$), and Cov($\langle \mathcal{X}, \mathcal{Y} \rangle, \langle \mathcal{Y}, \mathcal{N} \rangle$) in terms of a CP decomposition are keys to our estimation algorithms. Its proof is in the Supporting Information.

Proposition 1—*Let* $\mathcal{X} \in \mathbb{R}^{p_1 \times \ldots \times p_{D_x}}$ *and* $\mathcal{Y} \in \mathbb{R}^{q_1 \times \ldots \times q_{D_y}}$ *be two random tensors and* $\mathcal{V} = [V_1, \ldots, V_{D_x}]$ *and* $\mathcal{W} = [W_1, \ldots, W_{D_y}]$ *be two constant tensors of the same size as* $\mathcal{X}$ and $\mathcal{Y}$ *respectively. Define*

$$\mathbf{V}_{(-1)} = \left[\mathbf{V}_{D_x} \odot \ldots \odot \mathbf{V}_{i+1} \odot \mathbf{V}_{i-1} \odot \ldots \odot \mathbf{V}_1\right] \otimes \mathbf{I}_{p_i}, \tag{3}$$

$$\mathbf{W}_{(-j)} = \left[\mathbf{W}_{D_x} \odot \ldots \odot \mathbf{W}_{j+1} \odot \mathbf{W}_{j-1} \odot \ldots \odot \mathbf{W}_1\right] \otimes \mathbf{I}_{q_j}, \tag{4}$$

and let $\Sigma_{x_{(j)}}$ *denote the covariance of* $\mathbf{x}_{(i)} = vec(\mathbf{X}_{(i)})$. *Define* $\Sigma_{\mathbf{y}_{(j)}}$ *and* $\Sigma_{\mathbf{x}_0, \mathbf{y}_{(j)}}$ *analogously. Then*

$$\mathrm{Var}(\langle \mathcal{X}, \mathcal{V} \rangle) = \mathbf{v}_i^\top \mathbf{V}_{(-i)}^\top \Sigma_{\mathbf{x}_{(i)}} \mathbf{V}_{(-i)} \mathbf{v}_i$$

$$\mathrm{Var}\,(\langle \mathcal{Y}, \mathcal{W} \rangle) = \mathbf{w}_j^\top \mathbf{W}_{(-j)}^\top \Sigma_{y_{(j)}} \mathbf{W}_{(-j)} \mathbf{w}_j$$

$$Cov(\langle \mathcal{X}, \mathcal{V} \rangle, \langle \mathcal{Y}, \mathcal{W} \rangle) = \mathbf{v}_i^\top \mathbf{V}_{(-i)}^\top \Sigma_{\mathbf{x}_{(j)}, \mathbf{y}_{(j)}} \mathbf{W}_{(-j)} \mathbf{w}_j$$

for any $i \in [D_x]$ and $j \in [D_y]$, where $\mathbf{v}_i = vec(\mathbf{V}_i)$ *and* $\mathbf{W}j = vec(\mathbf{W}_j)$.

3.2 | Sample TCCA

Suppose we observe N pairs of i.i.d. tensor data ($\mathcal{X}_n, \mathcal{Y}_n$), and we estimate $\mathcal{V}$ and $\mathcal{W}$ by solving the optimization problem

$$\text{maximize } \hat{\rho}(\mathscr{V}, \mathscr{W}) \equiv \frac{\mathbf{v}^\top \widehat{\Sigma}_{\mathbf{x},\mathbf{y}} \mathbf{W}}{\sqrt{\mathbf{v}^\top \widehat{\Sigma}_{\mathbf{x}} \mathbf{v}} \sqrt{\mathbf{w}^\top \widehat{\Sigma}_{\mathbf{y}} \mathbf{w}}}. \tag{5}$$

where $\widehat{\Sigma}_{\mathbf{x}}$, $\widehat{\Sigma}_{\mathbf{y}}$ and $\widehat{\Sigma}_{\mathbf{x},\mathbf{y}}$ are sample estimates of the corresponding covariances. Recall that CCA models can be estimated numerically by computing the solution to the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \widehat{\Sigma}_{x,y} \\ \widehat{\Sigma}_{y,x} & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = \rho \begin{pmatrix} \widehat{\Sigma}_x & 0 \\ 0 & \widehat{\Sigma}_y \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}.$$

This problem is guaranteed to have a solution if and only/ if the covariance matrices $\widehat{\Sigma}_{\mathbf{x}}$ and $\widehat{\Sigma}_{\mathbf{y}}$ are nonsingular. In practice, the sample size N is smaller than the size of $\widehat{\Sigma}_{\mathbf{x}}$ and $\widehat{\Sigma}_{\mathbf{y}}$, $\left(\Pi_{i=1}^{D_x} p_i\right)\left(\times \Pi_{i=1}^{D_x} p_i\right), \left(\Pi_{j=1}^{D_x} q_j\right) \times \left(\Pi_{j=1}^{D_y} q_j\right)$, respectively. Therefore, the sample covariance matrices $\widehat{\Sigma}_{\mathbf{x}}$ and $\widehat{\Sigma}_{\mathbf{y}}$ are singular, and a solution cannot be obtained. As a remedy, several regularized estimation methods for obtaining nonsingular sample covariance matrices (Vinod, 1976; Ledoit & Wolf, 2004; González, Déjean, Martin, & Baccini, 2008; Ledoit & Wolf, 2012; Cai & Yuan, 2012; Bickel & Levina, 2008; 2008; Kubokawa et al., 2013; Srivastava & Reid, 2012) can be used.

If we take Var ($\langle \mathscr{X}, \mathscr{V} \rangle$), Var ($\langle \mathscr{Y}, \mathscr{W} \rangle$) and Cov($\langle \mathscr{X}, \mathscr{V} \rangle, \langle \mathscr{Y}, \mathscr{W} \rangle$) to be $\mathbf{v}^T \widehat{\Sigma}_{\mathrm{x}} \mathbf{v}$, $\mathbf{w}^T \widehat{\Sigma}_{\mathrm{y}} \mathbf{w}$, and $\mathbf{v}^\top \widehat{\Sigma}_{\mathrm{x,y}} \mathbf{w}$ respectively, then Proposition 1 suggests a block coordinate ascent algorithm where we update the factor matrices *in pairs* ($\mathbf{V}_i$, $\mathbf{W}_j$), for different combinations of $i \in [D_{\mathrm{x}}]$ and $j \in [D_{\mathrm{y}}]$. To update the pair ($\mathbf{V}_i$, $\mathbf{W}_j$), we solve the following problem:

$$\begin{aligned} &\text{maximize } \mathbf{v}_i^\top \mathbf{V}_{(-i)}^\top \widehat{\Sigma}_{\mathbf{x}_{(i)},\mathbf{y}_{(j)}} \mathbf{W}_{(-j)} \mathbf{w}_j \text{ subject to } \mathbf{v}_i^\top \mathbf{V}_{(-i)}^\top \widehat{\Sigma}_{\mathbf{x}_{(i)}} \mathbf{V}_{(-i)} \mathbf{v}_i = 1, \\ &\mathbf{w}_j^\top \mathbf{W}_{(-j)}^\top \widehat{\Sigma}_{\mathbf{y}_{(j)}} \mathbf{W}_{(-j)} \mathbf{w}_j = 1, \end{aligned} \tag{6}$$

which is a substantially smaller optimization problem over $p_i R_x + q_j R_y$ variables compared with any alternative "all-at-once" strategies to iteratively optimize over all $\left(\Pi_{i=1}^{D_{\mathrm{x}}} p_i\right)\left(\times \Pi_{i=1}^{D_{\mathrm{x}}} p_i\right) + \left(\Pi_{j=1}^{D_{\mathrm{y}}} q_j\right) \times \left(\Pi_{j=1}^{D_{\mathrm{y}}} q_i\right)$ parameters simultaneously. Problem (6) includes $\widehat{\Sigma}_{\mathbf{x}_{(i)}}$, a permuted version of $\widehat{\Sigma}_{\mathbf{x}}$ with size $\left(\Pi_{i=1}^{D_{\mathrm{x}}} p_i\right)\left(\times \Pi_{i=1}^{D_{\mathrm{x}}} p_i\right)$. However, we work on the "compressed" covariance matrix

$$\mathbf{v}_{(-i)}^\top \widehat{\Sigma}_{\mathbf{x}_{(i)}} \mathbf{v}_{(-i)} \in \mathbb{R}^{p_i R_x \times p_i R_x},$$

which is likely to be full rank when $N \geq p_i R_x$ instead of a singular matrix $\widehat{\Sigma}_{\mathbf{x}_{(i)}}$. This approach enables us to solve the generalized eigenvalue problem with matrices that are $(p_i R_x) + (q_j R_y)$-by-$(p_i R_x) + (q_{\mathrm{j}} R_{\mathrm{y}})$. Algorithm 1 summarizes the estimation procedure that

comes with the following convergence guarantees. The proof is in the Supporting Information.

Proposition 2—If the matrices $\widehat{\Sigma}_{\mathbf{x}}$ and $\widehat{\Sigma}_{y}$ are nonsingular, then the limit Points of the iterate sequence generated by Algorithm 1 are canonical tensors of the sample TCCA Problem.

Algorithm 1

TCCA, with (i) assumptions on CP structure on canonical tensors ($\mathscr{V}, \mathscr{W}$) and (ii) no additional assumptions on the structures of the covariances Var(vec($\mathscr{X}$)) and Var(vec($\mathscr{Y}$))

Initialize $v_i^{(0)}$ and $w_j^{(0)}$, for $i \in [D_x]$ and $j \in [D_y]$

$t \leftarrow 0$

repeat

Select $(i, j) \in [D_x] \times [D_y]$

$$\mathbf{V}_{(-i)}^{(t)} \leftarrow \left[\mathbf{v}_{D_x}^{(t)} \odot \cdots \odot \mathbf{V}_{i+1}^{(t)} \odot \mathbf{V}_{i-1}^{(t)} \odot \cdots \odot \mathbf{V}_{1}^{(t)}\right] \otimes \mathbf{I}_{p_i}$$

$$\mathbf{W}_{(-j)}^{(t)} \leftarrow \left[\mathbf{W}_{D_y}^{(t)} \odot \cdots \odot \mathbf{W}_{j+1}^{(t)} \odot \mathbf{W}_{j-1}^{(t)} \odot \cdots \odot \mathbf{W}_{1}^{(t)}\right] \otimes \mathbf{I}_{q_j}$$

$$\mathbf{C}_{\mathbf{x}}^{(t)} \leftarrow \mathbf{V}_{(-i)}^{(t)T} \widehat{\mathbf{\Sigma}}_{\mathbf{x}(i)} \mathbf{V}_{(-i)}^{(t)}$$

$$\mathbf{C}_{\mathbf{y}}^{(t)} \leftarrow \mathbf{W}_{(-j)}^{(t)\top} \widehat{\mathbf{\Sigma}}_{y(j)} \mathbf{W}_{(-j)}^{(t)}$$

$$\mathbf{C}_{\mathbf{x},\mathbf{y}}^{(t)} \leftarrow \mathbf{V}_{(-i)}^{(t)T} \widehat{\mathbf{\Sigma}}_{\mathbf{x}(i)\mathbf{y}(j)} \mathbf{W}_{(-j)}^{(t)}$$

Generalized eigen-decomposition: $\begin{pmatrix} 0 & C_{x,y}^{(t)} \\ C_{xy}^{(t)T} & 0 \end{pmatrix} \begin{pmatrix} w_i^{(t+1)} \\ w_j^{(t+1)} \end{pmatrix} = \rho^{(t+1)} \begin{pmatrix} C_x^{(t)} & 0 \\ 0 & C_y^{(t)} \end{pmatrix} \begin{pmatrix} w_i^{(t+1)} \\ w_j^{(t+1)} \end{pmatrix}$

$t \leftarrow t+1$

until $\rho^{(t)}$ converges

4 | TCCA WITH SEPARABLE COVARIANCE STRUCTURE

4.1 | Population TCCA with separable covariances

Hoff (2011) proposed the array normal distribution with separable covariance structure. Separable marginal covariances are defined as

$$\Sigma_{\mathrm{x}} = \mathrm{Var}\,(\mathbf{x}) = \Sigma_{\mathrm{x},\mathrm{D_x}} \otimes \ldots \otimes \Sigma_{\mathrm{x},1} \text{ and } \Sigma_{\mathrm{y}} = \mathrm{Var}\,(\mathrm{y}) = \Sigma_{\mathrm{y},\mathrm{D_y}} \otimes \ldots \otimes \Sigma_{\mathrm{y},1}, \tag{7}$$

where $\Sigma_{\mathrm{x},j} \in \mathbb{R}^{p_i \times p_i}$ for $i \in [D_{\mathrm{x}}]$ and $\Sigma_{\mathrm{y},j} \in \mathbb{R}^{q_j \times q_j}$ for $j \in [D_{\mathrm{y}}]$. Then the overall covariance of population model is

$$\mathrm{Var}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{x},\mathbf{D}_x} \otimes \ldots \otimes \Sigma_{\mathbf{x},1} & \Sigma_{\mathbf{x},\mathbf{y}} \\ \Sigma_{y_{\mathbf{x}}} & \Sigma_{\mathbf{y},\mathbf{D}_y} \otimes \ldots \otimes \Sigma_{\mathbf{y},1} \end{pmatrix}. \tag{8}$$

Intuitively, $\Sigma_{\mathbf{x},i}$ summarizes the covariance along the mode i fibres of tensor $\mathcal{X}$ and $\Sigma_{y,j}$ summarizes the covariance along the mode j fibres of tensor $\mathcal{Y}$. The following result shows a representation of $\mathrm{Var}(\langle \mathcal{X}, \mathcal{V} \rangle)$ and $\mathrm{Var}(\langle \mathcal{Y}, \mathcal{W} \rangle)$ in presence of separable covariance structure (7) that we will leverage in our estimation algorithm.

Proposition 3—Let $\mathcal{X} \in \mathbb{R}^{p_1 \times \ldots \times p_{D_x}}$ and $\mathcal{Y} \in \mathbb{R}^{q_1 \times \ldots \times q_{D_y}}$ be two random tensors admitting the separable covariance structure (7) and $\mathcal{V} = \left[v_1, \ldots, \mathbf{v}_{D_x}\right]$ and $\mathcal{W} = \left[\mathbf{W}_1, \ldots, \mathbf{W}_{D_y}\right]$ be two constant tensors. Define

$$\mathrm{H}_{\mathrm{x},-i} = {}^{*}_{i' \neq i}\left(\mathrm{V}_{i'}^{\top}\Sigma_{\mathrm{x},i'}\mathrm{V}_{i'}\right) \text{ and } \mathrm{H}_{\mathrm{y},-j} = {}^{*}_{j' \neq j}\left(\mathrm{W}_{j'}^{\top}\Sigma_{y,j'}\mathrm{W}_{j'}\right).$$

Then

$$\mathrm{Var}(\langle \mathcal{X}, \mathcal{V} \rangle) = \mathbf{v}_i^{\top}\left(\mathbf{H}_{\mathbf{x},-i} \otimes \Sigma_{\mathbf{x},i}\right)\mathbf{v}_i \text{ and } \mathrm{Var}(\langle \mathcal{Y}, \mathcal{W} \rangle) = \mathbf{w}_j^{\top}\left(\mathbf{H}_{\mathbf{y},-j} \otimes \Sigma_{\mathbf{y}_j}\right)\mathbf{w}_j,$$

for any $i \in [D_x]$ *and* $j \in [D_y]$, *where* $\mathbf{v}_i = vec(\mathbf{V}_i)$ *and* $\mathbf{w}_j = vec(\mathbf{W}_j)$.

With the separable covariance structure and the CP structure on canonical tensors $(\mathcal{V}, \mathcal{W})$, the objective function of TCCA population model (1) greatly simplifies. Note that the separable covariance structure may not hold for real data, in which case the covariance estimates are biased. Despite this drawback, this parsimonious structure is worth considering due to the stability that it can impart by reducing estimation variance.

4.2 | Sample TCCA with separable covariances

Given data $(\mathcal{X}_n, \mathcal{Y}_n)$, $n \in [N]$, the goal is to maximize the sample canonical correlation (5) under assumptions that (a) $\mathcal{V}$ and $\mathcal{W}$ have the CP decomposition structure and (b) $\mathcal{X}$ and $\mathcal{Y}$ admit the separable covariance structure. We follow the same strategy as sample TCCA in Section 3.2, updating parameters in pairs of factor matrices $(\boldsymbol{V}_i, \boldsymbol{W}_j)$. To update the pair $(\boldsymbol{V}_i, \boldsymbol{W}_j)$, we solve the subproblem

$$\text{maximize } \mathbf{v}_i^{\top}\mathbf{V}_{(-i)}^{\top}\widehat{\Sigma}_{\mathbf{x}_{(i)},\mathbf{y}_{(j)}}\mathbf{W}_{(-j)}\mathbf{w}_j, \text{ subject to } \mathbf{v}_i^{\top}\left(\mathbf{H}_{\mathbf{x},-i} \otimes \widehat{\Sigma}_{\mathbf{x},i}\right)\mathbf{v}_i = 1, \mathbf{w}_j^{\top}\left(\mathbf{H}_{\mathbf{y},-j} \otimes \widehat{\Sigma}_{\mathbf{y},j}\right)\mathbf{w}_j = 1.$$

where $\mathbf{H}_{\mathbf{x},-i}$ and $\mathbf{H}_{\mathbf{y},-j}$, defined in Proposition 3, are evaluated at current iterates $\mathbf{V}_{i'}$ where $i' \neq i$, and $\mathbf{W}_{j'}$, where $j' \neq j$.

By assuming the separable covariance structure, Proposition 3 enables us to greatly simplify the variance calculations for $\mathrm{Var}(\langle \mathcal{X}, \mathcal{V} \rangle)$ and $\mathrm{Var}(\langle \mathcal{Y}, \mathcal{W} \rangle)$. Note that the calculation of $\mathrm{V}_{(-i)}^{\top}\widehat{\Sigma}_{x_0}\mathrm{V}_{(-i)}$ and $\mathrm{W}_{(-j)}^{\top}\widehat{\Sigma}_{y_{(j)}}\mathrm{W}_{(-j)}$ in the subproblem of the sample TCCA costs $\left(R_x\prod_{i=1}^{D_x} p_i\right)^2 + \left(R_y\prod_{j=1}^{D_y} q_j\right)^2$ flops. In contrast, the calculation of matrices $\mathrm{H}_{x,-i} \otimes \widehat{\Sigma}_{x,j}$ and $\mathrm{H}_{y,-j} \otimes \widehat{\Sigma}_{y,j}$ in the subproblem of the sample TCCA with the separable covariance structure only costs $(R_x p_j)^2 + (R_y q_j)^2$ flops. Algorithm 2 summarizes the estimation

procedure under the separable covariance assumption (7). Like Algorithm 1, Algorithm 2 also comes with convergence guarantees. The proof is in the Supporting Information.

Proposition 4—If the matrices $\widehat{\Sigma}_{\mathbf{x}}$ and $\widehat{\Sigma}_{y}$ are nonsingular, then the limit points of the iterate sequence generated by Algorithm 2 are canonical tensors of the sample TCCA problem with separable covariances.

Algorithm 2

TCCA for two tensors of modes D_x and D_y, respectively, assuming (i) CP structure on canonical correlation tensors ($\mathscr{V}, \mathscr{W}$) and (ii) separable covariances Var(vec($\mathscr{X}$)) and Var(vec($\mathscr{Y}$))

Initialize $\mathbf{V}_i^{(0)}, i \in [D_x]$

$\mathbf{W}_j^{(0)}, j \in [D_y]$

$\mathbf{H}_{\mathbf{x}}^{(0)} \leftarrow *_i\left(\mathbf{V}_i^{(0)T}\widehat{\mathbf{\Sigma}}_{\mathbf{x},i}\mathbf{V}_i^{(0)}\right)$

$\mathbf{H}_{\mathbf{y}}^{(0)} \leftarrow *_i\left(\mathbf{W}_j^{(0)\top}\widehat{\Sigma}_{y,j}\mathbf{W}_j^{(0)}\right)$

$t \leftarrow 0$

repeat

Select $(i, j) \in [D_x] \times [D_y]$

$\mathbf{H}_{\mathbf{x},-i}^{(t)} \leftarrow \mathbf{H}_{\mathbf{x}}^{(t)}\tau\left(\mathbf{V}_i^{(t)\top}\widehat{\mathbf{\Sigma}}_{\mathbf{x},i}\mathbf{V}_i^{(t)}\right)$

$\mathbf{C}_{\mathbf{x}}^{(t)} \leftarrow \mathbf{H}_{\mathbf{x},-i}^{(t)} \otimes \widehat{\mathbf{\Sigma}}_{\mathbf{x},i}$

$\mathbf{H}_{\mathbf{y},-j}^{(t)} \leftarrow \mathbf{H}_{\mathbf{y}}^{(t)}\tau\left(\mathbf{W}_j^{(t)T}\widehat{\Sigma}_{\mathbf{y},j}\mathbf{W}_j^{(t)}\right)$

$\mathbf{C}_{\mathbf{y}}^{(t)} \leftarrow \mathbf{H}_{\mathbf{y},-j}^{(t)} \otimes \widehat{\Sigma}_{\mathbf{y},j}$

$\mathbf{V}_{(-i)}^{(t)} \leftarrow \mathbf{V}_{D_{\mathrm{x}}}^{(t)} \odot \cdots \odot \mathbf{V}_{i+1}^{(t)} \odot \mathbf{V}_{i-1}^{(t)} \odot \cdots \odot \mathbf{V}_1^{(t)}$

$\mathbf{W}_{(-j)}^{(t)} \leftarrow \mathbf{W}_{D_{\mathrm{y}}}^{(t)} \odot \cdots \odot \mathbf{W}_{j+1}^{(t)} \odot \mathbf{W}_{j-1}^{(t)} \odot \cdots \odot \mathbf{W}_1^{(t)}$

$\mathbf{C}_{\mathbf{x}y}^{(\mathrm{t})} \leftarrow \left(\mathbf{V}_{-i}^{(\mathrm{t})T} \otimes \mathbf{I}_{p_i}\right)\widehat{\mathbf{\Sigma}}_{\mathbf{x}_{(i)},\mathbf{y}_{(j)}}\left(\mathbf{W}_{-j}^{(\mathrm{t})\mathrm{T}} \otimes \mathbf{I}_{q_j}\right)$

Solve following generalized eigenvalue decomposition, $\begin{bmatrix} \mathbf{0} & \mathrm{C}_{\mathrm{xy}}^{(\mathrm{t})} \\ \mathrm{C}_{\mathrm{xy}}^{(\mathrm{t})} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathrm{v}_i^{(\mathrm{t}+1)} \\ \mathrm{W}_j^{(\mathrm{t}+1)} \end{bmatrix} = \rho^{(\mathrm{t}+1)}\begin{bmatrix} \mathrm{C}_{\mathrm{x}}^{(\mathrm{t})} & \mathbf{0} \\ \mathbf{0} & \mathrm{C}_{\mathrm{y}}^{(\mathrm{t})} \end{bmatrix}\begin{bmatrix} \mathrm{V}_i^{(\mathrm{t}+1)} \\ \mathrm{W}_j^{(\mathrm{t}+1)} \end{bmatrix}$

$\mathbf{H}_{\mathbf{x}}^{(t+1)} \leftarrow \mathbf{H}_{\mathbf{x},-i}^{(t)} * \left(\mathbf{V}_i^{(t+1)T}\widehat{\mathbf{\Sigma}}_{\mathbf{x},i}\mathbf{V}_i^{(t+1)}\right)$

$\mathbf{H}_{\mathbf{y}}^{(t+1)} \leftarrow \mathbf{H}_{\mathbf{y},-j}^{(t)} * \left(\mathbf{W}_i^{(t+1)T}\widehat{\mathbf{\Sigma}}_{\mathbf{y},j}\mathbf{W}_j^{(t+1)}\right)$

$t \leftarrow t+1$

until $\rho^{(t)}$ converges

4.2.1 | Estimation of separable covariance matrices—

Algorithm 2 relies on the sample estimate of separable covariances $\widehat{\Sigma}_{\mathbf{x},i}$ and $\widehat{\Sigma}_{\mathrm{y},j}$, and the unstructured cross-covariance $\widehat{\Sigma}_{\mathbf{x},\mathrm{y}}$. The following lemma will be useful in computing a consistent estimator for covariance matrices with the separable structure.

Lemma 1: If a random tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \ldots \times p_{D_X}}$ has mean zero and separable covariance $\mathrm{Var}(\mathbf{x}) = \Sigma_{\mathbf{x},D_\mathrm{x}} \otimes \ldots \otimes \Sigma_{\mathbf{x},1}$, then

$$\mathbb{E}\left(\mathbf{X}_{(i)}\mathbf{X}_{(i)}^{\top}\right) = \left(\prod_{i' \neq i} tr\left(\Sigma_{\mathbf{x},i'}\right)\right)\Sigma_{\mathbf{x},i} \text{ and } \mathbb{E}\left(\|\mathbf{x}\|_2^2\right) = \prod_{i=1}^{D_\mathrm{x}} tr\left(\Sigma_{\mathbf{x},i}\right).$$

Proof. For the first identity, see Hoff, 2011 (2011, Proposition 2.1). For the second identity, $\mathbb{E}\left(\|\mathbf{x}\|_2^2\right) = \mathrm{tr}(\mathrm{Var}(\mathbf{x})) = \mathrm{tr}\left(\Sigma_{\mathbf{x},\mathbf{D}_x} \otimes \ldots \otimes \Sigma_{\mathbf{x},1}\right) = \Pi_i \mathrm{tr}\left(\Sigma_{\mathbf{x},1}\right)$. □

Given N i.i.d. observations $(\mathbf{x}_n, \mathbf{y}_n)$, the estimators $\hat{r}_x = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \bar{\mathbf{x}}\|_2^2$ and $\hat{r}_y = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}_n - \bar{\mathbf{y}}\|_2^2$ consistently estimate $\Pi_{i=1}^{D_\mathrm{x}}\mathrm{tr}\left(\Sigma_{\mathrm{x},i}\right)$ and $\prod_{j=1}^{D_\mathrm{y}}\mathrm{tr}\left(\Sigma_{\mathrm{y},j}\right)$, respectively, where $\bar{x}$ and $\bar{y}$ are the sample means of the vectorized tensors $\mathbf{x}_n$ and $\mathbf{y}_n$. We propose the following covariance estimators:

$$\widehat{\Sigma}_{\mathrm{x},i} = \frac{1}{N\hat{r}_\mathrm{x}^{(D_\mathrm{x}-1)/D_\mathrm{x}}}\sum_{n=1}^{N}\left(x_{n(i)} - \bar{x}_{(i)}\right)\left(x_{n(i)} - \bar{x}_{(i)}\right)^{\top},\ \widehat{\Sigma}_{y,j}$$
$$= \frac{1}{N\hat{r}_y^{(D_y-1)/D_y}}\sum_{n=1}^{N}\left(y_{n(j)} - \bar{y}_{(j)}\right)\left(y_{n(j)} - \bar{y}_{(j)}\right)^{T},$$

$$\widehat{\Sigma}_{\mathrm{x,y}} = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x}_n - \bar{x})(\mathrm{y}_n - \bar{y})^{T},$$

where $i \in [D_\mathrm{x}]$ and $j \in [D_\mathrm{y}]$. The vectors $\mathbf{x}_{n(\mathrm{i})}$ and $\bar{x}_{(\mathrm{i})}$ denote the mode i vectorization of the nth observation and the sample mean of the mode i vectorized tensors $\mathbf{x}_n$, respectively. The vectors $\mathbf{y}_{n(\mathrm{j})}$ and $\bar{y}_{(\mathrm{j})}$ denote analogous vectorizations.

Unfortunately, the separable covariance structure (8) is not identifiable in the individual $\Sigma_{\mathrm{x},i}$ due to scaling indeterminacy. Therefore, $\Sigma_{\mathrm{x},i}$ cannot be consistently estimated. Note, however, that we do not need to consistently estimate the individual $\Sigma_{\mathrm{x},i}$ in order to consistently estimate their Kronecker product. To see this, note that by Slutsky's theorem,

$$\widehat{\Sigma}_{\mathrm{x},D_\mathrm{x}} \otimes \ldots \otimes \widehat{\Sigma}_{\mathrm{x},1} \to \frac{\left(\prod_i \mathrm{tr}\left(\Sigma_{\mathrm{x},j}\right)\right)^{D_\mathrm{x}-1}}{\left(\prod_i \mathrm{tr}\left(\Sigma_{\mathrm{x},j}\right)^{D_\mathrm{x}-1}\right.}\Sigma_{\mathrm{x},D_\mathrm{x}} \otimes \ldots \otimes \Sigma_{\mathrm{x},1} = \Sigma_{\mathrm{x},D_\mathrm{x}} \otimes \ldots \otimes \Sigma_{\mathrm{x},1}$$

consistently estimates Var(**x**).

Note that Hoff (2011) proposes an iterative algorithm for finding the maximum likelihood estimation (MLE) on the basis of the array normal assumption. The maximum likelihood estimation may improve upon the above estimates when data actually come from an array normal distribution.

5 | SPARSE TCCA

We may also recover sparse canonical tensors for both TCCA and scTCCA to enhance the interpretability of the estimated canonical tensors. Following the iterative thresholding strategy introduced by Ma (202013) for sparse principal component analysis, by Wang, Gu, Ning, and Liu (2015) for sparse expectation-maximization algorithms, and by Tan et al. (2018) for generalized eigenvalue problems, we incorporate a hard-thresholding step in Algorithms 1 and 2, as follows:

$$\mathbf{w}_i^{(t+1)} \leftarrow \Theta_\lambda\left(\mathbf{w}_i^{(t+1)}\right) \text{ and } \mathbf{w}_j^{(t+1)} \leftarrow \Theta_\lambda\left(\mathbf{w}_j^{(t+1)}\right),$$

where $\Theta_\lambda(\mathbf{v})$ performs element-wise hard-thresholding on **v**, namely, the ith element of $\Theta_\lambda(\mathbf{v})$ is v_i if $|v_i| > \lambda$ and 0 otherwise. In the simulation studies of Section 7, we employ fivefold cross-validation with a grid point search to choose the tuning parameter λ, following Tan, Wang, Liu, and Zhang (2018). Instead of a grid point search, a random search would be another choice (Bergstra & Bengio, 2012).

6 | PROBABILISTIC MODEL FOR TCCA

Bach and Jordan (2006) give a probabilistic interpretation for the classic CCA, which enables us to simulate data with desired canonical vectors and canonical correlations. For TCCA without any assumption on the covariance structure, it is the same as the regular CCA by treating vectorized canonical tensors as canonical vectors. In this section, we first discuss how to generate data from given d canonical correlations $\rho_d = (\rho_1, \ldots, \rho_d)$ and their corresponding canonical vectors in columns of matrices $(\mathbf{V}_d, \mathbf{W}_d)$, and then we extend it to TCCA with the separable covariance assumption. Let Σ_x and Σ_y be two covariance matrices such that $V_d^\top \Sigma_x V_d = W_d^\top \Sigma_y W_d = I_d$. Define two linear transformations $A_x = \Sigma_x V_d M_x$ and $A_y = \Sigma_y W_d M_y$, where $M_x, M_y \in \mathbb{R}^{d \times d}$ are arbitrary matrices such that $M_x M_y^\top = \text{diag}(\rho_d)$ and their spectral norms are less 1. We consider the latent factor model

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_d)$$

$$\mathbf{x} \mid \mathbf{Z} = \mathbf{z} \sim N\left(\mathbf{A}_\mathbf{x}\mathbf{z} + \boldsymbol{\mu}_\mathbf{x}, \Sigma_\mathbf{x} - \mathbf{A}_\mathbf{x}\mathbf{A}_\mathbf{x}^\top\right)$$

$$\mathbf{y} \mid \mathbf{Z} = \mathbf{z} \sim N\left(\mathbf{A}_\mathbf{y}\mathbf{z} + \boldsymbol{\mu}_\mathbf{y}, \Sigma_\mathbf{y} - \mathbf{A}_\mathbf{y}\mathbf{A}_\mathbf{y}^\top\right).$$

The joint distribution of $(\mathbf{x}, \mathbf{y})$ is

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim N\left(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_x & \Sigma_{xy}\\ \Sigma_{xy}^\top & \Sigma_y\end{bmatrix}\right),$$

where $\Sigma_{xy} = \Sigma_x V_d \mathrm{diag}\,(\rho_d) W_d^\top \Sigma_y$.

Now, we discuss how to construct Σ_x and Σ_y from $(\mathbf{V}_d, \mathbf{W}_d)$, which is not described in Bach and Jordan (2006). Let $\mathbf{V}_d = \mathbf{Q}_\mathbf{x}\mathbf{R}_\mathbf{x}$ be the thin QR decomposition of $\mathbf{V}_d$. Then

$$\Sigma_x = Q_x R_x^{-T} R_x^{-1} Q_x^T + T_x\left(I_p - Q_x Q_x^T\right)T_x^T$$

satisfies $\mathbf{V}_d^\top \Sigma_x \mathbf{V}_d = I_d$ for arbitrary $T_x \in \mathbb{R}^{p \times p}$. Similarly, let $\mathbf{W}_d = \mathbf{Q}_\mathbf{y}\mathbf{R}_\mathbf{y}$ be the thin QR decomposition of $\mathbf{W}_d$. Then

$$\Sigma_y = Q_y R_y^{-T} R_y^T Q_y^T + T_y\left(I_q - Q_y Q_y^T\right)T_y^T$$

satisfies $\mathbf{W}_d^\top \Sigma_y \mathbf{W}_d = \mathbf{I}_d$ for arbitrary $\mathrm{T_y} \in \mathbb{R}^{q \times q}$. In this notation, the joint covariance is

$$\begin{aligned}\mathrm{Var}\begin{pmatrix}\mathrm{x}\\ \mathrm{y}\end{pmatrix} &= \begin{pmatrix}\mathrm{Q_x R_x^{-T} R_x^T Q_x^T} & \mathrm{Q_x R_x^{-T}\,diag}\,(\rho_d)\mathrm{R_y^T Q_y^T}\\ \mathrm{Q_y R_y^{-T}\,diag}\,(\rho_d)\mathrm{R_y^T Q_y^T} & \mathrm{Q_x R_x^{-T} R_y^T Q_y^T}\end{pmatrix}\\ &+ \begin{pmatrix}\mathrm{T_x\left(I_p - Q_x Q_x^\top\right)T_x^T} & 0\\ 0 & \mathrm{T_y\left(I_q - Q_y Q_y^T\right)T_y^T}\end{pmatrix}.\end{aligned} \tag{9}$$

$\mathbf{T}_\mathbf{x}$ and $\mathbf{T}_\mathbf{y}$ are free parameters that adjust the noise level in $\mathbf{x}$ and $\mathbf{y}$, respectively. The normal generative model is

$$\mathbf{z} \sim N(0, I_d)$$

$$\mathbf{x} \mid \mathbf{z} = \mathbf{z} \sim N\left(\mathbf{Q}_\mathbf{x}\mathbf{R}_\mathbf{x}^{-T}\mathbf{M}_\mathbf{x}\mathbf{z} + \boldsymbol{\mu}_\mathbf{x}, \Sigma_\mathbf{x}\right)$$

$$\mathbf{y} \mid \mathbf{z} = \mathbf{z} \sim N\left(\mathbf{Q}_\mathbf{y}\mathbf{R}_\mathbf{y}^{-T}\mathbf{M}_\mathbf{y}\mathbf{z} + \boldsymbol{\mu}_\mathbf{y}, \Sigma_\mathbf{y}\right)$$

where Σ_x and Σ_y are as in Equation (9). The normality is not essential for construction of this covariance structure.

The separable covariance structure brings complication. To account for this parsimonious structure in the generative model, we construct marginal factor matrices $\Sigma_{x,i}$ and $\Sigma_{y,j}$ that satisfy $\mathbf{V}_i^\top \Sigma_{\mathbf{x},i}\mathbf{V}_i = R_x^{1/D_x}\mathbf{I}_{R_x}$ and $\mathbf{W}_j^\top \Sigma_{y,j}\mathbf{W}_j = R_y^{1/D_Y}\mathbf{y}_{R_y}$. Let $\mathrm{V}_i = \mathrm{Q}_{x,i}\mathrm{R}_{x,j}$ and $\mathrm{W}_j = \mathrm{Q}_{y,j}\mathrm{R}_{y,j}$ be the thin QR decompositions. Then

$$\Sigma_{x,i} = R_x^{-D_x} Q_{x,j} R_{x,y}^{-T} R_{x,j}^{-1} Q_{x,j}^T + T_{x,j}\left(I_{D_i} - Q_{x,i}Q_{x,j}^T\right)T_{x,j}^T$$

$$\Sigma_{y,j} = R_y^{-D_y} Q_{y,j} R_{y,j}^{-T} R_{y,j}^{-1} Q_{y,j}^T + T_{y,j}\left(I_{q_j} - Q_{y,j}Q_{y,j}^T\right)T_{y,j}^T$$

satisfy the conditions with arbitrary $\mathbf{T}_{x,i} \in \mathbb{R}^{p_i \times p_i}$ and $\mathbf{T}_{y,j} \in \mathbb{R}^{q_j \times q_j}$. And we have the desired property

$$\begin{aligned}
&\mathrm{vec}(\mathbf{V})^T\left(\Sigma_{\mathbf{x},D_x} \otimes \ldots \otimes \Sigma_{\mathbf{x},1}\right)\mathrm{vec}(\mathbf{V}) \\
&= 1_{R_x}^T\left(V_{D_x} \odot \ldots \odot V_1\right)^T\left(\Sigma_{x,D_x} \otimes \ldots \otimes \Sigma_{x,1}\right)\left(V_{D_x} \odot \ldots \odot V_1\right)1_{R_x} \\
&= 1_{R_x}^\top\left(*_i V_i^\top \Sigma_{x,i} V_i\right)1_x \\
&= R_x^{-1} 1_{R_x}^T I_{R_x} 1_{R_x} \\
&= 1 .
\end{aligned}$$

Similarly, $\mathrm{vec}(\mathbf{W})^T\left(\Sigma_{\mathbf{y},D_y} \otimes \ldots \otimes \Sigma_{\mathbf{y},1}\right)\mathrm{vec}(\mathbf{W}) = 1$.

7 | NUMERICAL EXPERIMENTS

We use the generative model described in Section 6 to evaluate the methods discussed in this paper: classic CCA, TCCA, scTCCA, sparse TCCA, and sparse TCCA with the separable covariance. We assess these methods on their ability to recover the true latent population parameters $(\mathscr{V}, \mathscr{W})$ used to generate i.i.d. samples of tensor data pairs $(\mathscr{X}_n, \mathscr{Y}_n)$ for $n \in [1000]$. In all examples, the true $\mathscr{V}$ is a vector of length 100 with six entries set to 1 and the rest to 0. We use three different latent $\mathscr{W} \in \mathbb{R}^{64\times 64}$ shown in Figure 1: $\mathscr{W}_{\text{rectangle}}$, $\mathscr{W}_{\text{cross}}$ and $\mathscr{W}_{\text{butterfly}}$. White pixels indicate values of 1, and black pixels indicate values of 0. The rectangle and cross-population canonical tensors $\mathscr{W}_{\text{rectangle}}$ and $\mathscr{W}_{\text{cross}}$ are low rank; specifically, they are Rank 1 and Rank 2, respectively. The butterfly population canonical tensor $\mathscr{W}_{\text{butterfly}}$ is a high rank. At the first glance, these illustrative examples do not come across as challenging estimation problems in the high dimension-low sample size regime, but if we were to vectorize the data and perform CCA, the number of parameters to fit is $100 + 64^4 = 4196$ whereas the sample size is 1,000. Because the number of parameters exceeds the sample size, the sample covariance matrix is singular. Consequently, we add a small ridge term $10^{-3}|$ to sample covariance matrices so that the generalized eigenvalue problem has a unique solution. The code for generating the simulation results is provided in the Supporting Information.

7.1 | Evaluation criteria and selection of tuning parameter

Canonical tensors can only be estimated up to a scaling factor. Thus, we measure estimation accuracy by the angle between population canonical tensors used to generate the data $(\mathcal{V}, \mathcal{W})$ and estimated canonical tensors $(\widehat{\mathcal{V}}, \widehat{\mathcal{W}})$, as follows:

$$\angle(\mathcal{V}, \widehat{\mathcal{V}}) = \frac{\langle \mathrm{v}, \widehat{\mathrm{V}} \rangle}{\|\mathrm{v}\|_2 \|\widehat{\mathrm{V}}\|_2}, \text{ and } \angle(\mathcal{W}, \widehat{\mathcal{W}}) = \frac{\langle \mathrm{w}, \widehat{\mathrm{W}} \rangle}{\|\mathrm{w}\|_2 \|\widehat{\mathrm{W}}\|_2}.$$

The angles can take on values from −1 to 1 where an angle closer to 1 indicates better recovery of the true canonical tensors.

We use k-fold cross-validation to perform model selection, for example, selection of the tensor rank or the sparsity level. We split our data into K equally sized groups; for $k \in [\mathrm{K}]$, we estimate a pair $(\mathcal{V}_{-k}, \mathcal{W}_{-k})$ on all but the kth fold of data for a sequence of models of varying complexity, $\mathcal{M}_1, \mathcal{M}_2, \ldots$, where model $\mathcal{M}_1$ has a fixed pair of tensor ranks R_x and R_y and sparsity inducing parameter λ. We denote the fitted canonical correlation using model $\mathcal{M}_1$ by $\hat{\rho}_{-k}(\mathcal{V}_{-k}, \mathcal{W}_{-k}; \mathcal{M}_l)$ Using $(\mathcal{V}_{-k}, \mathcal{W}_{-k})$, we compute the canonical correlation on the held out jth fold, which we denote $\hat{\rho}_{-j}(\mathcal{V}_{-k}, \mathcal{W}_{-k})$. We choose the model that minimizes the average discrepancy between the empirical canonical correlations on the training sets and testing sets (Waaijenborg, Verselewel de Witt Hamer, & Zwinderman, 2008).

$$\widehat{\mathcal{M}} = \mathcal{M}_l \operatorname{argmin} \frac{1}{\mathrm{K}} \sum_{k=1}^{K} |\hat{\rho}_k(\mathcal{V}_{-k}, \mathcal{W}_{-k}) - \hat{\rho}_{-k}(\mathcal{V}_{-k}, \mathcal{W}_{-k}; \mathcal{M}_l)|.$$

We then use all the data to estimate $(\mathcal{V}, \mathcal{W})$ using model $\widehat{\mathcal{M}}$.

7.2 | Results

Figure 2 compares estimation accuracy of various methods over 100 replicates. First of all, we observe that the various versions of TCCA outperform the CCA on the vectorized data for all three choices of the latent canonical tensor $\mathcal{W}$. Second, we notice that there appears to be some overfitting in the case of rectangle and cross problems. In the rectangle problem, where the population canonical tensor $\mathcal{W}_{\text{rectangle}}$ is a Rank 1 tensor, all four TCCA methods show the best performance when we fix the rank R_y for estimating $\mathcal{W}$ to be 1, and the performance deteriorates as higher ranks are used. The same trend can be seen in the result of the cross problem where the population canonical tensor $\mathcal{W}_{\text{cross}}$ is a Rank 2 tensor. All TCCA methods give smaller values of angle when $R_y = 3$ is used compared with when $R_y = 2$ is used. In contrast, in the case of the butterfly image, the rank of population canonical tensor $\mathcal{W}_{\text{butterfly}}$ is much greater than 3. Thus, we do not expect an overfitting problem as well as good estimation performance as in the low-rank cases. We confirm that in Figure 2, calculated angles from the butterfly problem are lower than those from rectangles or cross problem.

Another interesting observation is that results from TCCA methods with a separable covariance structure show better performance when we use a higher value for the rank parameter R_y than the true value than do other two models in cross and butterfly problems. This results implies that assuming a separable covariance structure improves the estimation accuracy in the two problems. This can be explained by the true image $\mathcal{X}$, which has a symmetric structure. Due to this special structure, we possibly expect improved performance from the more parsimonious model. However, this structured model does not have much effect on the rectangle problem. This may be because the true canonical tensor $\mathcal{W}$ already possess relatively few parameters. Note that the rectangle image is also symmetric but Rank 1, which has very few parameters.

Table 1 shows computation times taken for each method. In most cases, the computation times of TCCA methods are better or comparable with those of the CCA method. Especially, there is huge improvement when the underlying true canonical tensor has a low rank. Also, sparse models take less time than do nonsparse models, and separable covariance structure models take less time than do the models without the assumption for computation in general, which is expected.

8 | DISCUSSION

We have proposed a TCCA approach for finding the relationship between linear combinations of two tensor datasets. Our method combines the classic CCA approach and the low-rank tensor decomposition to reduce the vast dimensionality of tensor parameters. The proposed estimation algorithm scales well with the tensor data size and is easy to implement using existing statistical software. Our algorithms also support convergence guarantees, properties arguably lacking in the alternatives in the literature. We briefly highlight some open problems for further investigation.

In our illustrative examples, we did not incorporate a procedure for selecting the rank parameter. A main motivation for the low-rank formulation of the canonical tensors is reduced computation. Thus, one may wish to choose as high a rank parameter as one's computational budget allows. Nonetheless, we leave for future work a more principled approach to rank selection. One promising angle of pursuit would be to leverage the equivalence between CCA and a least squares problem (Sun, Ji, & Ye, 2008) and then derive an information criterion, a strategy commonly used to select the rank parameter in several recently proposed tensor estimation procedures (Zhou et al., 2013; Sun et al., 2017; Sun & Li, 2019).

An additional direction for future work is to generalize our TCCA framework to handle the analysis of multiple tensor datasets. There are numerous proposed methods to extend CCA for multiple vector-measurement datasets (Carroll, 1968; Kettenring, 1971; Hanafi, 2007; Witten & Tibshirani, 2009; Luo et al., 2015) but not for multiple tensor datasets to the best of our knowledge.

Another problem not tackled in the paper includes verifying the consistency of the proposed estimation procedures. Our clear specification of the population models provides a

framework for studying the consistency property in both the large n fixed p (Zhou et al., 2013) and the large n diverging p settings (Zhang, Li, Zhou, Zhou, & Shen, 2019). Finally, we are currently investigating our methods on a real dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

Bach FR, & Jordan MI (2006). A probabilistic interpretation of canonical correlation analysis, Technical Report.

Bergstra J, & Bengio Y (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281–305.

Bickel PJ, & Levina E (2008). Covariance regularization by thresholding. The Annals of Statistics, 36(6), 2577–2604.

Bickel PJ, & Levina E (2008). Regularized estimation of large covariance matrices. The Annals of Statistics, 36(1), 199–227.

Cai TT, & Yuan M (2012). Adaptive covariance matrix estimation through block thresholding. The Annals of Statistics, 40(4), 2014–2042.

Carroll JD (1968). Generalization of canonical correlation analysis to three or more sets of variables, Proceedings of the 76th Annual Convention of the American Psychological Association, Washington, DC: Vol. 3, pp. 227–228.

Carroll JD, & Chang J-J (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. Psychometrika, 35, 283–319.

Gang L, Yong Z, Yan-Lei L, & Jing D (2011). Three dimensional canonical correlation analysis and its application to facial expression recognition, International Conference on Intelligent Computing and Information Science Berlin, Heidelberg: Springer, pp. 56–61.

González I, Déjean S, Martin P, & Baccini A (2008). CCA: An R package to extend canonical correlation analysis. Journal of Statistical Software, 23(1), 1–14.

Hanafi M (2007). PLS path modelling: Computation of latent variables with the estimation mode b. Computational Statistics, 22(2), 275–292.

Harshman RA (1970). Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis, UCLA Working Papers in Phonetics, 16, 1–84.

Hoff PD (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. Bayesian Anal, 6(2), 179–196.

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

Kettenring JR (1971). Canonical analysis of several sets of variables. Biometrika, 58(3), 433–451.

Kolda TG, & Bader BW (2009). Tensor decompositions and applications. SIAM Review, 51(3), 455–500.

Kubokawa T, Srivastava MS, et al. (2013). Optimal ridge-type estimators of covariance matrix in high dimension: CIRJE, Faculty of Economics, University of Tokyo.

Ledoit O, & Wolf M (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411.

Ledoit O, &Wolf M (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics, 40(2), 1024–1060.

Lee SH, & Choi S (2007). Two-dimensional canonical correlation analysis. Signal Processing Letters, IEEE, 14(10), 735–738.

Lu H (2013). Learning canonical correlations of paired tensor sets via tensor-to-vector projection, Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, IJCAI ‘13 Beijing, China: AAAI Press, pp. 1516–1522.

Luo Y, Tao D, Ramamohanarao K, Xu C, & Wen Y (2015). Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Transactions on Knowledge and Data Engineering, 27(11), 3111–3124.

Ma Z (2013). Sparse principal component analysis and iterative thresholding. The Annals of Statistics, 41(2), 772–801.

Srivastava MS, & Reid N (2012). Testing the structure of the covariance matrix with fewer observations than the dimension. Journal of Multivariate Analysis, 112, 156–171.

Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW & Thompson PM (2010). Voxelwise genome-wide association study (vGWAS). Neuroimage, 53(3), 1160–1174. [PubMed: 20171287]

Sun L, Ji S, & Ye J (2008). A least squares formulation for canonical correlation analysis, Proceedings of the 25th International Conference on Machine Learning, ICML '08 New York, NY, USA: ACM, pp. 1024–1031.

Sun WW, & Li L (2019). Dynamic tensor clustering. Journal of the American Statistical Association, in press.

Sun WW, Lu J, Liu H, & Cheng G (2017). Provable sparse tensor decomposition. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3), 899–916.

Tan KM, Wang Z, Liu H, & Zhang T (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(5), 1057–1086.

Vinod HD (1976). Canonical ridge and econometrics of joint production. J Econ, 4(2), 147–166.

Waaijenborg S, Verselewel de Witt Hamer PC, & Zwinderman AH (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Statistical Applications in Genetics and Molecular Biology, 7(1), 3.

Wang H (2010). Local two-dimensional canonical correlation analysis. Signal Processing Letters, IEEE, 17(11), 921–924.

Wang Z, Gu Q, Ning Y, & Liu H (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality In Cortes C, Lawrence ND, Lee DD, Sugiyama M, & Garnett R (Eds.), Advances in Neural Information Processing Systems 28: Curran Associates, Inc, pp. 2521–2529.

Wang S-J, Yan W-J, Sun T, Zhao G, & Fu X (2016). Sparse tensor canonical correlation analysis for micro-expression recognition. Neurocomputing, 214, 218–232.

Witten DM, & Tibshirani RJ (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology, 8(1), 1–27.

Yan J, Zheng W, Zhou X, & Zhao Z (2012). Sparse 2-D canonical correlation analysis via low rank matrix approximation for feature extraction. Signal Processing Letters, IEEE, 19(1), 51–54.

Zhang X, Li L, Zhou H, Zhou Y, & Shen D (2019). Tensor generalized estimating equations for longitudinal imaging analysis. Statistica Sinica, 29, 1977–2005. [PubMed: 32523321]

Zhou H, Li L, &Zhu H (2013). Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association, 108(502), 540–552. [PubMed: 24791032]

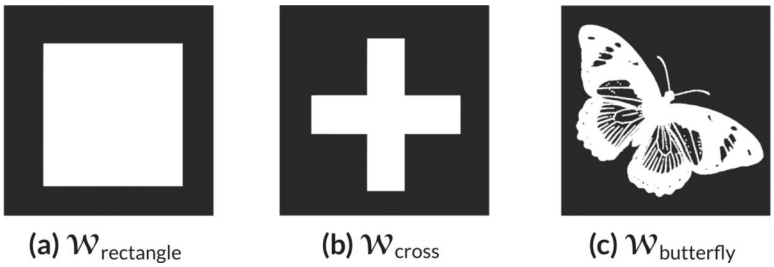

FIGURE 1.
True latent population $\mathcal{W}$ canonical tensors for numerical experiments[dummy]

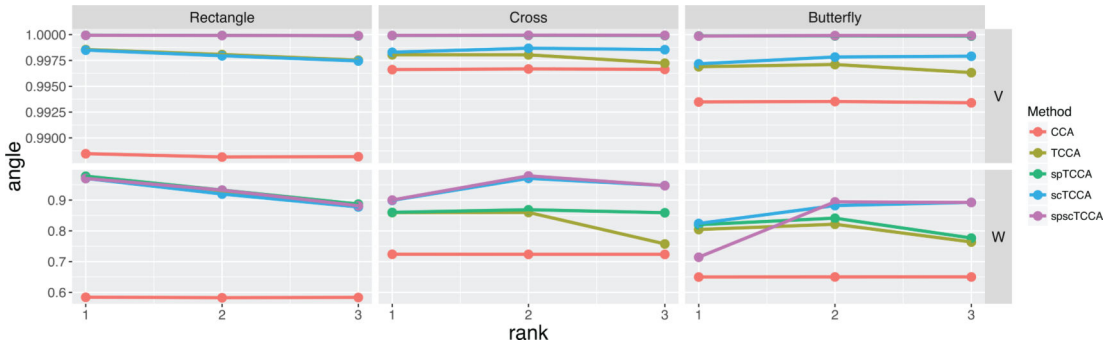

FIGURE 2.

Angles between the recovered canonical vector/tensors and the true canonical vector/tensors. CCA, canonical correlation analysis; scTCCA, tensor canonical correlation analysis with separable covariance; spscTCCA, sparse tensor canonical correlation analysis with separable covariance; spTCCA, sparse tensor canonical correlation analysis; TCCA, tensor canonical correlation analysis

TABLE 1

Computation time: The mean run times are reported in seconds with the standard deviation in parentheses

Problem	Rank	CCA	TCCA	spTCCA	scTCCA	spscTCCA
Rectangle	1	22.90 (3.62)	7.52 (1.10)	7.36 (0.95)	3.14 (0.55)	3.09 (0.44)
	2	22.80 (3.18)	15.32 (2.51)	15.12 (2.17)	11.03 (2.03)	30.11 (8.78)
	3	21.85 (3.36)	24.07 (3.91)	18.61 (3.14)	15.89 (2.95)	37.09 (12.09)
Cross	1	18.49 (2.32)	8.05 (1.34)	7.78 (1.40)	3.39 (0.53)	13.19 (12.14)
	2	17.33 (2.43)	17.31 (3.32)	15.13 (5.02)	6.96 (1.11)	33.55 (12.78)
	3	17.10 (2.55)	24.44 (3.79)	54.86 (25.73)	10.04 (1.84)	9.63 (2.05)
Butterfly	1	18.92 (3.48)	7.08 (1.60)	61.18 (28.58)	2.90 (0.64)	31.66 (6.36)
	2	18.23 (2.92)	20.01 (4.53)	30.10 (14.87)	9.89 (2.15)	48.57 (13.14)
	3	19.99 (2.55)	32.27 (4.68)	26.09 (3.89)	11.84 (2.06)	13.58 (5.89)

Note. Experiments were performed on a computer cluster consisting of machines with Intel Xeon-based processors (I5 or I7 processors) with RAM ranging from 1,600 to 2,100 MHz. Abbreviations: CCA, canonical correlation analysis; scTCCA, tensor canonical correlation analysis with separable covariance; spscTCCA, sparse tensor canonical correlation analysis with separable covariance; spTCCA, sparse tensor canonical correlation analysis; TCCA, tensor canonical correlation analysis