SPARSE GENERALIZED CANONICAL CORRELATION ANALYSIS VIA LINEARIZED BREGMAN METHOD

JIA CAI

School of Statistics and Mathematics
Collaborative Innovation Development Center of
Pearl River Delta Science & Technology Finance Industry
Guangdong University of Finance & Economics
Guangzhou, Guangdong, 510320, China

JUNYI HUO

School of Electronics and Computer Science, University of Southampton
University Road, Southampton, SO17 1BJ, United Kingdom

ABSTRACT. Canonical correlation analysis (CCA) is a powerful statistical tool for detecting mutual information between two sets of multi-dimensional random variables. Unlike CCA, Generalized CCA (GCCA), a natural extension of CCA, could detect the relations of multiple datasets (more than two). To interpret canonical variates more efficiently, this paper addresses a novel sparse GCCA algorithm via linearized Bregman method, which is a generalization of traditional sparse CCA methods. Experimental results on both synthetic dataset and real datasets demonstrate the effectiveness and efficiency of the proposed algorithm when compared with several state-of-the-art sparse CCA and deep CCA algorithms.

1. Introduction. Canonical correlation analysis (CCA) [16], is a celebrated data analysis tool that allows to detect the correlation between two sets of multidimensional variables. The two sets of multidimensional variables can be considered as two distinct objects or two views of the same object. CCA aims at finding two canonical variates (weight vectors) such that the projected variables in the lower-dimensional space are maximally correlated. Therefore, CCA has been widely used in many branches of science and technology, but not limited to, for instance, cross language document retrieval [23], genomic data analysis [26], functional magnetic resonance imaging [11]. However, CCA could not detect the mutual information of datasets that more than two. A natural extension of CCA, namely, generalized canonical correlation analysis (GCCA) [15, 18, 22] was proposed to handle the limitation of CCA. Tensor CCA was proposed [19] to deal with multiple datasets, too.
In practice, a large portion of features are not informative for high-dimensional data analysis. Typically, canonical loadings are not sparse. To make the canonical loadings more interpretable, a lot of sparse CCA algorithms were proposed extensively in the literature. Sparse penalized CCA algorithm [24], penalized matrix decomposition approach based sparse CCA method [25], and primal-dual framework [13] based sparse CCA were investigated. Moreover, Chu et al. selected the sparsest CCA solutions from a subset of all solutions via linearized Bregman method [9]. Chen et al. developed a precision adjusted iteration thresholding method to estimate the sparse canonical weights [8], while Gao et al. investigated a two stage based sparse CCA method [12], where the first initialization stage was solved by applying Alternating Direction Method with Multipliers (ADMM) method, and then a group-Lasso based method was utilized to find the sparse weights in the second refinement stage. However, how to achieve sparse GCCA algorithm is far less mature although it was introduced in [17].

In this paper, based on the model demonstrated in [22], we propose a novel sparse GCCA algorithm. Firstly, we convert GCCA problem into that of linear system of equations, and then apply linearized Bregman method to pursue sparsity. Experiments on both simulated dataset and real world datasets demonstrate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. We first introduce some preliminaries about CCA and GCCA, then present the sparse GCCA algorithm. Next we show the experimental studies. Finally, we conclude this paper and discuss the future work.

2. Preliminaries.

2.1. Canonical correlation analysis. In this section, we give a brief review of CCA. Let \( x \in \mathbb{R}^{n_1}, y \in \mathbb{R}^{n_2} \) be two random variables. Given \( m \) observations, denote \( X = (x_1, \cdots, x_m) \in \mathbb{R}^{n_1 \times m}, Y = (y_1, \cdots, y_m) \in \mathbb{R}^{n_2 \times m} \). Then CCA solves

\[
\max w_1^T X Y^T w_2 \quad s.t. \quad w_1^T X X^T w_1 = 1, \quad w_2^T Y Y^T w_2 = 1.
\] (2.1)

Multiple version of CCA takes form [9, 14]:

\[
\max_{W_1, W_2} \text{Trace}(W_1^T X Y^T W_2) \quad s.t. \quad \begin{cases}
W_1^T X X^T W_1 = I, & W_2^T Y Y^T W_2 = I, \\
W_1 \in \mathbb{R}^{n_1 \times \ell}, & W_2 \in \mathbb{R}^{n_2 \times \ell},
\end{cases}
\] (2.2)

where \( I \) stands for identity matrix, \( \ell \) is the number of columns of \( W_i \) (\( i = 1, 2 \)). When \( \ell = 1 \), problem (2.2) reduces to problem (2.1). Obviously, CCA could only deal with two datasets.

2.2. Generalized canonical correlation analysis. To analyze the case with more than two views, sum-of-correlations (SUMCOR) based generalized CCA was proposed [7],

\[
\min \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \|W_i^T X_i - W_j^T X_j\|_F^2,
\]

which has been shown to be NP-hard, where \( J \) stands for the number of views. Another formulation of GCCA is try to seek a common latent representation of different views [18]:

\[
\min_{\{W_j\}_{j=1}^J, G} \sum_{j=1}^J \|G - W_j^T X_j\|_F^2 \quad s.t. \quad GG^T = I_\ell,
\] (2.3)
where \( G \in \mathbb{R}^{f \times m} \) is a common latent representation of the different views. Problem (2.3) is also called MAX-VAR formulation of GCCA, since the optimal solution amounts to taking principal eigenvectors of a matrix aggregated from the correlation matrices of the views.

In fact, when \( J=2 \), Equation (2.3) becomes
\[
\min_{W_1,W_2,G} \| G - W_1^T X_1 \|_F^2 + \| G - W_2^T X_2 \|_F^2 \quad \text{s.t.} \quad GG^T = I_\ell. \tag{2.4}
\]
Notice that
\[
\| W_1^T X_1 - W_2^T X_2 \|_F^2 = \text{Trace}(W_1^T X_1 X_1^T W_1 + W_2^T X_2 X_2^T W_2 - W_1^T X_1 X_2^T W_1 W_2^T X_2 X_1^T W_2)
\]
\[
= 2\ell - 2\text{Trace}(W_1^T X_1 X_2^T W_2),
\]
by considering the constraints in Equation (2.2) and
\[
\| W_1^T X_1 - W_2^T X_2 \|_F^2 \leq 2\| W_1^T X_1 - G \|_F^2 + 2\| G - W_2^T X_2 \|_F^2.
\]
Therefore the target function in Equation (2.4) is a relaxation of the target function in (2.2).

3. Sparse GCCA algorithm. In order to pursue sparse GCCA algorithm, we need to reformulate problem (2.3). Firstly, we consider how to minimize \( \| G - W_j^T X_j \|_F^2 \) for each fixed \( j = 1, \cdots, J \). Obviously,
\[
\| G - W_j^T X_j \|_F^2 = \text{Trace}(G^T - W_j^T X_j G - G X_j^T W_j + W_j^T X_j X_j^T W_j),
\]
then taking derivatives with respect to \( W_j \), we can see that
\[
X_j X_j^T W_j = X_j G^T.
\]
In practice, \( X_j \) is normally a low rank matrix, e.g., the term-document matrix in Natural Language Processing (NLP). Therefore, to give a compact solution of \( X_j \), we use the compact approximation forms of \( X_j \). Let the reduced singular value decompositions (SVDs) of \( X_j \) \( (j = 1, \cdots, J) \) be
\[
X_j = P_j \Sigma_j Q_j^T, \tag{3.1}
\]
where \( P_j \in \mathbb{R}^{n_j \times r_j}, Q_j \in \mathbb{R}^{m \times r_j} \). Then \( X_j X_j^T = P_j \Sigma_j^2 P_j^T \). We can convert \( X_j X_j^T W_j = X_j G^T \) into \( P_j^T W_j = \Sigma_j^{-1} Q_j^T G^T \). If we denote
\[
A = \begin{pmatrix} P_1^T \\ \vdots \\ P_J^T \end{pmatrix}, \quad W = \begin{pmatrix} W_1 \\ \vdots \\ W_J \end{pmatrix}, \quad B = \begin{pmatrix} \Sigma_1^{-1} \\ \vdots \\ \Sigma_J^{-1} \end{pmatrix} \begin{pmatrix} Q_1^T \\ \vdots \\ Q_J^T \end{pmatrix} G^T, \tag{3.2}
\]
then \( P_j^T W_j = \Sigma_j^{-1} Q_j^T G^T \) becomes \( AW = B \), and we comes to
\[
\min_{W \in \mathbb{R}^{n \times \ell}} \| W \|_1 \quad \text{s.t.} \quad AW = B \tag{3.3}
\]
where \( n = n_1 + \cdots + n_J, \| W \|_1 = \sum_{s=1}^{n} \sum_{\tau=1}^{\ell} W(s, \tau) \). However, there is one problem that how to compute \( G \). Here, we consider the situation that \( G \) is fixed and employ the conclusion described in [18] directly. That is the rows of \( G \) are the eigenvectors of the matrix \( M = \sum_{j=1}^{J} X_j^T (X_j X_j^T)^{-1} X_j \). Normally, \( X_j X_j^T \) is not invertible in most situations. Thus, we use regularization technique or other schemes to modify this item. Notice that Reference [22] and this paper are both based on Equation (2.3). However, reference [22] converts Equation (2.3) into generalized
eigenvalue problem and uses least squares method to find the canonical variates. Here, we import sparsity and convert Equation (2.3) into Equation (3.3), then utilize linearized Bregman to obtain the canonical variates, which will be elucidated in the sequel. Recall the SVDs of $X_j$ (problem (3.1)), then $X_j X_j^T = P_j \Sigma_j^2 P_j^T$, simple computation leads to

$M = \sum_{j=1}^{J} Q_j Q_j^T$.

Now the problem is how to solve (3.3). Many numerical methods has been used to solve the following problem

$z^* = \operatorname{argmin}\{ \|z\|_1, z \in \mathbb{R}^\tau, Az = b\}$. (3.4)

Among these methods, linearized Bregman method [27] is one of the most powerful approach to solve this type of problem. The iteration procedure of linearized Bregman method can be stated as

$$
\begin{cases}
\epsilon^{k+1} = \nu^k + A^T (b - Az^k), \\
\nu^{k+1} = \delta S_\mu (\nu^{k+1}),
\end{cases}
$$

where $\delta, \mu > 0$, $z^0 = \nu^0 = 0$, $S_\mu (\cdot)$ is a soft-thresholding operator defined by

$S_\mu (x) = [S_\mu (x(1)), \cdots, S_\mu (x(\tau))]^T$,

with

$S_\mu (x) = \begin{cases}
0, & |x| \leq \mu \\
\operatorname{sgn}(x)(|x| - \mu), & |x| > \mu
\end{cases}$

**Algorithm 1** Sparse Generalized CCA algorithm.

**Input:**
Training data $X_j \in \mathbb{R}^{n_j \times m}$ ($j = 1, \cdots, J$), and tolerance parameter $\varepsilon$.

**Output:**
Sparse canonical variates $W$. 
1: Compute reduced SVD for each $X_j$ via Equation (3.1).
2: Compute top $\ell$ eigenvectors of matrix $M$.
3: Construct $G$ (setting the top $\ell$ eigenvectors of $M$ as the rows of $G$).
4: Let $W^0 = U^0 = 0$.
5: while error > $\varepsilon$ do
6: Compute $(U^{k+1}, W^{k+1})$ via (3.6),
7: error = $\|AW - B\|_F$,
8: end while
9: return $W = (W^T_1, \cdots, W^T_J)^T$.

Here, we extend the linearized Bregman method in problem (3.5) to solve the one stated in (3.3), and comes to Algorithm 1. In Algorithm 1, the following stopping criterion can be adopted:

$\|AW - B\|_F \leq \varepsilon$,

where $0 < \varepsilon < 1$ is a tolerance parameter. Actually, one can choose distinct $\varepsilon$ for different views. This is consistent with the two views case when choosing $J = 2$. 
Definition 1. The linearized Bregman based sparse GCCA (SGCCA) algorithm for computing the elements of matrix $W$ is defined by

$$
\begin{align*}
U^{k+1} &= U^k + A^T(B - AU^k), \\
W^{k+1} &= \delta S_\mu(U^{k+1}),
\end{align*}
$$

(3.6)

where $W^0 = U^0 = 0$, $S_\mu(U^k) = S_\mu(U(s, \tau))$.

The convergence of the proposed algorithm 1 is presented in the following theorem.

Theorem 1. Assume that $0 < \delta < \frac{1}{\|A^T A\|_2} = 1$, then the sequence $\{W^{k+1}\}$ generated by (3.6) converges to the unique solution of

$$
W^*_\mu = \arg\min \left\{ \mu \|W\|_1 + \frac{1}{2\delta} \|W\|_F \right\} \text{ s.t. } AW = B,
$$

(3.7)

e.i. \lim_{k \to \infty} \|W^k - W^*_\mu\| = 0. \text{ Denote } W^* \text{ as the solution of optimization problem in (3.3) that has the minimal Frobenius norm among all solutions of the optimization problem in (3.3), that is,}

$$
W^* = \arg\min \|W\|_F \text{ s.t. } W \text{ is a solution of problem (3.6)},
$$

then, \lim_{\mu \to \infty} \|W^*_\mu - W^*\| = 0.

The second part of the theorem is a trivial extension of [5, Proposition 2.3], we will omit the proof of it and only give the proof of the first part. This shall be done in the Appendix.

4. Experimental results.

4.1. Experiment on synthetic dataset. In this part, we constructed synthetic dataset to investigate the consistency of the proposed sparse algorithms for GCCA problem. We constructed three matrices $X$, $Y$, and $Z$, which are corrupted with noise and take the form:

$$
X = v_1 u^T + \epsilon_1, \quad Y = v_2 u^T + \epsilon_2, \quad Z = v_3 u^T + \epsilon_3,
$$

where

$$
v_1 = \begin{pmatrix} 1, \ldots, 1, -1, \ldots, -1, 0, \ldots, 0 \end{pmatrix}^T,
$$

$$
v_2 = \begin{pmatrix} 0, \ldots, 0, 1, \ldots, 1, -1, \ldots, -1 \end{pmatrix}^T,
$$

$$
v_3 = \begin{pmatrix} 1, \ldots, 1, 0, \ldots, 0, -1, \ldots, -1 \end{pmatrix}^T,
$$

$\epsilon_1 \in \mathbb{R}^{10000 \times 100} \sim \mathcal{N}(0, 0.3^2)$, $\epsilon_2 \in \mathbb{R}^{15000 \times 100} \sim \mathcal{N}(0, 0.4^2)$, and $\epsilon_3 \in \mathbb{R}^{18000 \times 100} \sim \mathcal{N}(0, 0.5^2)$, are three random noise matrices. $u \in \mathbb{R}^{100}$ is a random vector with all entries being drawn from the normal distribution, i.e.,

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

Settings We randomly selected half of the data as training data and used the rest for testing, this procedure was repeated 30 times, and we report the average results. GCCA was used as the baseline algorithm. For WGCCA [2], Gaussian weight were utilized, while for Deep GCCA [3] method, we chose epoch = 100,
batch size as the integer part of ratio × #trnum, where the ratio is chosen from
the candidate set \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}, and #trnum stands for the number of training samples.
For the proposed algorithm, we choose \( \delta = 0.5, \mu = 20 \).

**Results** Plots of \( X, Y \) and \( Z \) are presented in Figure 1. Table 1 shows the comparison results for the sparsity, the reconstruction error both for training data and testing data, where the reconstruction error for testing is defined as

\[
\frac{1}{J} \sum_{j=1}^{J} \| W_j^T X_{ts}^j - G \|_F^2.
\]

Note that for multiple datasets, summation of correlation coefficients is not appropriate for measuring the property of a good GCCA algorithm. To give a more comprehensive understanding of generalized CCA methods, we also indicate the above quantity for training data, i.e. replacing \( X_{ts}^j \) with \( X_{tr}^j \) for \( j = 1, \cdots, J \). We can see that only WGCCA and sparse GCCA can achieve sparsity results. Table 1 demonstrates that Sparse GCCA can obtain the highest sparsity result and comparable small reconstruction error. Deep CCA obtains the smallest reconstruction testing error, but it achieves higher reconstruction training error, which means Deep GCCA does not learn a good representation of \( G \) and canonical variates \( W_j \) (\( j = 1, \cdots, J \)).

4.2. **Experiment on real datasets.** Our experiments are performed on the following 5 datasets: the first four datasets are from the gene database\(^1\), and the rest one is from the JRC-Acquis database. The details are explained below and the statistics can be found in Table 2.

- Leukemia: gene expression values for 72 samples (47 samples from patients with acute lymphoblastic leukemia, and 25 from patients with acute myeloblastic leukemia).

\(^1\) [http://stat.ethz.ch/~dettling/bagboost.html](http://stat.ethz.ch/~dettling/bagboost.html)
Table 1. Comparison results on synthetic dataset over 30 training-testing replications: reconstruction error for training data (Trerror), reconstruction error for testing data (Terror), and sparsity (Spar1, Spar2, Spar3 stand for the sparsity of the first view, second view, and third view, respectively) obtained by GCCA, Deep GCCA, WGCCA and SGCCA algorithms.

| Method    | GCCA       | Deep GCCA  | WGCCA     | SGCCA     |
|-----------|------------|------------|-----------|-----------|
| Trerror   | 3.7277e-30 | 0.5043     | 1.1118e-07| 1.6354e-10|
| Tterror   | 0.4756     | 0.2955     | 0.6860    | 0.4554    |
| Spar1 (%) | 2.73       | 0.01       | 70.01     | 99.47     |
| Spar2 (%) | 5.15       | 0.03       | 78.19     | 99.67     |
| Spar3 (%) | 7.54       | 0.09       | 83.02     | 99.72     |

- Lymphoma: 42 samples of diffuse large B-cell lymphoma, 9 observations of follicular lymphoma and 11 cases of chronic lymphocytic leukemia. The total sample size is 62, and the expression of 4026 well-measured genes.
- Prostate: gene expression values measured by Affymetrix human 95Av2 arrays for 102 samples (52 prostate tumors and 50 nontumors prostate samples).
- Brain: 42 microarray gene expression profiles from five different tumors of the central nervous system.
- JRC-Acquis [21]: a collection of parallel texts in the following 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish. (see Figure 2)

Table 2. Data Structures: data dimension (n), number of data (m), number of classes (K), number of columns in $W_1$ and $W_2$ (ℓ).

| Type      | Data       | n   | m   | K  | ℓ  |
|-----------|------------|-----|-----|----|----|
| Gene Data | Leukemia   | 3571| 72  | 2  | 1  |
|           | Lymphomia  | 4026| 62  | 3  | 2  |
|           | Prostate   | 6033| 102 | 2  | 1  |
|           | Brain      | 5597| 42  | 5  | 4  |

Settings The preprocessing procedure for gene data is described in [10]. For JRC-Acquis database, we first extracted the “body” part of the XML file, and then obtained a bag-of-words representation by using Term Frequency Inverse Document Frequency (TFIDF) approach, which is efficient in the document retrieval task. After removing numbers, stop-words (English, Spanish, and French, respectively), and rare words (appearing less than twice), we obtained $12421 \times 1000$ “English” matrix, $17203 \times 1000$ “Spanish” matrix and $15650 \times 1000$ “French” matrix. For experiments on gene datasets, we compared sparse GCCA with CCA, PMD CCA [24], Deep CCA ([11]) and GCCA, where CCA was considered as the baseline algorithm. There are two parameters controlling the sparsity results of PMD CCA method, which were selected from the candidate set $0.01 : 0.01 : 0.5$ (between 0.01 and 0.5 with a step increment of 0.01). For Deep CCA algorithm, we chose epoch= 100, learning rate from the candidate set $\{1e^{-3}, 1e^{-2}, 1e^{-1}\}$. The batch size was chosen as the integer part of ratio $\times$ #trnum, where the ratio was chosen from the candidate set $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$. For the proposed algorithm, we choose $\delta = 0.9$, $\mu = 10$.
(where $\delta$ is the indicator function that equals 1 if $y_1 = y_2$ and 0 otherwise. For a given sample point $(x_j, y_j)$, $\delta$ is obtained as $\delta(y_1, y_2) = 1$ if $y_1 = y_2$ and 0 otherwise. Accurate reconstruction error is then achieved for the reconstruction error, that is, $\delta(y_1, y_2)$.

The proposed algorithm achieves the highest sparsity results followed by PDM CCA method (except Prostate dataset, the difference is subtle, only 0.05%), and comparable accuracy results. Accuracy is defined as follows:

$$\text{Accuracy} := \frac{\sum_{i=1}^{N} \delta(o_i, p_i)}{N} \times 100\%,$$

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 $(x_j, y_j)$, $\delta$ is obtained as $\delta(y_1, y_2) = 1$ if $y_1 = y_2$ and 0 otherwise. Accurate reconstruction error is then achieved for the reconstruction error, that is, $\delta(y_1, y_2)$.

The proposed algorithm achieves the highest sparsity results followed by PDM CCA method (except Prostate dataset, the difference is subtle, only 0.05%), and comparable accuracy results. Accuracy is defined as follows:

$$\text{Accuracy} := \frac{\sum_{i=1}^{N} \delta(o_i, p_i)}{N} \times 100\%,$$

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 $(x_j, y_j)$, $\delta$ is obtained as $\delta(y_1, y_2) = 1$ if $y_1 = y_2$ and 0 otherwise. Accurate reconstruction error is then achieved for the reconstruction error, that is, $\delta(y_1, y_2)$.
Table 3. Comparison results on gene data over 30 training-testing replications: the average of sparsity (Spar1 and Spar2 stand for the sparsity of the first view and second view, respectively), classification accuracy, and the summation of correlation coefficients (SCORR).

| Methods   | Spar1(%) | Spar2(%) | Accuracy (%) | SCORR  |
|-----------|----------|----------|--------------|--------|
| Leukemia  |          |          |              |        |
| CCA       | 0.03     | 0        | 97.20        | 0.8945 |
| PMD CCA   | 98.48    | 0        | 72.22        | 0.8944 |
| GCCA      | 5.80     | 50       | 97.20        | 0.9227 |
| Deep CCA  | 0        | 0        | 94.40        | 0.8662 |
| SGCCA     | 98.99    | 50       | 100          | 0.8916 |
| Lyphomia  |          |          |              |        |
| CCA       | 0.04     | 0        | 83.87        | 1.8206 |
| PMD CCA   | 85.38    | 0        | 77.97        | 1.7911 |
| GCCA      | 5.77     | 50       | 90.32        | 1.7423 |
| Deep CCA  | 0        | 0        | 80.65        | 0.9754 |
| SGCCA     | 99.23    | 50       | 96.77        | 1.6933 |
| Prostate  |          |          |              |        |
| CCA       | 0.1      | 0        | 88.24        | 0.7646 |
| PMD CCA   | 99.19    | 0        | 60.78        | 0.2978 |
| GCCA      | 4.59     | 50       | 88.24        | 0.7645 |
| Deep CCA  | 0        | 0        | 79.41        | 0.7610 |
| SGCCA     | 99.14    | 50       | 86.27        | 0.7758 |
| Brain     |          |          |              |        |
| CCA       | 0.07     | 0        | 71.43        | 2.9749 |
| PMD CCA   | 76.40    | 74.65    | 47.62        | 3.0669 |
| GCCA      | 4.52     | 34.9     | 61.90        | 2.8307 |
| Deep CCA  | 0        | 0        | 61.90        | 2.9674 |
| SGCCA     | 99.62    | 34.9     | 66.67        | 2.5729 |

Table 4. Comparison results on JRC-Acquis database over 30 training-testing replications for \( \ell = 100 \): reconstruction error for training data (Trerror), reconstruction error for testing data (Tsererror), and sparsity (Spar1, Spar2, Spar3 stand for the sparsity of the first view, second view, and third view, respectively) obtained by GCCA, WGCCA and SGCCA algorithms.

|          | GCCA     | Deep GCCA | WGCCA    | SGCCA    |
|----------|----------|-----------|----------|----------|
| Trerror  | 2.2078e-30  | 3.0859e-02 | 8.4776e-12 | 5.5140e-11 |
| Tsererror| 4.1496e-02  | 1.7648e-02 | 0.7138   | 3.8799e-02 |
| Spar1 (%)| 7.00      | 0.1140    | 10.27    | 96.56    |
| Spar2 (%)| 7.24      | 0.6850    | 9.96     | 96.85    |
| Spar3 (%)| 6.09      | 0.1100    | 8.49     | 95.64    |

error for training data, which means Deep GCCA method may lead to over-fitting results. Since the proposed Sparse GCCA algorithm is achieved based upon the traditional GCCA method, the experimental results indicate that sparse GCCA obtains higher sparsity result, and meanwhile it achieves lower accuracy or higher reconstruction error when comparing sparse GCCA with GCCA methods.

5. Conclusion. CCA is a widely discussed and used statistical tool. But CCA could only deal with two datasets. Therefore, GCCA is studied to handle the
we achieve sparse GCCA algorithm by imposing \( \ell_1 \) norm on canonical variates. Employing the idea of linearized Bregman iteration, for high dimensional data. To make the canonical variates more interpretable, in this paper, under the framework of GCCA, we reformulate GCCA problem into that of linear system of equations. Employing the idea of linearized Bregman iteration, we achieve sparse GCCA algorithm by imposing \( \ell_1 \) norm on canonical variates. Experimental results on both synthetic dataset and real world datasets demonstrate the effectiveness of the proposed sparse GCCA algorithm. For future work we will investigate how to obtain sparse deep CCA algorithm.

**Appendix.** In this part, we first give a brief review of Bregman distance and linearized Bregman iteration method, then present the proof of Theorem 1. Bregman Distance is defined by \( [4] \)

\[
D_F^u(w, v) = F(w) - F(v) - \langle w - v, u \rangle,
\]

where \( F \) is a convex function. \( u \in \partial F(v) \) is a subgradient in the subdifferential of \( F \) at the point \( v \). \( D_F^u(w, v) \) is not a distance in the usual sense since \( D_F^u(w, v) \neq D_F^u(v, w) \), and it is nonnegative obviously. Bregman iteration for general problem \( \min_{w \in \mathbb{R}^n} \{ F(w) : Aw = b \} \) is given by \( [20] \)

\[
\begin{aligned}
w^{k+1} &= \arg \min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Aw - b\|^2 + \mu D_F^k(w, w^k) \right\}, \\
u^{k+1} &= u^k - \frac{1}{\mu} A^T (Aw^k - b),
\end{aligned}
\]

for \( w^0 = u^0 = 0 \). Letting \( b^k = \sum_{i=0}^{k-1} (b - Aw^i) \), if \( F(w) = \|w\|_1 \), then (5.2) becomes

\[
\begin{aligned}
b^{k+1} &= b^k + b - Aw^k, \\
w^{k+1} &= \arg \min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Aw - b^{k+1}\|^2 + \mu \|w\|_1 \right\},
\end{aligned}
\]

In order to solve the second step of (5.3), based on the Taylor expansion of \( \|Aw-b\|^2 \) around \( w^k \),

\[
\|Aw - b\|^2 \approx \|Aw^k - b\|^2 + 2(w - w^k, A^T (Aw^k - b)) + \frac{1}{\delta} \|w - w^k\|^2,
\]

where \( \delta \) is a fixed parameter. We have the following linearized Bregman iteration for given \( w^0 = u^0 = 0 \), i.e.,

\[
\begin{aligned}
w^{k+1} &= \arg \min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2\delta} \|w - w^k + \delta A^T (Aw^k - b)\|^2 + \mu D_F^k(w, w^k) \right\}, \\
u^{k+1} &= u^k - \frac{1}{\mu} (w^{k+1} - w^k) - \frac{1}{\mu} A^T (Aw^k - b),
\end{aligned}
\]

where the second expression for updating \( w^{k+1} \) is to ensure that \( w^{k+1} \in \partial J(w^{k+1}) \). Details could be found in \( [6] \). It was shown in \( [6, 27] \) that if \( F(w) = \|w\|_1 \), (5.4) can be reformulated into its compact form of (3.5). In fact, (5.4) is equivalent to gradient descent applied to the dual of problem

\[
\min_{w \in \mathbb{R}^n} \left\{ \mu F(w) + \frac{1}{2\delta} \|w\|^2 : Aw = b \right\}.
\]

We need the following three lemmas before giving the proof of Theorem 1.

**Lemma 1** \( [27] \). Assume that \( \delta A^T A < I \), then

\[
\|Aw^{k+1} - b\|^2 + \left( \frac{1}{\delta} - \|A^T A\| \right) \|w^{k+1} - w^k\|^2 \leq \|Aw^k - b\|^2.
\]
**Lemma 2** ([5]). Assume that \( \delta A^T A < I \), then the sequences \( \{w^k\}_{k \in \mathbb{N}} \) and \( \{u^k\}_{k \in \mathbb{N}} \) are bounded.

**Lemma 3** ([5]). Assume that \( \delta A^T A < I \), then
\[
\lim_{k \to \infty} \| A^T (A w^k - b) \| = 0.
\]

In particular, when \( A \) is rectangular and surjective, \( A w^k \) converges to \( b \), i.e.,
\[
\lim_{k \to \infty} \| A w^k - b \| = 0.
\]

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

**Proof.** Denote \( U^k = (u_1^k, \ldots, u_{\ell}^k) \), \( B = (b_1, \ldots, b_{\ell}) \), \( W^k = (w_1^k, \ldots, w_{\ell}^k) \), \( W^* = (w_{\mu,1}^*, \ldots, w_{\mu,\ell}^*) \). Obviously, (3.5) is the generalization of linearized Bregman in the vector case. Furthermore, \( w_{\mu,i}^* (i = 1, \ldots, \ell) \) is the minimizer of
\[
w_{\mu,i}^* = \arg \min \left\{ \mu \| w_i \|_1 + \frac{1}{2\delta} \| w_i \|_2^2 \right\} \quad \text{s.t.} \quad A w_i = b_i
\]
for each fixed \( i = 1, \ldots, \ell \). This can be proved by contradiction. On one hand, \( A w = B \) is equivalent to \( A w_i = b_i (i = 1, \ldots, \ell) \). On the other hand, \( W^* = (w_{\mu,1}^*, \ldots, w_{\mu,\ell}^*) \) yields that (5.6) holds true. Or else, there exist \( \tilde{w}_{\mu,s}^* \) for some \( s \in \{1, \ldots, \ell\} \) such that
\[
\tilde{w}_{\mu,s}^* = \arg \min \left\{ \mu \| w_s \|_1 + \frac{1}{2\delta} \| w_s \|_2^2 \right\} \quad \text{s.t.} \quad A w_s = b_s,
\]
and \( \tilde{w}_{\mu,s}^* \neq w_{\mu,s}^* \). Then construct \( \tilde{W}_{\mu}^* = (w_{\mu,1}^*, \ldots, \tilde{w}_{\mu,s}^*, \ldots, w_{\mu,\ell}^*) \), which leads to
\[
\tilde{W}_{\mu,s}^* = \arg \min \left\{ \mu \| W \|_1 + \frac{1}{2\delta} \| W \|_F \right\} \quad \text{s.t.} \quad A W = B.
\]

This is a contradiction to that \( W^* \) is the unique solution of (5.6). Therefore, we only need to prove that \( \lim_{k \to \infty} \| w^k_i - w_{\mu,i}^* \| = 0 \) for each fixed \( i = 1, \ldots, \ell \). We shall employ the idea described in the proof of [5, Theorem 2.4]. To allow a slightly abuse of notations, we omit the subscript \( i \) for \( u_i^k, b_i, w_i^k \) and \( w_{\mu,i}^* \) for simplicity. Lemma 2 implies that \( w^k \) is bounded, then there exist convergence subsequences. We present that the limit of an arbitrary convergent subsequence is the unique solution \( w_{\mu,i}^* \) of (5.5). By the uniqueness of \( w_{\mu,i}^* \), we obtain \( \lim_{k \to \infty} \| w^k_i - w_{\mu,i}^* \| = 0 \). Let \( \{w^k_i\}_{i \in \mathbb{N}} \) be an arbitrary given convergent subsequence and \( \tilde{w} \) its limit, we prove \( \tilde{w} = w_{\mu,i}^* \) in the sequel. Note that \( u^0 = w^0 = 0 \), the second equation of (5.4) yields that
\[
\mu u^k + \frac{1}{\delta} w^k = A^T \sum_{j=0}^{k-1} (b - A w^j).
\]

Define \( s^k = \sum_{j=0}^{k-1} (b - A w^j) \), since \( A \) is surjective and both \( \{u^k \in [-1, 1]^n\}_{k \in \mathbb{N}} \) and \( \{w^k \}_{k \in \mathbb{N}} \) are bounded. We can see that \( \| s^k \| \leq C \) for some constant \( C \) for all \( k \). Denote \( F(w) = \mu \| w \|_1 + \frac{1}{2\delta} \| w \|_2^2 \). Obviously, \( u^0 \in \| \partial w^0 \|_1 \). Since \( u^0 = w^0 = 0 \), by the definition of \( u^k \) and \( w^k \) in (5.4), we have \( u^k \in \partial \| w^k \|_1 \). By induction and differentiating the first part in the first equation of (5.4), thus, \( u^k \in \partial \| w^k \|_1, \mu u^k_i + \frac{1}{\delta} w^k_i \in \partial F(w^k_i) \). The nonnegativity of the Bregman distance \( D_F(\mu u^k_i + \frac{1}{\delta} w^k_i, (w_{\mu,i}^*, w_{\mu,i}^*)) \) in Equation (5.1) for \( F(w) \) leads to
\[
F(w^k_i) \leq F(w_{\mu,i}^*) - \langle w_{\mu,i}^* - w^k_i, \mu u^k_i + \frac{1}{\delta} w^k_i \rangle.
\]
By Cauchy-Schwartz inequality, we have
\[ |\langle A(w^*_\mu - w^{k_i}), s^{k_i} \rangle| \leq \| A(w^*_\mu - w^{k_i}) \| \cdot \| s^{k_i} \| \leq C \| A(w^*_\mu - w^{k_i}) \|.
\]
Letting \( i \to \infty \), and noticing that
\[ \lim_{i \to \infty} A w^{k_i} = A \tilde{w} = b = A w^*_\mu. \]
We can see that
\[ \lim_{i \to \infty} |\langle A(w^*_\mu - w^{k_i}), s^{k_i} \rangle| = 0. \]
In (5.7), let \( i \to \infty \), we have \( F(\tilde{w}) \leq F(w^*_\mu) \). This together with \( A \tilde{w} = b \) and the uniqueness of \( w^*_\mu \) implies that \( \tilde{w} = w^*_\mu \), which completes the proof. \( \square \)

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes and K. Livescu, Deep canonical correlation analysis, in International Conference on Machine Learning, (2013), 1247–1255.
[2] A. Benton, R. Arora, and M. Dredze, Learning multiview embeddings of twitter users, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2 (2016), 14–19.
[3] A. Benton, H. Khayrallah, B. Gujral and et al., Deep Generalized Canonical Correlation Analysis, in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), (2019), 1–6.
[4] L. M. Brègman, A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming, Zh. Vychisl. Mat. Mat. Fiz., 7 (1967), 620–631.
[5] J. F. Cai, S. Osher and Z. Shen, Convergence of the linearized bregman iteration for \( \ell_1 \)-norm minimization, Math. Comp., 78 (2009), 2127–2136.
[6] J. F. Cai, S. Osher and Z. Shen, Linearized bregman iterations for compressed sensing, Math. Comp., 78 (2009), 1515–1536.
[7] J. Carroll, 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, 3 (1968), 227–228.
[8] M. Chen, C. Gao, Z. Ren and et al., Sparse cca via precision adjusted iterative thresholding, in Proceedings of International Congress of Chinese Mathematicians, (2016).
[9] D. Chu, L. Z. Liao, M. K. Ng and X. W. Zhang, Sparse canonical correlation analysis: new formulation and algorithm, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2013), 3050–3065.
[10] M. Dettling, Bagboosting for tumor classification with gene expression data, Bioinformatics, 20 (2004), 3583–3593.
[11] O. Frieman, J. Cedeño-Mann, P. Lundberg, H. Borga and H. Knutsson, Detection of neural activity in functional mri using canonical correlation analysis, Magn. Reson. Med., 45 (2001), 323–330.
[12] C. Gao, Z. Ma and H. H. Zhou, Sparse cca: Adaptive estimation and computational barriers, Ann. Statist., 45 (2017), 2074–2101.
[13] D. R. Hardoon and J. Shawe-Taylor, Sparse canonical correlation analysis, Mach. Learn., 83 (2011), 331–353.
[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput., 16 (2004), 2639–2664.
[15] P. Horst, Generalized canonical correlations and their applications to experimental data, J. Clin. Psychol., 17 (1961), 331–347.
[16] H. Hotelling, Relations between two sets of variates, Biometrika, 2 (1936), 321–377.
[17] M. Kang, B. Zhang, X. Wu, C. Y. Liu and J. Gao, Sparse generalized canonical correlation analysis for biological model integration: a genetic study of psychiatric disorders, in 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (2013), 1490–1493.
[18] J. R. Kettenring, Canonical analysis of several sets of variables, Biometrika, 58 (1971), 433–451.
[19] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu and Y. G. Wen, Tensor canonical correlation analysis for multi-view dimension reduction, IEEE Trans. Knowl. Data Eng., 27 (2015), 3111–3124.

[20] S. Osher, M. Burger, D. Goldfarb, J. J. Xu and W. T. Ying, An iterative regularization method for total variation-based image restoration, Multiscale Model. Simul., 4 (2005), 460–489.

[21] R. Steinberger, B. Pouliquen, A. Widiger and et al., The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), (2006), 2142–2147.

[22] J. Vía, I. Santamaría, and J. Pérez, A learning algorithm for adaptive canonical correlation analysis of several data sets, Neural Netw., 20 (2007), 139–152.

[23] A. Vinokourov, N. Cristianini and J. Shawe-Taylor, Inferring a semantic representation of text via cross-language correlation analysis, in Advances in Neural Information Processing Systems, (2003), 1497–1504.

[24] S. Waaijenborg, P. C. V. de Witt Hamer and A. H. Zwinderman, Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis, Stat. Appl. Genet. Mol. Biol., 7 (2008), Art. 3.

[25] D. M. Witten, R. Tibshirani and T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 10 (2009), 515–534.

[26] Y. Yamani, J. P. Vert, A. Nakaya and M. Kanehisa, Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, Bioinformatics, 19 (2003), i323–i330.

[27] W. Yin, S. Osher, D. Goldfarb and J. Darbon, Bregman iterative algorithms for 1-minimization with applications to compressed sensing, SIAM J. Imaging Sci., 1 (2008), 143–168.

Received February 2019; revised August 2019.

E-mail address: jiacai1999@gdufe.edu.cn
E-mail address: jh4a19@soton.ac.uk.