Sparse Generalized Canonical Correlation Analysis: Distributed Alternating Iteration-Based Approach

Kexin Lv
Kelen_Lv@sjtu.edu.cn
Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

Jia Cai
jiacai1999@gdufe.edu.cn
School of Digital Economics, Guangdong University of Finance and Economics, Guangzhou 510320, China

Junyi Huo
jh4a19@soton.ac.uk
ByteDance Ltd., Beijing 100089, China

Chao Shang
c-shang@tsinghua.edu.cn
Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

Xiaolin Huang
xiaolinhuang@sjtu.edu.cn

Jie Yang
jieyang@sjtu.edu.cn
Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

Sparse canonical correlation analysis (CCA) is a useful statistical tool to detect latent information with sparse structures. However, sparse CCA, where the sparsity could be considered as a Laplace prior on the canonical variates, works only for two data sets, that is, there are only two views or two distinct objects. To overcome this limitation, we propose a sparse generalized canonical correlation analysis (GCCA), which could detect the latent relations of multiview data with sparse structures. Specifically, we convert the GCCA into a linear system of equations and impose $\ell_1$ minimization penalty to pursue sparsity. This results in a nonconvex problem on the Stiefel manifold. Based on consensus optimization, a

Jie Yang and Xiaolin Huang are the corresponding authors.

Neural Computation 36, 1380–1409 (2024) © 2024 Massachusetts Institute of Technology
https://doi.org/10.1162/neco_a_01673
distributed alternating iteration approach is developed, and consistency is investigated elaborately under mild conditions. Experiments on several synthetic and real-world data sets demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Canonical correlation analysis (CCA), launched by Hotelling (1936), is a celebrated statistical tool for finding the correlation between two sets of multidimensional variables. The two sets can be considered as two views of one object or a view of two distinct objects. The main aim of CCA is to find two sets of canonical variables (weight vectors) such that the projected variables in the lower-dimensional space are maximally correlated. Due to its efficiency of finding latent information, CCA has been widely used in many branches of signal processing and data analysis, including cross-language document retrieval (Vinokourov et al., 2003), genomic data analysis (Yamanishi et al., 2003), video surveillance application (Hu, 2014; Sun et al., 2023), and functional magnetic resonance imaging (Friman et al., 2001). Both theoretical and algorithmic aspects of CCA have been extensively studied in the literature, as demonstrated by works such as Sun et al. (2023), Cai and Sun (2017), Gao et al. (2019), Ge et al. (2016), Ma et al. (2015), Michaeli et al. (2016), Koide-Majima and Majima (2021), Tan et al. (2018), and Wang et al. (2015).

However, a large portion of features are not informative for high-dimensional data in the field of data analysis. When the canonical variables involve all features in the original space, the canonical variates are usually not sparse, making the interpretation of the solutions difficult. To facilitate the interpretation of canonical variates in high-dimensional data analysis, introducing sparsity has been a common approach. Specifically, since the advent of compressive sensing, sparsity has also been found to enhance the performance of various learning methods, provided that a suitable sparse structure can be identified. Similarly, in the field of CCA, there have been many efforts to impose sparsity to pursue higher performance, among them the sparse penalized CCA algorithm (Waaijenborg et al., 2008), the penalized matrix decomposition approach-based sparse CCA method (Witten et al., 2009), and the sparse CCA under primal-dual framework (Hardoon & Shawe-Taylor, 2011). In addition, Chu et al. (2013) selected the sparsest CCA solution from a subset of all solutions via the linearized Bregman method. Chen et al. (2017) developed a precision-adjusted iteration thresholding method to estimate the sparse canonical weights, while Gao et al. (2017) investigated a two-stage-based sparse CCA method, where the first initialization stage was solved by alternating direction method of multipliers (ADMM; Boyd et al., 2011), and then a group-Lasso based method was used to find the sparse weights in the second refinement stage.
In spite of great success, CCA and sparse CCA can handle only two data sets, which heavily limits the applications on multiview analysis and multimodal learning. To overcome this drawback, generalized CCA (GCCA) methods have been proposed. Among several attempts, tensor CCA (Luo et al., 2015), GCCA (Kang et al., 2013), weighted GCCA (Benton et al., 2016), scalable MAX-VAR GCCA (Fu et al., 2017), and Deep GCCA (DGCCA; Benton et al., 2019) have shown good performance to deal with multiple data sets. Similar to CCA, suitably imposing sparsity on GCCA could also improve the performance.

However, the sparsity pursuit method designed for CCA cannot be readily extended to GCCA, and simply copying the technique from CCA to sparse CCA is not applicable for GCCA. To the best of our knowledge, only Kang et al. (2013), Kanatsoulis et al. (2019), and Li et al. (2022) have discussed sparse GCCA methods. Kang et al. (2013) designed a sparse GCCA under the special constraints that the data matrices and the projected variables have multiple regression relationships. Li et al. (2022) integrated the $\ell_2,0$-norm-constrained optimization into GCCA and solved it by the Newton-based method, and Kanatsoulis et al. (2019) discussed a primal-dual decomposition-based approach (PDD) using the ADMM for GCCA in large-scale problems, and its theoretical convergence is guaranteed by introducing Robinson’s condition, which requires the number of canonical components to be far fewer than the number of samples or features. However, the performance of these sparse methods related to GCCA is far from satisfactory in accuracy.

This article aims to introduce a sparse GCCA approach to address the aforementioned challenges. The contributions of this article are summarized as follows:

- We formulate GCCA into the form of a linear system of equations, which serves as the basis for imposing sparsity. This leads to a nonconvex problem on the Stiefel manifold.
- Based on the developed GCCA-related equations and the model demonstrated in Vía et al. (2007), we elegantly develop a novel sparse GCCA algorithm using the distributed alternative iteration method and self-consistent-field (SCF; Zhang et al., 2022) iteration technique, which is a generalization of Boyd’s consensus problem (Boyd et al., 2011).
- The theoretical consistency of the proposed algorithm is judiciously investigated using optimization theory under mild conditions.
- Extensive experiments on gene data and the Europarl data set demonstrate the effectiveness and efficiency of the proposed method.

The remainder of the article is organized as follows. In section 2, we give a brief review of CCA and GCCA. Section 3 is devoted to the design of the new sparse GCCA and its solving algorithm. Section 4 discusses the experimental results. The proof of the main results are in the appendix.
In this section, we briefly review the classic and the generalized CCA.

### 2.1 Canonical Correlation Analysis

Let $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$ be two random variables. Denote $X = (x_1, \ldots, x_m) \in \mathbb{R}^{n_1 \times m}$, $Y = (y_1, \ldots, y_m) \in \mathbb{R}^{n_2 \times m}$. Without loss of generality, we assume both $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$ have zero mean: $\sum_{i=1}^m x_i = 0$ and $\sum_{i=1}^m y_i = 0$. Then CCA solves the following problem,

$$
\max_{w_1 \neq 0, w_2 \neq 0} \quad w_1^T X Y^T w_2,
\quad \text{s.t.} \quad w_1^T X X^T w_1 = 1, \quad w_2^T Y Y^T w_2 = 1.
$$

(2.1)

In equation 2.1, only one pair of canonical variables could be found. To obtain more pairs, Chu et al. (2013) and Hardoon et al. (2004) extended CCA to the multiple version,

$$
\max_{W_1 \neq 0, W_2 \neq 0} \quad \text{Trace}(W_1^T X Y^T W_2)
\quad \text{s.t.} \quad W_1^T X X^T W_1 = I_\ell, \quad W_1 \in \mathbb{R}^{n_1 \times \ell},
\quad W_2^T Y Y^T W_2 = I_\ell, \quad W_2 \in \mathbb{R}^{n_2 \times \ell},
$$

(2.2)

where $I_\ell$ denotes the $\ell \times \ell$ identity matrix, and $\ell$ also stands for the number of columns of $W_i$ ($i = 1, 2$). When $\ell = 1$, equation 2.2 reduces to equation 2.1. Obviously, both CCA and multiple CCA could only deal with two data sets or a two-view data set.

### 2.2 Generalized Canonical Correlation Analysis

To detect the relations of multiple multivariate data sets (more than two), a generalized CCA that considers the sum of correlations was proposed by Carroll (1968),

$$
\min_{W_1 \neq 0, W_2 \neq 0} \quad \sum_{i=1}^J \sum_{j=1, j \neq i}^J \|W_i^T X_i - W_j^T X_j\|_F^2
\quad \text{s.t.} \quad W_i^T X_i X_i^T W_i = I_\ell,
$$

(2.3)

where $F$ denotes the Frobenius norm of a matrix and $J$ stands for the number of views. Equation 2.3 minimizes the sum of distances of each two low-dimensional representations $W_i^T X_i$ and $W_j^T X_j$, which however, is an NP-hard problem. To efficiently study the latent information of multiple-view data sets, MAX-VAR formulation of GCCA was proposed in Kettenring (1971), as follows,
where $G \in \mathbb{R}^{l \times m}$ is a common latent representation of the different views. Different from equation 2.3, equation 2.4 optimizes the latent representation $G$ and the multiple-view canonical variables $\{W_j\}_{j=1}^J$ simultaneously. In the literature, equation 2.4 is solved by selecting the principal eigenvectors of a matrix aggregated from the correlation matrix of different views, that is, the rows of the optimal $G$ are the eigenvectors of the following matrix:

$$M = \sum_{j=1}^J X_j^T (X_j X_j^T)^{-1} X_j.$$ 

Here, we discuss the relationship between the optimization of GCCA, equation 2.4, and multiple CCA, equation 2.3. When $J = 2$, the view of equation 2.4 is deduced to be the same as that of equation 2.3 and becomes

$$\min_{W_1, W_2, G} \|G - W_1^T X_1\|_F^2 + \|G - W_2^T X_2\|_F^2$$

s.t. $GG^T = I_l,$

$$\min_{W_1, W_2, G} \|G - W_1^T X_1\|_F^2 + \|G - W_2^T X_2\|_F^2$$

s.t. $GG^T = I_l.$

$$\min_{W_1, W_2, G} \|G - W_1^T X_1\|_F^2 + \|G - W_2^T X_2\|_F^2$$

s.t. $GG^T = I_l.$

Then its objective function has the following relationship,

$$2\|G - W_1^T X_1\|_F^2 + 2\|G - W_2^T X_2\|_F^2 \geq (a) \|W_1^T X_1 - W_2^T X_2\|_F^2$$

$$= \text{Trace}(W_1^T X_1 X_1^T W_1 + W_2^T X_2 X_2^T W_2 - W_1^T X_1 W_1^T X_2 W_2 - W_1^T X_1 X_1^T W_2 - W_1^T X_2 W_2),$$

where $(a)$ holds true by applying the triangle inequality. One can find that GCCA with $J = 2$ in equation 2.5 is a relaxation of that in multiple CCA, equation 2.2.

3 Sparse GCCA: Model and Algorithm

3.1 New Formulation of Sparse GCCA. In this section, we propose a sparse GCCA model and develop its solving algorithm. Directly adding $\|W_j\|_1$ as an $L_1$ regularization term to the optimization problem in equation 2.4 is a straightforward approach. However, this leads to a complex nonconvex problem in sparse GCCA, involving an orthogonal constraint,
Sparse Generalized Canonical Correlation Analysis

an $L_1$ regularization term, and the Frobenius norm. Since the ADMM framework excels in decomposing problems into easily solvable subproblems, particularly when dealing with multiple diverse constraints, we appropriately reformulate the original problem to make it more compatible with the ADMM framework, enabling a more effective resolution of the sparse GCCA problem. The basic idea is to convert equation 2.4 into a linear system of equations by considering the optimality conditions and employing the singular value decomposition (SVD) technique. Specifically, according to the expansion of the objective function of equation 2.4,

$$\|G - W_j^T X_j\|_F^2 = \text{Trace}(GG^T - W_j^T X_j G^T - GX_j^T W_j + W_j^T X_j X_j^T W_j),$$

we can take the derivatives with respect to $W_j$ and obtain the following optimality conditions,

$$X_j X_j^T W_j = X_j G^T. \quad (3.1)$$

Typically, $X_j$ is a low-rank matrix, and there is redundancy in equation 3.1. Denote the reduced SVD of $X_j$ as

$$X_j = P_j \Sigma_j Q_j^T, \quad (3.2)$$

where $P_j \in \mathbb{R}^{n_j \times r_j}$ and $Q_j \in \mathbb{R}^{m \times r_j}$ are unitary matrices with $r_j \ll \min(n_j, m)$. Then we have $X_j X_j^T = P_j \Sigma_j^2 P_j^T (j = 1, \ldots, J)$ and convert the optimality condition, equation 3.1, into

$$P_j^T W_j = \Sigma_j^{-1} Q_j^T G^T, \quad \forall j = 1, \ldots, J.$$

With notation $A_j = P_j^T$, $B_j = -\Sigma_j^{-1} Q_j^T$, and $Z = G^T$, the above optimality condition containing the SVD of $X_j$ could be further written as the linear equation

$$A_j W_j + B_j Z = 0, \quad \forall j = 1, \ldots, J.$$

According to these optimality conditions, we now formulate the sparse solution of the canonical variates $W_j$ as follows,

$$\min_{\{W_j\}_{j=1}^J, Z} \|W\|_1 = \sum_{j=1}^J \|W_j\|_1 \quad \text{s.t.} \quad A_j W_j + B_j Z = 0, \quad Z^T Z = I_\ell, \quad (3.3)$$

where the $\ell_1$ norm of a matrix is defined as the summation of the $\ell_1$ norm of its columns: $\|W\|_1 = \sum_{j=1}^J \|W_j\|_1$. Unlike equation 2.4, equation 3.3
imposes sparsity constraints to interpret canonical variables. On the other hand, from the Bayesian inference viewpoint, equation 3.3 could be considered as a maximum a posteriori (MAP) estimate of $\|W\|_1$ with Laplace prior under special constraints. Besides, the linear constraints in equation 3.3 distinguish our novelty from other sparse GCCA formulations. Due to the constraint $Z^T Z = I_\ell$, equation 3.3 is a nonconvex problem on a Stiefel manifold. Before addressing the convergence analysis, we present the following first-order optimality conditions:

**Lemma 1.** Let $(W_1^*, W_2^*, \ldots, W_J^*, Z^*)$ be a local minimizer of equation 3.3, for each fixed $j = 1, \ldots, J$. Then there exist Lagrange multipliers $\Lambda_j^* \in \mathbb{R}^{r_j \times \ell}$, $\tilde{\Lambda}^* \in \mathbb{R}^{\ell \times \ell}$, $(j = 1, \ldots, J)$ such that

$$
A_j^T \Lambda_j^* \in \partial \|W_j^*\|_1, \quad \sum_{j=1}^J (B_j Z^*)^T \Lambda_j^* + \tilde{\Lambda}^* = 0,
$$

$$
A_j W_j^* + B_j Z^* = 0, \quad (Z^*)^T Z^* = I_\ell,
$$

where $\partial \|W\|_1$ stands for the subdifferential of $\| \cdot \|_1$ at $W^*$.

In the sequel, we call $(W_1^*, W_2^*, \ldots, W_J^*, Z^*, \Lambda_1^*, \ldots, \Lambda_J^*, \tilde{\Lambda}^*)$ satisfying equation 3.4 a Karush-Kuhn-Tucker (KKT) point of equation 3.3.

### 3.2 Sparse GCCA Algorithm via Distributed ADMM.

To solve the proposed sparse GCCA, equation 3.3, we develop an efficient algorithm, mainly in the framework of distributed ADMM. We first provide the augmented Lagrangian of equation 3.3 as follows,

$$
L_{\beta}(W_1, \ldots, W_j, Z, \Lambda_1, \ldots, \Lambda_j) = \sum_{j=1}^J \|W_j\|_1 + I_{O_\ell}(Z) - \sum_{j=1}^J \langle \Lambda_j, A_j W_j + B_j Z \rangle
$$

$$
+ \frac{\beta}{2} \sum_{j=1}^J \|A_j W_j + B_j Z\|_F^2,
$$

where $O_\ell = \{Z | Z^T Z = I_\ell\}$, $\Lambda_j \in \mathbb{R}^{r_j \times \ell}$ $(j = 1, \ldots, J)$ are the Lagrange multipliers corresponding to the constraints $A_j W_j + B_j Z$ $(j = 1, \ldots, J)$. $I_{O_\ell}(Z)$ is an indicator function defined as

$$
I_{O_\ell}(Z) = \begin{cases} 
0, & Z \in O_\ell, \\
+\infty, & \text{otherwise}.
\end{cases}
$$

However, it is difficult to obtain $W_j^{k+1}$ $(j = 1, \ldots, J)$ and $Z^{k+1}$ simultaneously for any $k$th iteration. Moreover, the existence of $A_j$ and $B_j$ with different
ranks makes it more challenging, and the orthogonality constraint $Z^T Z = I$ leads the nonconvex problem. Hence, we need to employ the iteration idea stated in distributed ADMM with slight modifications.

First, we decouple the update of primal variables $W_j^{k+1}$ ($j = 1, \ldots, J$) and $Z^{k+1}$; that is, $L_{\beta_k}$ is optimized with respect to variables $W_j$ and $Z$ one at a time, while fixing the others. Mathematically, it alternatively optimizes \{Z, W_j, \Lambda_j\}$_{j=1}^J$ in the $k$th iteration as follows,

$$
Z^{k+1} = \arg\min_{Z \in \Omega} \left\{ - \sum_{j=1}^{J} \langle \Lambda_j, A_j W_j^k + B_j Z \rangle + \frac{\beta_k}{2} \sum_{j=1}^{J} \|A_j W_j^k + B_j Z\|_F^2 \right\},
$$

(3.5)

$$
W_j^{k+1} = \arg\min_{W_j} \left\{ L_{\beta_k}(W_j, Z^{k+1}, \Lambda_1^k, \ldots, \Lambda_J^k) \right\},
$$

(3.6)

$$
\Lambda_j^{k+1} = \Lambda_j - \beta_k (A_j W_j^{k+1} + B_j Z^{k+1}),
$$

(3.7)

where $L_{\beta_k}(W_j, Z, \Lambda_1, \Lambda_2, \ldots, \Lambda_J)$ stands for $L_{\beta_k}(W_1, \ldots, W_J, Z, \Lambda_1, \ldots, \Lambda_J)$ with fixed $W_i$ ($i \neq j$). Among the subproblems above, $Z$-subproblem 3.5 could be simplified as

$$
Z^{k+1} = \arg\min_{Z \in \Omega} \left\{ - \sum_{j=1}^{J} \|A_j W_j^k + B_j Z\|_F^2 \right\}.
$$

(3.8)

Recalling the global consensus problem (Boyd et al., 2011),

$$
\min_{i=1}^{N} f_i(x_i) \quad \text{s.t. } x_i - z = 0, \quad i = 1, \ldots, N,
$$

(3.9)

which describes an optimization problem under the constraints that promote agreement or consistency denoted as $z$ among all variables $x_i$. Specifically, the constraints in equation 3.3 could be viewed as a weighted matrix version of those in equation 3.9. And the variable $Z$ plays the role of a central collector. When $J = 1$, equation 3.8 reduces to the famous unbalanced Procrustes problem (Lippert & Edelman, 2000; Eldén & Park, 1999; Hurley & Cattell, 1962), and has analytical solutions. However, when $J > 1$, it cannot be easily solved. Then we show how to solve the $Z$-subproblem 3.8 in detail.

Denote $C = \sum_{j=1}^{J} B_j^T B_j$, $D^k = - \sum_{j=1}^{J} B_j^T (\frac{\Lambda_j^k}{\beta_k} - A_j W_j^k)$; then equation 3.8 could be converted into the problem addressed in the following proposition, whose proof is in the appendix.
**Proposition 1.** For equation 3.8, we have

\[
\min_{Z \in O_{\ell}} \sum_{i=1}^{\ell} \lambda_i (C + Z(D^k)^T + D^kZ^T) = \min_{Z \in O_{\ell}} \left\{ \sum_{j=1}^{J} \left\| A_j W_j^k + B_j Z - \frac{\Lambda_j^k}{\beta_k} \right\|_F^2 \right\},
\]

for each fixed \( k \), where \( \lambda_i(E) \) \((i = 1, \ldots, \ell)\) stand for the eigenvalues of a matrix \( E \), where \( E = C + Z(D^k)^T + D^kZ^T \) in the Z-subproblem.

Proposition 1 indicates that equation 3.8 could be solved by minimizing the \( \ell \) smallest eigenvalues of the matrix \( C + Z(D^k)^T + D^kZ^T \). Motivated by Zhang et al. (2022, 2020), we could find the solution of equation 3.8 by the SCF iteration approach as described in algorithm 1.

Then we go on to solve the \( \{W_j\}_{j=1}^{J} \)-subproblem. Since \( \{W_j^k\} \) \((i \neq j)\), \( \ldots, W_j^k \) are fixed, we can further simplify the \( \{W_j\}_{j=1}^{J} \)-subproblem 3.6 as

\[
W_j^{k+1} = \arg \min_{W_j} \left\{ \|W_j\|_1 - (\Lambda_j^k, A_j W_j + B_j Z^{k+1}) + \frac{\beta_k}{2} \|A_j W_j + B_j Z^{k+1}\|_F^2 \right\}.
\]

\[
= \arg \min_{W_j} \left\{ \|W_j\|_1 + \frac{\beta_k}{2} \|A_j W_j + B_j Z^{k+1} - \frac{\Lambda_j^k}{\beta_k}\|_F^2 \right\}. \quad (3.10)
\]
Building on lemma 1, we can directly obtain the following theorem:

**Theorem 1.** Assume that $A_j$ is of full row rank for each fixed $j (j = 1, \ldots, J)$. Let $(W_j^k, \ldots, W_j^k, Z^k, \Lambda_j^k, \ldots, \Lambda_j^k)$ be generated by solving subproblems equations 3.8, 3.10, and 3.7 exactly and $\beta_{k+1} = \rho \beta_k (\rho > 1)$. Then, the sequence $(W_1^k, \ldots, W_J^k, Z^k, \Lambda_1^k, \ldots, \Lambda_J^k)$ is bounded, and

$$\lim_{k \to \infty} A_j W_j^k + B_j Z^k = 0 (j = 1, \ldots, J).$$

Moreover, any accumulation point $(W_1^*, \ldots, W_J^*, Z^*, \Lambda_1^*, \ldots, \Lambda_J^*, -\sum_{j=1}^J (B_j Z^*)^T \Lambda_j^*)$ of \{$(W_1^k, \ldots, W_J^k, Z^k, \Lambda_1^k, \ldots, \Lambda_J^k, -\sum_{j=1}^J (B_j Z^k)^T \Lambda_j^k)$\}_{k=1}^\infty satisfies KKT conditions 3.4 in lemma 1.

**Remark 1.** Theorem 1 provides the theoretical analysis of general ADMM process in equations 3.8, 3.10, and 3.7; however, optimization task 3.10 is an $\ell_1$-norm regularized least squares problem, which does not admit a closed-form solution. In the literature, various approaches (Tan et al., 2021; Chartrand & Wohlberg, 2013; Liu et al., 2019; Lin et al., 2011) have been employed to address this challenge. In the following, we prefer the linearization approximation method.

To avoid an exhaustive iterative process, we approximate it by linearizing the Frobenius norm term (for each fixed $j$),

$$\min_{W_j} \|W_j\|_1 + \beta_k \left\{ (A_j^T (A_j W_j^k + B_j Z^{k+1} - \Lambda_j^k / \beta_k), W_j - W_j^k) + \frac{||W_j - W_j^k||_2^2}{2\delta} \right\},$$

(3.11)

where $\delta > 0$ is a proximity parameter. For each fixed $j, j = 1, \ldots, J$, equation 3.11 can be solved by applying the famous proximal gradient descent (PGD) method,

$$W_j^{k+1} = S \left( W_j^k - \delta A_j^T \left( A_j W_j^k + B_j Z^{k+1} - \frac{\Lambda_j^k}{\beta_k} \right), \frac{\delta}{\beta_k} \right),$$

(3.12)

where $S(x, \mu)$ is the component-wise soft-thresholding shrinkage operator defined as

$$S(x, \mu) = \text{sgn}(x) \odot \max(|x| - \mu, 0),$$

with $\odot$ denoting the component-wise products of vectors or matrices. This update via PGD gives an approximate but closed-form solution for equation 3.10 so that every accumulation point is a critical point for general
Algorithm 2 Distributed alternative iteration based sparse generalized CCA algorithm (SGCCA).

**Input:**

Training data $X_j \in \mathbb{R}^{n_j \times m}$ ($j = 1, \cdots, J$), parameter $\delta > 0$, $\rho > 1$, $\beta_{\text{max}}$ and tolerance parameter $\varepsilon_1, \varepsilon_2$, initial factor $\tau$.

**Output:**

Sparse canonical variates $W = (W_1^T, \cdots, W_J^T)^T$.

1: Compute reduced SVD for each $X_j$ ($j = 1, \cdots, J$) via (7) and obtain $A_j, B_j$.

2: Let $W_j^0 = 0$, $A_j^0 = 0$, $\beta_0 = \tau \max(1/\|A_j^T B_j\|_\infty)$, ($j = 1, \cdots, J$).

3: while $\|A_j^{k+1} - A_j^k\|_F/\beta_k > \varepsilon_1$ and $\beta_k\|W_j^{k+1} - W_j^k\|_F/\max\{1, \|W_j^k\|_F\} > \varepsilon_2$ do

4: Compute $Z_j^{k+1}$ via Algorithm 1, and $W_j^{k+1}$ and $A_j^{k+1}$ by (17) and (12), respectively.

5: Update $\beta_{k+1}$ by $\beta_{k+1} = \min(\beta_{\text{max}}, \rho \beta_k)$.

6: end while

nonconvex, nonsmooth programs according to Li and Lin (2015). The similar theoretical analysis of such nonconvex linearized ADMM problems has been addressed in Lin et al. (2011), and we will provide the analysis of our method and show it is consistent with theorem 1 later.

Now, we come to the following augmented distributed alternative iteration-based sparse GCCA algorithm, as summarized in algorithm 2.

3.2.1 Selection of Parameters. In equation 3.11, $\delta$ controls the convergence in $\{W_j\}_{j=1}^J$ subproblems, and the smaller $\delta$ indicates the faster convergence. Besides, we employ a strategy of dynamically adjusting $\beta_k$, gradually increasing the step size by a factor of $\rho$ to accelerate the convergence of the algorithm. Simultaneously, we empirically set a maximum value $\beta_{\text{max}}$ to ensure the stability of the algorithm. Specifically, we determine one parameter by cross-validation while keeping the other parameters fixed in our experiments, as detailed in the subsequent section.

3.2.2 Computational Complexity. Algorithm 2 provides efficient updates with analytical expressions. To solve the $Z$-subproblem, we employ the SCF
iteration, which incurs a computational cost of $\mathcal{O}(m^3 + m\ell^2 + \ell^3)$ flops. The computational complexity of the $W_j$-subproblem and $\Lambda_j$-update is approximately $\mathcal{O}(r_1n_1\ell)$, primarily due to matrix multiplication. The most computationally intensive part is the SVD decomposition, but it can be replaced with an iterative solver, such as the locally optimal block preconditioned conjugate gradient method (Knyazev & Neymeyr, 2003), reducing the complexity from $\mathcal{O}(m^3)$ to $\mathcal{O}(m\ell^2)$.

3.2.3 Theoretical Convergence Analysis. The convergence of algorithm 2 cannot be directly inherited from classical ADMM, as the original problem, equation 3.3, is nonconvex and the update for equation 3.10 is approximate (specifically, equation 3.11 is a linearization approximation of equation 3.10). Based on the convergence conditions of matrix series, we remove the “F” symbol and get the following convergence analysis for algorithm 2:

**Theorem 2.** Let $\{(W_1^k, \ldots, W_j^k, Z^k, \Lambda_1^k, \ldots, \Lambda_j^k)\}_{k=1}^\infty$ be the sequence generated by algorithm 2. For each fixed $j (j = 1, \ldots, J)$, assume that (1) $\lim_{k \to \infty} \beta_k (W_j^{k+1} - W_j^k) = 0$, and (2) $\lim_{k \to \infty} (\Lambda_j^{k+1} - \Lambda_j^k)/\beta_k = 0$. Then any accumulation point $(W_1^*, \ldots, W_j^*, Z^*, \Lambda_1^*, \ldots, \Lambda_j^*)$ of $\{(W_1^k, \ldots, W_j^k, Z^k, \Lambda_1^k, \ldots, \Lambda_j^k, -\sum_{j=1}^J (B_jZ_j^*)^T \Lambda_j^*)\}_{k=1}^\infty$ satisfies KKT conditions 3.4 in lemma 1. If $\{(W_1^k, \ldots, W_j^k, Z^k, \Lambda_1^k, \ldots, \Lambda_j^k, -\sum_{j=1}^J (B_jZ_j^*)^T \Lambda_j^*)\}_{k=1}^\infty$ converges, it converges to a KKT point of problem 3.3.

**Remark 2.** The two assumptions in theorem 2 are mild. Let us elaborate those conditions. The first condition, $\lim_{k \to \infty} \beta_k (W_j^{k+1} - W_j^k) = 0$, ensures that the discrepancy between the two adjacent items, $W_j^{k+1}$ and $W_j^k$, should converge to 0, which will lead to the results $A_j^T \Lambda_j^* \in \partial \|W_j^*\|_1$ by noticing that $W_j^{k+1} (j = 1, \ldots, J)$ solves equation 3.11 (details are in the appendix). The second condition, $\lim_{k \to \infty} (\Lambda_j^{k+1} - \Lambda_j^k)/\beta_k = 0$, yields that $\lim_{k \to \infty} A_jW_j^k + B_jZ_j^k = 0$ by noticing equation 3.7, which means the generated sequence is feasible point of problem 3.3. Therefore, the above analysis indicates that under these two assumptions, the accumulation point $(W_1^*, \ldots, W_j^*, Z^*, \Lambda_1^*, \ldots, \Lambda_j^*)$ of the sequence $\{(W_1^k, \ldots, W_j^k, Z^k, \Lambda_1^k, \ldots, \Lambda_j^k, -\sum_{j=1}^J (B_jZ_j^*)^T \Lambda_j^*)\}_{k=1}^\infty$ generated by algorithm 2 is a KKT point of problem 3.3, which satisfies the first-order necessary conditions stated in equation 3.4. However, to the best of our knowledge, it is impossible to find the local optima of problem 3.3 directly. The detailed proofs of lemma 1 and theorems 1 and 2 are in the appendix.

3.3 The Simplified Version: Sparse GCCA with Fixed $G$. We presented the complete algorithm in the previous subsection, and when the variable $G$ is fixed in equation 3.3, we can get a much simpler version for the
proposed sparse GCCA. This enables us to use different versions in various cases. We can denote $W_j = (\alpha^1_j, \ldots, \alpha^\ell_j)$, $Z = (z_1, \ldots, z_\ell)$, where the columns of $Z = G^T$ are the eigenvectors of the matrix

$$M = \sum_{j=1}^J X_j^T (X_j X_j^T)^{-1} X_j.$$ 

For each fixed $i$, we need to solve $J$ problems:

$$P_j^T \alpha^i_j = \Sigma_j^{-1} Q_j^T z_i, \ (i = 1, \ldots, \ell).$$

To achieve sparsity of canonical variates, we establish the following model,

$$\min \|\alpha^i_j\|_1 \quad \text{s.t.} \quad P_j^T \alpha^i_j = \Sigma_j^{-1} Q_j^T z_i, \ j = 1, \ldots, J. \quad (3.13)$$

To reduce notational burden and allow a slight abuse of notation, we omit the index $i, j$ in the sequel unless specified. As a classical $\ell_1$ problem, there are lots of effective algorithms for equation 3.13, such as iterative shrinkage-thresholding algorithm (ISTA; Hale et al., 2007; Wright et al., 2008), and least angle regression (LARS), subgradient descent. Here we use fast ISTA (FISTA; Beck & Teboulle, 2009). Let $v_1 = \alpha_0, t_1 = 1$. The update for the sparse GCCA with fixed $G$ is

$$\begin{align*}
\alpha_s &= \arg\min_{\alpha \in \mathbb{R}^n} \left\{ \|\alpha\|_1 + \frac{L}{2} \|\alpha - (v_s - \frac{\ell}{L}(PP^T v_s - P\Sigma^{-1}Q^T z))\|^2_2 \right\}, \\
t_{s+1} &= \frac{1 + \sqrt{1 + 4t_s^2}}{2}, \\
v_{s+1} &= \alpha_s + \frac{t_{s-1}}{t_{s+1}}(\alpha_s - \alpha_{s-1}),
\end{align*}$$

where $L = \|PP^T\|$ differs for a distinct view and $t_s$ is the step-size of the $s$th iteration.

4 Experimental Results

In this section, we carry out numerical experiments on both synthetic data set and real-world data sets to evaluate our proposed sparse GCCA algorithm (SGCCA) by comparing it with other algorithms, including the traditional GCCA, weighted GCCA (Benton et al., 2016; WGCCA), deep GCCA (DGCCA; Benton et al., 2019), and the sparse solutions like structured GCCA under a penalty–dual decomposition framework with $\ell_1$ penalty (PDD-$\ell_1$; Kanatsoulis et al., 2019) and sparsity-constrained GCCA (Li et al., 2022; SCGCCA). All experiments are performed under Ubuntu.
4.1 Experiments on a Synthetic Data Set. First, we consider the proposed SGCCA on synthetic data to evaluate its convergence, sparsity, and construction accuracy.

4.1.1 Data Sets. We construct the synthetic three-view data set as three matrices, $X_1$, $X_2$, and $X_3$, in the following,

$$X_1 = v_1 u^T + \epsilon_1, \quad X_2 = v_2 u^T + \epsilon_2, \quad X_3 = v_3 u^T + \epsilon_3,$$

where the three sparse projected vectors are

$$v_1 = (1, \ldots, 1, -1, \ldots, -1, 0, \ldots, 0)^T,$$

$$v_2 = (0, \ldots, 0, 1, \ldots, 1, -1, \ldots, -1)^T,$$

$$v_3 = (1, \ldots, 1, 0, \ldots, 0, -1, \ldots, -1)^T,$$

and $\epsilon_1 \in \mathbb{R}^{10000 \times 100} \sim \mathcal{N}(0, 0.3^2)$, $\epsilon_2 \in \mathbb{R}^{15000 \times 100} \sim \mathcal{N}(0, 0.4^2)$, and $\epsilon_3 \in \mathbb{R}^{17000 \times 100} \sim \mathcal{N}(0, 0.5^2)$ are three random noise matrices. $u \in \mathbb{R}^{100}$ is a random vector with all entries drawn from the normal distribution,

$$u(i) \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, 100.$$

Then the ground-truth sparsity of each view, denoted as the percentage of zero entries of the projected vector, is 50%, 66.67%, 70.59%, respectively.

4.1.2 Experimental Settings. We randomly select half of the data for training and use the rest for testing. GCCA is used as the baseline algorithm. We also consider weighted GCCA (WGCCA) with gaussian initialization weights and DGCCA with 100 epochs and batch size tuned by cross-validation. For the PDD-$\ell_1$ method, we select $in_{iter} = 5$ (iteration number for the update of $W_i$, $i = 1, \ldots, J$ and $G$), and $out_{iter} = 50$ (iteration number for the Lagrange multipliers). For the SCGCCA method using Newton hard-thresholding pursuit (NHTP), we set the maximum iterations to 2000 and the sparsity level to 1. For the parameters of the proposed SGCCA algorithm, we set $\beta_{max} = 10^4$, tolerance $\epsilon_1 = \epsilon_2 = 10^{-5}$, and $\delta, \rho$ tuned by cross-validation around one.
Table 1: Performance of the Algorithms on Synthetic Data Set: Averaged Sparsity of Canonical Variates, Correlations for Training (corr\textsubscript{tr}) and Testing Data (corr\textsubscript{ts}), Construction Error for Training (error\textsubscript{tr}) and Testing Data (error\textsubscript{ts}).

| Criterion     | GCCA | WGCCA | DGCCA | PDD-\ell_{1} | SCGCCA | SGCCA |
|---------------|------|-------|-------|--------------|--------|-------|
| sparsity      | 4.78%| 61.58%| 68.16%| 63.18%       | 99.98% | 99.58%|
| corr\textsubscript{tr} | 3.0000 | 3.0000 | 0.8727 | 2.9943 | 1.0141 | 3.0000 |
| error\textsubscript{tr}  | 0.0000 | 0.0000 | 0.0846 | – | 0.8544 | 0.0000 |
| corr\textsubscript{ts}  | 2.6528 | 2.9995 | 0.9233 | 0.3828 | 0.9834 | 2.4141 |
| error\textsubscript{ts}  | 0.5112 | 0.7325 | 4.4295 | – | 1.0209 | 0.4528 |

Note: The proposed SGCCA achieves the best trade-off between sparsity and construction error.

4.1.3 Metrics. We employ the concept of averaged sparsity of canonical variates to assess the ability to identify the sparse structure within the synthetic data. This metric represents the mean of the percentages of zero entries among the projected vectors. Specifically, we consider the value of an entry to be treated as zero if it is less than 1E-5 in the experiment. We define the construction error of the common latent representation \( G \) in equation 2.4, measured by the Frobenius norm \( \frac{1}{J} \sum_{j=1}^{J} \| W_j^T X_j - G \|_F^2 \) with \( J = 3 \) views, for training data (referred to as error\textsubscript{tr}) and testing data (referred to as error\textsubscript{ts}). Furthermore, we calculate the correlation as \( \frac{1}{2} \sum_{i \neq j} \text{Trace}(W_i^T X_i X_j^T W_j) \) for training (denoted as corr\textsubscript{tr}) and testing (denoted as corr\textsubscript{ts}). And the correlation is up to the number of the data views \( J \). To sum up, the aim of the experiments is to discover a sparse decomposition structure that allows us to achieve a lower construction error, as well as higher sparsity, ultimately leading to improved signal recovery performance.

4.1.4 Performance. We compare our SGCCA to the baseline methods and report the results in Table 1, where “−” means “not applicable” since the objective function of PDD-\ell_{1} is different from other methods. We observe, first, that all the considered algorithms exhibit experimentally efficient training and yield promising training outcomes. Specifically, although the baseline methods are implemented in different programming languages, our approach runs efficiently in the same programming language (Python), for example, 10 times faster than WGCCA and nearly 100 times faster than DGCCA. Second, traditional GCCA and WGCCA fall short in identifying the sparse structure, as their sparsity is lower than the ground-truth sparsity of 62.42%. Third, for the sparse solutions, our proposed SGCCA can achieve a lower construction error of the latent representation and demonstrates increased sparsity compared to most compared methods.
Sparse Generalized Canonical Correlation Analysis

Table 2: Data Structures: Data Dimension \((n)\), Number of Data \((m)\), Number of Classes \((K)\), Number of Columns in \(W_1\) and \(W_2\) \((\ell)\).

| Type   | Data      | \(n\) | \(m\) | \(K\) | \(\ell\) |
|--------|-----------|-------|-------|-------|---------|
| Gene Data | Leukemia  | 3571  | 72    | 2     | 1       |
|         | Prostate  | 6033  | 102   | 2     | 1       |
|         | Colon     | 2000  | 62    | 2     | 1       |
|         | SRBCT     | 2308  | 63    | 4     | 1       |

4.2 Experiments on Real Data Sets. Through experiments on synthetic data, we confirm that the proposed sparse GCCA is capable of finding sparse structures. Here, we present the experimental results on actual data sets.

4.2.1 Classification on Gene Expression Data. We consider four data sets from the gene expression database.\(^1\) The details are explained below and the statistics are in Table 2.

- Leukemia: Gene expression values for 72 samples with 3571 genes (47 samples from patients with acute lymphoblastic leukemia, and 25 from patients with acute myeloblastic leukemia)
- Prostate: Gene expression values measured by the Affymetrix technology for 102 samples (52 prostate tumors and 50 nontumor prostate samples)
- Colon: Gene expression values for 40 tumor and 22 normal colon tissues for 2000 selected human genes with highest minimal intensity
- SRBCT: Gene expression values from cDNA microarrays containing 2308 genes for 63 samples in four classes

In this case, we apply the proposed method to the application of classification problem. Specifically, there are two views, the data matrix \(X_1\) and the label matrix \(X_2\). The preprocessing procedure for gene data is described in Dettling (2004). In the experiment, we choose \(\ell\) as the rank of the matrix \(X_1X_2^T\).

We select 4/5 of the data for training and use the rest for test. The choices of the other parameters are the same as described in the previous section.

In addition to the previously employed evaluation metrics like the correlations for training and testing, reconstruction error for training, and averaged sparsity of canonical variates, we have introduced this classification accuracy,

\[
\text{Accuracy: } \frac{\sum_{i=1}^{m} \delta(o_i, p_i)}{m},
\]

\(^1\)http://stat.ethz.ch/~dettling/bagboost.html.
Table 3: Performance of the Algorithms on Gene Expression Data (Leukemia and Prostate): Sparsity of Canonical Variates of Data Matrix, Correlation for Training (corr\textsubscript{tr}) and Testing Data (corr\textsubscript{ts}), Construction Error for Training Data (error\textsubscript{tr}), Classification Accuracy for Training (acc\textsubscript{tr}) and Testing Data (acc\textsubscript{ts}).

| Data     | Metrics     | GCCA | WGCCA | DGCCA | PDD-\ell_1 | SGCCA |
|----------|-------------|------|-------|-------|------------|-------|
| Leukemia | sparsity    | 4.62%| 5.12% | 41.10%| 99.27%     | 97.86%|
|          | corr\textsubscript{tr} | 1.0000 | 0.9068 | 0.1043 | 0.9800     | 0.9996|
|          | error\textsubscript{tr} | 0.0000 | 0.0386 | 0.0188 | –          | 0.0000|
|          | acc\textsubscript{tr} | 100% | 98.25% | 64.91% | 64.91%     | 100%  |
|          | corr\textsubscript{ts} | 0.9143 | 0.8401 | 0.5015 | 0.2460     | 0.9376|
|          | acc\textsubscript{ts} | 100% | 93.33% | 73.33% | 66.67%     | 100%  |
| Prostate | sparsity    | 5.60%| 7.14% | 10.20%| 99.27%     | 98.72%|
|          | corr\textsubscript{tr} | 1.0000 | 0.8836 | 0.0614 | 0.9909     | 0.9962|
|          | error\textsubscript{tr} | 0.0000 | 0.0349 | 0.0881 | –          | 0.0000|
|          | acc\textsubscript{tr} | 100% | 97.53% | 51.85% | 56.79%     | 100%  |
|          | corr\textsubscript{ts} | 0.7741 | 0.6097 | 0.2838 | 0.1352     | 0.8148|
|          | acc\textsubscript{ts} | 90.48%| 85.71% | 57.14% | 61.91%     | 90.48%|

Notes: The proposed SGCCA achieves the best trade-off between sparsity and accuracy. The best results in sparsity and classification accuracy are in bold.

where \( \delta(y_1, y_2) \) is the indicator function that equals 1 if \( y_1 = y_2 \) and 0 otherwise. For a given sample point \( x_i \), \( ol_i \) and \( pl_i \) are the obtained and provided labels, respectively. Among those metrics, sparsity and classification accuracy are the most important.

The detailed comparison results are listed in Tables 3 and 4, and the best results in sparsity and classification accuracy are in bold. We observe that smaller errors related to the common latent representation during training lead to better classification accuracy. Besides, in the case of multiclassification tasks (e.g., in SRBCT), the performance of all methods shows a decline. However, the proposed SGCCA outperforms baselines remarkably in classification accuracy and significantly increases the sparsity of the canonical variable for the data matrix while maintaining accuracy. To visually showcase the training results, we perform t-SNE (Van der Maaten & Hinton, 2008) dimensionality reduction using the classification task of SRBCT as an illustration to demonstrate the outcomes in Figure 1.

4.2.2 Cross-Language Document Retrieval. In this section, we conduct experiments on Europarl parallel corpus (Koehn, 2005—Europarl for short), a collection of documents extracted from the proceedings of the European Parliament. It includes translated documents in 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic
Table 4: Performance of the Algorithms on Gene Expression Data (Colon and SRBCT): Sparsity of Canonical Variates of Data Matrix, Correlation for Training (corr)$_{tr}$ and Testing Data (corr)$_{ts}$, Construction Error for Training Data (error)$_{tr}$, Classification Accuracy for Training (acc)$_{tr}$ and Testing Data (acc)$_{ts}$.

| Data   | Metrics | GCCA | WGCCA | DGCCA | PDD-ℓ₁ | SGCCA |
|--------|---------|------|-------|-------|--------|-------|
| Colon  | sparsity| 1.9% | 1.7%  | 0.7%  | 9.26%  | 96.64%|
|        | corr$_{tr}$| 1.0000 | 0.8844 | 0.4397 | 0.8611 | 0.9990 |
|        | error$_{tr}$| 0.0000 | 0.0523 | 0.0437 | –      | 0.0000 |
|        | acc$_{tr}$| 100% | 97.95% | 32.65% | 93.87% | 100%  |
|        | corr$_{ts}$| 0.7053 | 0.6020 | 0.1146 | 0.1284 | 0.7908 |
|        | acc$_{ts}$| 92.30% | 76.92% | 61.53% | 69.23% | 92.30% |
| SRBCT  | sparsity| 2.42% | 3.16% | 53.76% | 99.13% | 97.66%|
|        | corr$_{tr}$| 1.0000 | 0.8885 | 0.0134 | 0.9752 | 0.9996 |
|        | error$_{tr}$| 0.0000 | 0.0420 | 0.0209 | –      | 0.0000 |
|        | acc$_{tr}$| 100% | 56.00% | 34.00% | 96.00% | 100%  |
|        | corr$_{ts}$| 0.9460 | 0.9326 | 0.1986 | 0.9359 | 0.9559 |
|        | acc$_{ts}$| 76.92% | 69.23% | 23.07% | 61.54% | 76.92% |

Notes: The proposed SGCCA achieves the best trade-off between sparsity and accuracy. The best results in sparsity and classification accuracy are in bold.

Figure 1: The visualization of SRBCT features by t-SNE.

(Latvian, Lithuanian), and Greek. The main aim is to learn the latent representations of the sentences, which reveals the correlations of the same sentences in different languages (views).

For the Europarl data set, we select three different types of language data (English, French, and Spanish) and obtain a bag-of-words representation using term frequency inverse document frequency (TFIDF) approach,
which is widely recognized as an efficient way in the document retrieval task. After removing numbers, stop words (English, French, and Spanish, respectively), and rare words (appearing less than twice), we achieve distinct sizes of matrices for different types of language data, which is demonstrated in Table 5. We choose half of the data as training and use the rest for testing. The parameters for the other methods are the same as those described in the previous section.

We compare the proposed SGCCA with baseline methods for $\ell = 1$ by considering several metrics: the correlations for training and testing, reconstruction error for training, sparsity of each canonical variate and retrieval precision (Chu et al., 2013) for both training and testing. The retrieval precision is computed by projecting the query document into a learned lower-dimensional space and then calculating the distance between the projected query document and all the testing documents. The distances are sorted in decreasing order, and the index location of the most relevant document is identified. Then the retrieval precision is determined by considering the percentage of top relevant documents among all documents.

The detailed performance compared with other algorithms is reported in Table 6. Specifically, we evaluate language document retrieval precision for each pair of languages in the sequence of English-French pair, English-Spanish pair, and French-Spanish pair. We find that the proposed SGCCA is considerably more prominent than other algorithms in retrieval precision for almost every view and also performs well in sparsity, showing that SGCCA could preferably find the sparse structure in the cross-language document retrieval task.

### Table 5: Data Structures for Three Languages Data Dimension ($n$), Number of Data ($m$).

| Type            | Data  | $n$  | $m$  |
|-----------------|-------|------|------|
| Document Data   | English | 12,421 | 1000 |
|                 | French  | 15,650 | 1000 |
|                 | Spanish | 17,203 | 1000 |

5 Conclusion

In this article, we present a novel GCCA framework by leveraging the MAX-VAR formulation and employing the SVD technique. Our approach offers a fresh perspective on GCCA as a linear system of equations and incorporates sparsity into this framework. Theoretical consistency is thoroughly investigated under mild conditions. We design a distributed alternating iteration-based sparse GCCA algorithm (SGCCA). Extensive experimental results on synthetic data sets, gene data, and the Europarl data set all demonstrated the effectiveness of the proposed algorithm, which is promising for CCA applications that involve more than two views and have a sparsity
Table 6: Performance of the Algorithms on Europarl Data Set When $\ell = 1$.

| Metrics       | GCCA | WGCCA | DGCCA | PDD-$\ell_1$ | SCGCCA | SGCCA |
|---------------|------|-------|-------|--------------|--------|-------|
| sparsity of each | 6.09% | 16.58% | 0.86% | 99.98% | **99.99%** | 99.93% |
| sparsity      | 6.76% | 17.14% | 2.60% | **99.99%** | 99.99% | 99.92% |
| corr$_{tr}$   | 5.85% | 14.43% | 5.90% | 99.98% | **99.99%** | 99.90% |
| error$_{tr}$  | 6.23% | 16.05% | 3.11% | 99.98% | **99.99%** | 99.92% |
| prec$_{tr}$ of each | 100% | 99.99% | 53.41% | 72.68% | 64.32% | 75.91% |
| prec$_{ts}$ of each | 100% | 99.98% | 48.71% | 70.88% | 58.92% | 81.91% |
| prec$_{tr}$   | 100% | 99.99% | 53.13% | 69.62% | 65.16% | 73.08% |
| prec$_{ts}$   | 100% | 100%  | 51.75% | 71.06% | 59.47% | 76.97% |
| corr$_{ts}$   | 74.29% | 56.74% | 50.68% | 73.23% | 65.31% | 75.07% |
| prec$_{ts}$   | 74.77% | 55.25% | 49.23% | 75.03% | 58.99% | **84.26%** |
| prec$_{ts}$   | 73.98% | 50.96% | 54.43% | 67.23% | 55.44% | 72.41% |
| prec$_{ts}$   | 74.35% | 54.32% | 51.45% | 71.83% | 59.91% | **77.25%** |

Notes: Sparsity of canonical variate of each view (sparsity of each) and averaged sparsity of canonical variates, correlations for training and testing data (corr$_{tr}$ and corr$_{ts}$), construction error (error$_{tr}$), and retrieval precision for training and testing data between each pair of languages (prec$_{tr}$ of each and prec$_{ts}$ of each) and their averaged results (prec$_{tr}$ and prec$_{ts}$). Our proposed SGCCA achieves the best trade-off between sparsity and retrieval precision.

In the future, investigating the integration of the proposed SGCCA with deep learning techniques could open avenues for enhanced feature extraction and representation learning.

Appendix: Details of the Proofs of Proposition 1, Lemma 1, and Theorems 1 and 2

We begin with the proof of proposition 1.

**Proof of Proposition 1.** Employing the properties of trace operators, we can see that

$$
\sum_{j=1}^{J} \left\| B_j Z - \left( \frac{\Lambda^k_j}{\beta_k} - A_j W_j^k \right) \right\|_F^2 = \text{Trace} \left[ \sum_{j=1}^{J} \left[ Z^T B_j^T B_j Z - Z^T B_j^T \left( \frac{\Lambda^k_j}{\beta_k} - A_j W_j^k \right) \right] 
- \left( \frac{\Lambda^k_j}{\beta_k} - A_j W_j^k \right)^T B_j Z + \left( \frac{\Lambda^k_j}{\beta_k} - A_j W_j^k \right)^T \left( \frac{\Lambda^k_j}{\beta_k} - A_j W_j^k \right) \right].
$$
Recall $C = \sum_{j=1}^{l} B_j^T B_j$, $D^k = -\sum_{j=1}^{l} B_j^T (\frac{A^k_j}{\beta_k} - A_j W^k_j)$, and the term $\left(\frac{A^k_j}{\beta_k} - A_j W^k_j\right)^T (\frac{A^k_j}{\beta_k} - A_j W^k_j)$ is independent of $Z$. Therefore,

$$\min_{Z \in O} J \sum_{j=1}^{l} \left\| B_j Z - \left(\frac{A^k_j}{\beta_k} - A_j W^k_j\right)\right\|_F^2 = \min_{Z \in O} \text{Trace}(Z^T C Z + Z^T D^k + (D^k)^T Z).$$

Directly applying the conclusion of Zhanget al. (2022; theorem 3.3), we can achieve the conclusion of the proposition:

$$\min_{Z \in O} \lambda_i (C + Z(D^k)^T + D^k Z^T) = \min_{Z \in O} \left\{ \sum_{j=1}^{l} A_j W^k_j + B_j Z - \frac{\Lambda^k_j}{\beta_k}\right\}_F^2.$$

Now we directly cite two lemmas that play significant roles in proving theoretical results. We first give some notations. The effective domain of a convex function $f$ on $S$ is denoted by $\text{dom } f$ and defined by

$$\text{dom } f = \{x | \exists \mu, (x, \mu) \in \text{epi } f \} = \{x | f(x) < +\infty\},$$

where epi$f$ stands for the epigraph of $f$. ri$C$ means the relative interior of a convex set $C$:

$$\text{ri } C = \{x \in \text{aff } C | \exists \epsilon > 0, (x + \epsilon B) \cap (\text{aff } C) \subset C\},$$

where aff$C$ denotes the affine hull of $C$, $B = \{x | \|x\| \leq 1\}$ stands for the Euclidean ball in $\mathbb{R}^n$. $f'(x; y)$ denotes the directional derivative of $f$ at $x$ with respect to a vector $y$.

**Lemma 2** (Rockafellar, 1970, theorem 23.4). Let $f$ be a proper convex function. For $x \notin \text{dom } f$, $\partial f(x)$ is empty. For $x \in \text{ri}(\text{dom } f)$, $\partial f(x)$ is nonempty, $f'(x; y)$ is closed and proper as a function of $y$, and

$$f'(x; y) = \sup \{ \langle x^*, y \rangle | x^* \in \partial f(x) \} = \delta^*(y | \partial f(x)).$$

Finally, $\partial f(x)$ is a nonempty bounded set if and only if $x \in \text{int}(\text{dom } f)$ (int stands for the interior of a set), in which case $f'(x; y)$ is finite for every $y$.

**Lemma 3** (Rockafellar, 1970, theorem 24.4). Let $f$ be a closed proper convex function on $\mathbb{R}^n$. If $x_1, x_2, \ldots$ and $x^*_1, x^*_2, \ldots$ are two sequences such that $x^*_i \in \partial f(x_i)$, where $x_i$ converges to $x$ and $x^*_i$ converges to $x^*$, then $x^* \in \partial f(x)$. In other words, the graph of $\partial f$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$. 
Following the technique route for ADMM convergence discussion (Zhang et al., 2015), we can prove lemma 1:

**Proof of Lemma 1.** Since \((W_1^*, \ldots, W_J^*, Z^*)\) is a local minimizer of 3.3, it should satisfy the optimality conditions,

\[
A_j W_j^* + B_j Z^* = 0, \quad (Z^*)^T Z^* = I_{\ell}.
\]

By proving the existence of \(\Lambda_j^* (j = 1, \ldots, J)\) such that \(A_j^T \Lambda_j^* \in \partial \|W_j^*\|_1\) for each fixed \(j\) and setting \(\Lambda_j^* = - \sum_{j=1}^J (B_j Z_j^*)^T \Lambda_j^*, \) we have the conclusion that \((W_1^*, \ldots, W_J^*, Z^*, \Lambda_1^*, \ldots, \Lambda_J^*, \tilde{\Lambda}^*)\) satisfies the optimality conditions in equation 3.3. For each fixed \(j\), to prove the existence of such \(\Lambda_j^*\), we only need to prove \(S_j \cap \partial J(W_j^*) \neq \emptyset\), where the cone \(S_j = \{A_j^T \Lambda_j, \Lambda_j \in \mathbb{R}^{r_j \times \ell}\}\), and \(J(W_j) = \|W_j\|_1\), \(j = 1, \ldots, J\).

Obviously, both \(S_j\) and \(\partial J(W_j^*)\) are nonempty, closed convex sets. Suppose that \(S_j \cap \partial J(W_j^*) = \emptyset\). According to the separation theorem of convex sets (Rockafellar, 1970), there exist nonzero \(Y_j \in \mathbb{R}^{n_j \times \ell}\) for each fixed \(j\) \((j = 1, \ldots, J)\) such that

\[
(Y_j, D_j) \leq (Y_j, A_j^T \Lambda_j) - 1, \quad \forall D_j \in \partial J(W_j^*), \quad \Lambda_j \in \mathbb{R}^{r_j \times \ell}.
\]

Thus, \(A_j Y_j = 0, j = 1, \ldots, J\); otherwise, let \(\Lambda_j = \alpha A_j Y_j\) and \(\alpha \to -\infty\) so that \((Y_j, D_j) \leq -\infty\), which is obviously false. Hence, we can see that

\[
(Y_j, D_j) \leq -1, \quad \forall D_j \in \partial J(W_j^*), \quad j = 1, \ldots, J.
\]

Let \(W_j(\alpha) = W_j^* + \alpha Y_j\) \((j = 1, \ldots, J)\); then we obtain a feasible point of equation 3.3 as \((W_1(\alpha), \ldots, W_J(\alpha), Z^*)\) and \(W_j(\alpha) \to W_j^*\) as \(\alpha \to 0^+\). Since \((W_1^*, \ldots, W_J^*, Z^*)\) is a local minimizer of equation 3.3, we have

\[
J(W_j(\alpha)) - J(W_j^*) \geq 0, \quad j = 1, \ldots, J
\]

for a sufficiently small \(\alpha\), and the directional derivative of \(J\) at \(W_j^*\) is defined as

\[
J'(W_j^*, Y_j) = \lim_{\alpha \to 0^+} \frac{J(W_j(\alpha)) - J(W_j^*)}{\alpha} \geq 0.
\]

However, lemma 2 tells us that

\[
J'(W_j^*, Y_j) = \max_{D_j \in \partial J(W_j^*)} (Y_j, D_j) \leq -1, \quad j = 1, \ldots, J,
\]
which is a contradiction. Hence, for each fixed $j$,

$$S_j \cap \partial \mathcal{J}(W^*_j) \neq \emptyset,$$

and then we prove the existence of such $\Lambda^*_j$. \hfill \Box

Now we are in position to give the proof of theorem 1.

**Proof of Theorem 1.** We only need to prove the boundedness of $\{W^k_j\}_{k=1}^\infty$ and $\{\Lambda^k_j\}$ for each $j = 1, \ldots, J$, since $\{Z^k\}$ is an orthogonal matrix that is obviously bounded. Because $W^k_j$ solves equation 3.10, it satisfies the optimality condition

$$0 \in \beta_k A_j^T (A_j W^k_j + B_j Z^{k+1} - \Lambda_j^k/\beta_k) + \partial \|W_j^{k+1}\|_1, \quad \forall k \geq 0,$$

or, equivalently

$$A_j^T \Lambda^k_j \in \partial \|W_j^{k+1}\|_1,$$

by noticing $\beta_k (A_j W^k_j + B_j Z^{k+1}) = \Lambda_j^k - \Lambda_j^{k+1}$. When $A_j$ is of full row rank,

$$\Lambda_j^{k+1} \in (A_j A_j^T)^{-1} \cdot \partial \|W_j^{k+1}\|_1, \quad \forall k \geq 1.$$

Obviously, for a fixed $j$, $\partial \|W_j^{k+1}\|_1$ is a compact set (Rockafellar, 1970), from which it follows that the sequence $\{\Lambda_j^{k}\}_{k=1}^\infty$ ($j = 1, \ldots, J$) is bounded (fixed $j$). From the iteration procedure of equations 3.8, 3.10, and 3.7, we can see that

$$L_{\beta_k}(W^{k+1}_1, W^{k+1}_2, W^{k+1}_3, \ldots, W^{k+1}_j, Z^{k+1}_j, \Lambda_1^k, \ldots, \Lambda_j^k)$$

$$\leq L_{\beta_k}(W^k_1, W^k_2, W^k_3, \ldots, W^k_j, Z^{k+1}_j, \Lambda_1^k, \ldots, \Lambda_j^k)$$

$$\leq L_{\beta_k}(W^k_1, W^k_2, W^k_3, \ldots, W^k_j, Z^{k+1}_j, \Lambda_1^k, \ldots, \Lambda_j^k)$$

$$\vdots$$

$$\leq L_{\beta_k}(W^k_j, Z^{k+1}_j, \Lambda_1^k, \ldots, \Lambda_j^k),$$

and

$$L_{\beta_k}(W^k_1, \ldots, W^k_j, Z^{k+1}_j, \Lambda_1^k, \ldots, \Lambda_j^k)$$

$$\leq L_{\beta_k}(W^k_1, \ldots, W^k_j, Z^k_j, \Lambda_1^k, \ldots, \Lambda_j^k).$$
\[ L_{\beta_{k-1}}(W_1^k, \ldots, W_J^k, Z^k, \Lambda_1^{k-1}, \ldots, \Lambda_J^{k-1}) + \frac{1}{2} \sum_{j=1}^J (\Lambda_j^k - \Lambda_j^{k-1})^2 \]

\[ + \frac{\beta_k - \beta_{k-1}}{2} \sum_{j=1}^J \|A_j W_j^k + B_j Z^k\|_F^2 \]

\[ = L_{\beta_{k-1}}(W_1^k, \ldots, W_J^k, Z^k, \Lambda_1^{k-1}, \ldots, \Lambda_J^{k-1}) + \frac{\beta_{k-1} + \beta_k}{2\beta_{k-1}} \sum_{j=1}^J \|\Lambda_j^k - \Lambda_j^{k-1}\|_F^2, \]

where the last equality is achieved with the relation \( \beta_{k-1}(A_j W_j^k + B_j Z^k) = \Lambda_j^{k-1} - \Lambda_j^k \). By repeating the above process for \( k \), we obtain

\[ L_{\beta_k}(W_1^{k+1}, \ldots, W_J^{k+1}, Z^{k+1}, \Lambda^k) \]

\[ \leq L_{\beta_0}(W_1^1, \ldots, W_J^1, Z^1, \Lambda^0) + \sum_{i=1}^k \frac{\beta_{i-1} + \beta_i}{2\beta_{i-1}^2} \sum_{j=1}^J \|\Lambda_j^i - \Lambda_j^{i-1}\|_F^2 \]

\[ \leq L_{\beta_0}(W_1^0, \ldots, W_J^0, Z^1, \Lambda^0) + \sum_{i=1}^k \frac{\beta_{i-1} + \beta_i}{2\beta_{i-1}^2} \sum_{j=1}^J \|\Lambda_j^i - \Lambda_j^{i-1}\|_F^2, \]

where

\[ L_{\beta_0}(W_1^0, \ldots, W_J^0, Z^1, \Lambda^0) \]

\[ = \sum_{j=1}^J \|W_j^0\|_1 - \sum_{j=1}^J \langle \Lambda_j^0, A_j W_j^0 + B_j Z^1 \rangle + \frac{\beta_0}{2} \sum_{j=1}^J \|A_j W_j^0 + B_j Z^1\|_F^2 \]

\[ = \tau \ast \max(1/\|A_j^T B_j\|_\infty) \sum_{j=1}^J \|B_j Z^1\|_F^2 \]

is bounded with \( W_j^0 = 0, \Lambda_j^0 = 0, \beta_0 = \tau \ast \max(1/\|A_j^T B_j\|_\infty) \). \( Z^1 \) an orthogonal matrix, and \( B_j = -\Sigma_j^{-1} Q_j \) (\( Q_j \) are unitary matrices), \( j = 1, \ldots, J \). Notice that \( \{\Lambda_j^k\}_{k=1}^\infty \) is bounded in theorem 1, and we can take \( M_\Lambda \) as the uniform bound of \( \|\Lambda_j^k\|, k = 1, \ldots, \infty, j = 1, \ldots, J \); then

\[ \|\Lambda_j^k - \Lambda_j^{k-1}\| \leq 2M_\Lambda, j = 1, \ldots, J. \]
The estimation of \( \sum_{i=1}^{k} \frac{\beta_{i-1}+\beta_i}{2\beta^2_{i-1}} \sum_{j=1}^{l} \| \lambda_j^i - \Lambda_{j-1}^i \|^2 \) reduces to

\[
\sum_{i=1}^{k} \frac{\beta_{i-1}+\beta_i}{2\beta^2_{i-1}} \sum_{j=1}^{l} \| \lambda_j^i - \Lambda_{j-1}^i \|^2 \leq 4JM^2_{\lambda} \sum_{i=1}^{k} \frac{\beta_{i-1}+\beta_i}{2\beta^2_{i-1}}.
\]

Simple computation concerned with the summation of geometric sequence leads to

\[
\sum_{i}^{\infty} \frac{\beta_{i-1}+\beta_i}{2\beta^2_{i-1}} = \frac{\rho(1+\rho)}{2(\rho-1)\beta_0},
\]

which is bounded obviously by considering the relations \( \beta_{k+1} = \rho \beta_k \) and \( \rho > 1 \). Thus, \( L_{\beta_k}(W_f^{k+1}, \ldots, W_f^{k+1}, \Lambda_f^{k+1}, \ldots, \Lambda_f^{k+1}) \) is upper-bounded. Moreover,

\[
\sum_{j=1}^{l} \| W_j^k \|_1 = L_{\beta_{k-1}}(W_f^1, \ldots, W_f^k, \Lambda_1^{k-1}, \ldots, \Lambda_f^{k-1})
\]

\[
+ \frac{1}{2\beta_{k-1}} \sum_{j=1}^{l} (\| \lambda_j^{k-1} \|^2 - \| \lambda_j^k \|^2),
\]

is upper-bounded, which could be achieved by noticing the expression for \( L_{\beta_{k-1}}(W_f^1, \ldots, W_f^k, \Lambda_1^{k-1}, \ldots, \Lambda_f^{k-1}) \) and \( \Lambda_j^{k-1} - \Lambda_j^k = \beta_{k-1}(A_j W_j^k + B_j Z_j^k) \). Thus, the sequence \( \{W_j^k\}_{k=1}^{\infty} \) is bounded. Also notice that

\[
A_j W_j^k + B_j Z_j^k = \frac{\Lambda_j^{k-1} - \Lambda_j^k}{\beta_{k-1}} \to 0 \quad \text{as} \quad k \to \infty.
\]

Hence for any accumulation point \( (W_f^1, \ldots, W_f^k, Z^*, \Lambda_f^1, \ldots, \Lambda_f^k, - \sum_{j=1}^{l} (B_j Z_j^*)^\top \Lambda_f^*) \) of \( \{(W_f^1, \ldots, W_f^k, Z^*, \Lambda_f^1, \ldots, \Lambda_f^k, - \sum_{j=1}^{l} (B_j Z_j^*)^\top \Lambda_f^*)_{k=1}^{\infty}\} \), without any loss of generality, we can assume that \( (W_f^1, \ldots, W_f^k, Z^*, \Lambda_f^1, \ldots, \Lambda_f^k, - \sum_{j=1}^{l} (B_j Z_j^*)^\top \Lambda_f^*) \) is the limit of \( \{(W_f^1, \ldots, W_f^k, Z_k, \Lambda_f^1, \ldots, \Lambda_f^k, - \sum_{j=1}^{l} (B_j Z_j^*)^\top \Lambda_f^*)_{k=1}^{\infty}\} \), where \( k_i \) is the subsequence of \( k, i = 1, \ldots, \infty \). Letting \( i \to \infty \) and applying lemma 3, we have \( A^\top_j \Lambda_f^* \in \partial \| W_f^j \|_1, \quad j = 1, \ldots, l \). Thus, \( (W_f^1, \ldots, W_f^k, Z^*, \Lambda_f^1, \ldots, \Lambda_f^k, - \sum_{j=1}^{l} (B_j Z^*)^\top \Lambda_f^*) \) satisfies conditions 3.4 in lemma 1.

**Proof of Theorem 2.** Since \( \lim_{k \to \infty} \frac{\Lambda_{j+1}^{k+1} - \Lambda_j^k}{\rho_k} = 0 \) and \( \frac{\Lambda_{j}^{k+1} - \Lambda_{j+1}^{k+1}}{\rho_k} = A_j W_j^{k+1} + B_j Z_j^{k+1} \), we have

\[
\lim_{k \to \infty} A_j W_j^k + B_j Z^k = 0.
\]
For any accumulation point \((W_1^*, \ldots, W_J^*, Z^*, \Lambda_1^*, \ldots, \Lambda_J^*, -\sum_{j=1}^{J} (B_j Z^*)^T \Lambda_j^*)\) of \(((W_1^k, \ldots, W_J^k, Z^k, \Lambda_1^k, \ldots, \Lambda_J^k, -\sum_{j=1}^{J} (B_j Z^k)^T \Lambda_j^k))_{k=1}^{\infty}\), there exists subsequence \(((W_1^{k_i}, \ldots, W_J^{k_i}, Z^{k_i}, \Lambda_1^{k_i}, \ldots, \Lambda_J^{k_i}, -\sum_{j=1}^{J} (B_j Z^{k_i})^T \Lambda_j^{k_i}))_{i=1}^{\infty}\) such that

\[
\lim_{i \to \infty} W_j^{k_i} = W_j^*, \quad \lim_{i \to \infty} Z^{k_i} = Z^*, \quad \lim_{i \to \infty} \Lambda_j^{k_i} = \Lambda_j^*.
\]

Hence, for each fixed \(j(j = 1, \ldots, J)\),

\[
A_j W_j^* + B_j Z^* = \lim_{i \to \infty} A_j W_j^{k_i} + B_j Z^{k_i} = 0,
\]

\[
(Z^*)^T Z^* = \lim_{i \to \infty} (Z^{k_i})^T Z^{k_i} = I,
\]

which means \((W_1^*, \ldots, W_J^*, Z^*)\) is a feasible point of problem 3.3. Since \(W_j^{k+1}\) solves equation 3.11, we have

\[
\frac{\beta_k}{\delta} \left( W_j^k - \delta A_j^T \left( A_j W_j^k + B_j Z^{k+1} - \frac{\Lambda_j^k}{\beta_k} \right) - W_j^{k+1} \right) \in \partial \| W_j^{k+1} \|_1
\]

for \(\forall k \geq 0\) or, equivalently,

\[
A_j^T \Lambda_j^k + \left( A_j^T A_j - \frac{1}{\delta} I \right) \beta_{k-1} (W_j^k - W_j^{k-1}) \in \partial \| W_j^k \|_1,
\]

for \(\forall k \geq 1\) by noticing

\[
B_j Z^k = \frac{1}{\beta_{k-1}} (\Lambda_j^{k-1} - \Lambda_j^j) - A_j W_j^k.
\]

Similarly, by passing to subsequence \(\{k_i\}\), letting \(i \to \infty\), and applying lemma 3, we can see that \(A_j^T \Lambda_j^* \in \partial \| W_j^* \|_1\). Therefore, \((W_1^*, \ldots, W_J^*, Z^*, \Lambda_1^*, \ldots, \Lambda_J^*, -\sum_{j=1}^{J} (B_j Z^*)^T \Lambda_j^*)\) satisfies conditions 3.4 in lemma 1. □

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376153, 62373211, 12271111, 11671171, 62376155), the Research Program of the Shanghai Municipal Science and Technology Committee (22511105600), Special Support Plan for High-Level Talents of Guangdong Province (2019TQ05X571), and Guangdong Basic and Applied Basic Research Foundation (2022A1515011726). We thank Xiaowei Zhang from OmniVision Technologies, Singapore, for discussing the theoretical
analysis of the proposed algorithm, Charilaos Kanatsoulis from the University of Minnesota for providing the PDD codes and useful discussions about the tuning parameters of PDD ($\ell_1$), and two anonymous reviewers for their valuable comments, which have greatly contributed to improving the quality of this article.

References

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2, 183–202. 10.1137/080716542

Benton, A., Arora, R., & Dredze, M. (2016). Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (vol. 2, pp. 14–19).

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., & Arora, R. (2019). Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP* (pp. 1–6).

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1–122. 10.1561/2200000016

Cai, J., & Sun, H. (2017). Constrained ERM learning of canonical correlation analysis: A least squares perspective. *Neural Computation*, 29(10), 2825–2859. 10.1162/neco_a_00996

Carroll, J. (1968). Equations and tables for a generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of Annual Convention of the American Psychological Association* (vol. 3, pp. 227–228).

Chartrand, R., & Wohlberg, B. (2013). A nonconvex ADMM algorithm for group sparsity with sparse groups. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6009–6013).

Chen, M., Gao, C., Ren, Z., & Zhou, H. H. (2017). Sparse CCA via precision adjusted iterative thresholding. In *Proceedings of the 7th International Congress of Chinese Mathematics ALM*, 44 (vol. 2, pp. 481–534).

Chu, D., Liao, L. Z., Ng, M. K., & Zhang, X. (2013). Sparse canonical correlation analysis: New formulation and algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 3050–3065. 10.1109/TPAMI.2013.104

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18), 3583–3593. 10.1093/bioinformatics/bth447

Eldén, L., & Park, H. (1999). A Procrustes problem on the Stiefel manifold. *Numerische Mathematik*, 82(4), 599–619.

Friman, O., Cedefamn, J., Lundberg, P., Borgia, M., & Knutsson, H. (2001). Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine*, 45(2), 323–330. 10.1002/1522-2594(200102)45:2(323::AID-MRM1041)3.0.CO;2-

Fu, X., Huang, K., Hong, M., Sidiropoulos, N. D., & So, A. M.-C. (2017). Scalable and flexible multiview MAX-VAR canonical correlation analysis. *IEEE Transactions on Signal Processing*, 65(16), 4150–4165. 10.1109/TSP.2017.2698365
Gao, C., Garber, D., Srebro, N., Wang, J., & Wang, W. (2019). Stochastic canonical correlation analysis. *Journal of Machine Learning Research, 20*(167), 1–46.

Gao, C., Ma, Z., & Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Annals of Statistics, 45*(5), 2074–2101. 10.1214/16-AOS1519

Ge, R., Jin, C., Netrapalli, P., Kakade, S. M., Netrapalli, P., & Sidford, A. (2016). Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning* (pp. 2741–2750).

Hale, E. T., Yin, W., & Zhang, Y. (2007). A fixed-point continuation method for $\ell_1$-regularized minimization with applications to compressed sensing. CAAM technical report TR07-07.

Hardoon, D. R., & Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Machine Learning, 83*(3), 331–353. 10.1007/s10994-010-5222-7

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation, 16*(12), 2639–2664. 10.1162/0899766042321814

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 2*(3/4), 321–377. 10.1093/biomet/28.3-4.321

Hu, H. (2014). Multiview gait recognition based on patch distribution features and uncorrelated multilinear sparse local discriminant canonical correlation analysis. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(4), 617–630. 10.1109/TCSVT.2014.2309834

Hurley, J. R., & Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7*(2), 258. 10.1002/bs.3830070216

Kanatsoulis, C. I., Fu, X., Sidiropoulos, N. D., & Hong, M. (2019). Structured SUMCOR multiview canonical correlation analysis for large-scale data. *IEEE Transactions on Signal Processing, 67*(2), 306–319. 10.1109/TSP.2018.2878544

Kang, M., Zhang, B., Wu, X., Liu, C., & Gao, J. (2013). Sparse generalized canonical correlation analysis for biological model integration: A genetic study of psychiatric disorders. In *Proceedings of the Engineering in Medicine and Biology Society* (pp. 1490–1493).

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika, 58*(3), 433–451. 10.1093/biomet/58.3.433

Knyazev, A. V., & Neymeyr, K. (2003). Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. *Electronic Transactions of Numerical Analysis, 15*, 38–55.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers* (vol. 5, pp. 79–86).

Koide-Majima, N., & Majima, K. (2021). Quantum-inspired canonical correlation analysis for exponentially large dimensional data. *Neural Networks, 135*, 55–67. 10.1016/j.neunet.2020.11.019

Li, H., & Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems, 28*, Curran.

Li, X., Xiu, X., Liu, W., & Miao, Z. (2022). An efficient Newton-based method for sparse generalized canonical correlation analysis. *IEEE Signal Processing Letters, 29*, 125–129. 10.1109/LSP.2021.3129170
Lin, Z., Liu, R., & Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low-rank representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), Advances in neural information processing systems, 24 (pp. 612–620). Curran.

Lippert, R., & Edelman, A. (2000). 9.4 Nonlinear eigenvalue problems with Orthogonality constraints. In Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, & H. van den Vorst (Eds.), Templates for the solution of algebraic eigenvalue problems: A practical guide. SIAM.

Liu, Q., Shen, X., & Gu, Y. (2019). Linearized ADMM for nonconvex nonsmooth optimization with convergence analysis. IEEE Access, 7, 76131–76144. 10.1109/ACCESS.2019.2914461

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. 10.1109/TKDE.2015.2445757

Ma, Z., Lu, Y., & Foster, D. (2015). Finding linear structure in large datasets with scalable canonical correlation analysis. In Proceedings of the International Conference on Machine Learning (pp. 169–1708).

Michaeli, T., Wang, W., & Livescu, K. (2016). Nonparametric canonical correlation analysis. In Proceedings of the International Conference on Machine Learning (pp. 1967–1976).

Rockafellar, R. T. (1970). Convex Analysis. Princeton University Press.

Sun, J., Xiu, X., Luo, Z., & Liu, W. (2023). Learning high-order multi-view representation by new tensor canonical correlation analysis. IEEE Transactions on Circuits and Systems for Video Technology, 33(10), 5645–5654. 10.1109/TCSVT.2023.3263853

Tan, K. M., 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. 10.1111/rssb.12291

Tan, M., Hu, Z., Yan, Y., Cao, J., Gong, D., & Wu, Q. (2021). Learning sparse PCA with stabilized ADMM method on Stiefel manifold. IEEE Transactions on Knowledge and Data Engineering, 33(3), 1078–1088.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).

Via, J., Santamaría, I., & Pérez, J. (2007). A learning algorithm for adaptive canonical correlation analysis of several data sets. Neural Networks, 20(1), 139–152.

Vinokourov, A., Cristianini, N., & Shawe-Taylor, J. (2003). Inferring a semantic representation of text via cross-language correlation analysis. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), Advances in neural information processing systems, 16 (pp. 1497–1504). MIT Press.

Waaijenborg, S., de Witt Hamer, P. C. V., & Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Statistical Applications in Genetics and Molecular Biology, 7(1). 10.2202/1544-6115.1329

Wang, W., Arora, R., Livescu, K., & Srebro, N. (2015). Stochastic optimization for deep CCA via nonlinear orthogonal iterations. In Proceedings of the 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (pp. 688–695).
Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics, 10*(3), 515–534. 10.1093/biostatistics/kxp008
Wright, S. J., Nowak, R. D., & Figueiredo, M. A. T. (2008). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing, 57*, 2479–2493. 10.1109/TSP.2009.2016892
Yamanishi, Y., Vert, J.-P., Nakaya, A., & Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics, 19*(suppl.), i323–i330. 10.1093/bioinformatics/btg1045
Zhang, L.-H., Wang, L., Bai, Z., & Li, R.-C. (2022). A self-consistent-field iteration for orthogonal canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(2), 890–904. 10.1109/TPAMI.2020.3012541
Zhang, L.-H., Yang, W. H., Shen, C., & Ying, J. (2020). An eigenvalue-based method for the unbalanced Procrustes problem. *SIAM Journal on Matrix Analysis and Applications, 41*(3), 957–983. 10.1137/19M1270872
Zhang, X., Chu, D., & Tan, R. C. (2015). Sparse uncorrelated linear discriminant analysis for undersampled problems. *IEEE Transactions on Neural Networks and Learning Systems, 27*(7), 1469–1485. 10.1109/TNNLS.2015.2448637

Received November 19, 2023; accepted February 21, 2024.