Chinese Journal of Electronics
Vol.27, No.3, May 2018

Sparse Weighted Canonical Correlation Analysis*

MIN Wenwen[1], LIU Juan[1] and ZHANG Shihua[2,3]

(1. *State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China*)

(2. *National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*)

(3. *School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*)

Abstract — Given two data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, Sparse canonical correlation analysis (SCCA) is to seek two sparse canonical vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to maximize the correlation between $\boldsymbol{Xu}$ and $\boldsymbol{Yv}$. Classical and sparse Canonical correlation analysis (CCA) models consider the contribution of all the samples of data matrices and thus cannot identify an underlying specific subset of samples. We propose a novel Sparse weighted canonical correlation analysis (SWCCA), where weights are used for regularizing different samples. We solve the $\boldsymbol{L_0}$-regularized SWCCA ($\boldsymbol{L_0}$-SWCCA) using an alternating iterative algorithm. We apply $\boldsymbol{L_0}$-SWCCA to synthetic data and real-world data to demonstrate its effectiveness and superiority compared to related methods. We consider also SWCCA with different penalties like Least absolute shrinkage and selection operator (LASSO) and Group LASSO, and extend it for integrating more than three data matrices.

Key words — Canonical correlation analysis (CCA), Sparse canonical correlation analysis (SCCA), Sparse weighted CCA (SWCCA), Group LASSO regularized SWCCA, Multi-view SWCCA.

I. Introduction

Canonical correlation analysis (CCA) is a powerful tool to integrate two data matrices[1−5], which has been comprehensively used in many diverse fields. Given two matrices $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$ from the same samples, CCA is used to find two sparse canonical vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to maximize the correlation between $\boldsymbol{Xu}$ and $\boldsymbol{Yv}$. However, in many real-world problems like those in bioinformatics[6−10], the number of variables in each data matrix is usually much larger than the sample size. The classical CCA leads to non-sparse canonical vectors which are difficult to interpret in biology. To conquer this issue, a large number of sparse CCA models[6−16] have been proposed by using regularized penalties (*e.g.*, LASSO and L_0-norm) to obtain sparse canonical vectors for variable selection. Parkhomenko *et al.*[11] first proposed a Sparse CCA (SCCA) model using Least absolute shrinkage and selection operator (LASSO) (L_1-norm) penalty to genomic data integration. Lê Cao *et al.*[8] further proposed a regularized CCA with Elastic-net penalty for a similar task. Witten *et al.*[6] proposed the Penalized matrix decomposition (PMD) algorithm to solve the Sparse CCA with two penalties: LASSO and Fused LASSO to integrate DNA copy number and gene expression from the same samples/individuals. Furthermore, a large number of generalized LASSO regularized CCA models have been proposed to consider prior structural information of variables[17−21]. For example, Lin *et al.*[17] proposed a Group LASSO regularized CCA to explore the relationship between two types of genomic data sets. If we consider a pathway as a gene group, then these gene pathways form an overlapping group structure[22]. Chen *et al.*[20] developed an overlapping group LASSO regularized CCA model to employ such group structure.

These existing sparse CCA models can find two sparse canonical vectors with a small subset of variables across all samples (Fig.1(a)). However, many real data such as the cancer genomic data show distinct heterogeneity[23,24]. Thus, the current CCA models fail to consider such heterogeneity and cannot directly identify a set of sample-specific correlated variables. To this end, we propose a novel Sparse weighted CCA (SWCCA) model, where weights are used for regularizing different samples with a typical penalty (*e.g.*, LASSO and L_0-norm) (Fig.1(b)). In

*Manuscript Received Dec. 2, 2016; Accepted June 6, 2017. This work is supported by the National Natural Science Foundation of China (No.61422309, No.61379092, No.61621003, No.11661141019), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No.XDB13040600), the National Science Foundation of Jiangsu Province (No.BK20161249), the Fundamental Research Funds for the Central Universities (No.2042017KF0233), CAS Frontier Science Research Key Project for Top Young Scientist (No.QYZDB-SSW-SYS008), and the Key Laboratory of Random Complex Structures and Data Science, CAS (No.2008DP173182).

 DOI:10.1049/cje.2017.08.004

this way, SWCCA can not only select two variable sets, but also select a sample set (Fig.1(b)). We further adopt an efficient alternating iterative algorithm to solve L_0 (or L_1) regularized SWCCA model. We apply L_0-SWCCA and related ones onto two simulated datasets and two real biological data to demonstrate its efficiency in capturing correlated variables across a subset of samples.

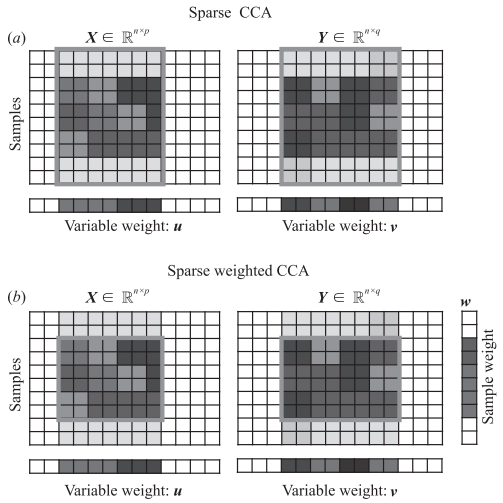

Fig. 1. Illustration of the difference between SWCCA and SCCA. (a) SCCA is used to extract two sparse canonical vectors ($\boldsymbol{u}$ and $\boldsymbol{v}$) to measure the association of two matrices; (b) SWCCA is used to consider two subset of sample-related sparse canonical vectors. The weights ($\boldsymbol{w}$) are used for regularizing different samples in SWCCA. SWCCA can not only obtain two sparse canonical vectors, but also identify a set of samples based on those non-zero elements of $\boldsymbol{w}$

II. L_0-Regularized SWCCA

Here, we assume that there are two data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ (n samples and p variables) and $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$ (n samples and q variables) across a same set of samples. The classical CCA seeks two components ($\boldsymbol{u}$ and $\boldsymbol{v}$) to maximize the correlation between linear combinations of variables from the two data matrices as Eq.(1).

$$\rho = \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{xy}\boldsymbol{v}}{\sqrt{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{x}\boldsymbol{u})(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_{y}\boldsymbol{v})}} \tag{1}$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered, we obtain the empirical covariance matrices $\boldsymbol{\Sigma}_{xy} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, $\boldsymbol{\Sigma}_{x} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{\Sigma}_{y} = \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}$. Thus we have the following equivalent criterion as Eq.(2).

$$\rho = \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}}{\sqrt{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u})(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v})}} \tag{2}$$

Obviously, ρ of Eq.(2) is invariant to the scaling of $\boldsymbol{u}$ and $\boldsymbol{v}$. Thus, maximizing criterion (2) is equivalent to solve the following constrained optimization problem as Eq.(3).

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u} = 1, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} = 1 \end{aligned} \tag{3}$$

Previous studies[6,12] have shown that considering the covariance matrix ($\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, $\frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$) as diagonal one can obtain better results. For this reason, Asteris *et al.*[13] assume that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}$ and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{I}$, and the L_0-regularized Sparse CCA (L_0-SCCA) (also called "diagonal penalized CCA") model can be presented as Eq.(4).

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} \\ \text{s.t.} \quad & \|\boldsymbol{u}\|_0 \le k_u, \|\boldsymbol{v}\|_0 \le k_v \\ & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \end{aligned} \tag{4}$$

where $\|\boldsymbol{u}\|_0$ is the L_0-norm penalty function, which returns to the number of non-zero entries of $\boldsymbol{u}$. Asteris *et al.*[13] applied a projection strategy to solve L_0-SCCA. Let $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, then the model of Eq.(4) is equivalent to rank-one L_0-SVD model[25].

Let $\boldsymbol{a} = \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{b} = \boldsymbol{Y}\boldsymbol{v}$, then the objective function $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} = \sum_{i=1}^{n} a_i b_i$. To consider the different contribution for samples, we modify the objective function of Eq.(4) to be $\sum_{i=1}^{n} w_i(a_i b_i)$ with $\boldsymbol{w} = [w_1, w_2, \cdots, w_n]^{\mathrm{T}}$. Thus, we obtain a new objective function as Eq.(5).

$$\sum_{i=1}^{n} w_i(a_i b_i) = \boldsymbol{u}\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} \tag{5}$$

Furthermore, we also force $\boldsymbol{w}$ to be sparse to select a limited number of samples. Finally we propose a L_0-regularized SWCCA (L_0-SWCCA) model as Eq.(6).

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} \\ \text{s.t.} \quad & \|\boldsymbol{u}\|_0 \le k_u, \|\boldsymbol{v}\|_0 \le k_v, \|\boldsymbol{w}\|_0 \le k_w \\ & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1 \end{aligned} \tag{6}$$

where $\mathrm{diag}(\boldsymbol{w})$ is a diagonal matrix and $\mathrm{diag}(\boldsymbol{w})_{ii} = w_i$. If $\mathrm{diag}(\boldsymbol{w}) = \frac{1}{\sqrt{n}}\boldsymbol{I}$, then L_0-SWCCA reduces to L_0-SCCA.

III. Optimization

In this section, we design an alternating iterative algorithm to solve Eq.(6) by using a sparse projection strategy. We start with the sparse projection problem corresponding to the sub-problem of Eq.(6) with fixed $\boldsymbol{v}$ and $\boldsymbol{w}$ as Eq.(7).

$$\begin{aligned} \max_{\boldsymbol{u}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1, \|\boldsymbol{u}\|_0 \le k \end{aligned} \tag{7}$$

For a given column vector $\boldsymbol{z} \in \mathbb{R}^{p\times 1}$ and $k \le p$, we define a sparse project operator $\Pi(\cdot, k)$ as Eq.(8).

$$[\Pi(\boldsymbol{z}, k)]_i = \begin{cases} z_i, & \text{if } i \in \mathrm{support}(\boldsymbol{z}, k) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathrm{support}(\boldsymbol{z}, k)$ is defined as a set of indexes corresponding to the largest k absolute values of $\boldsymbol{z}$. For example, if $\boldsymbol{z} = [-5, 3, 5, 2, -1]^{\mathrm{T}}$, then $\Pi(\boldsymbol{z}, 3) = [-5, 3, 5, 0, 0]^{\mathrm{T}}$.

Theorem 1 The solution of problem (7) is

$$\boldsymbol{u}^* = \frac{\Pi(\boldsymbol{z}, k)}{\|\Pi(\boldsymbol{z}, k)\|_2} \tag{9}$$

Note that $\|\cdot\|_2$ denotes the Euclidean norm. We can prove the Theorem 1 by contradiction (we omit the proof here). Based on Theorem 1, we design an alternating iterative approach to solve Eq.(6).

i) Optimizing $\boldsymbol{u}$ with fixed $\boldsymbol{v}$ and $\boldsymbol{w}$. Fix $\boldsymbol{v}$ and $\boldsymbol{w}$ in Eq.(6), let $\boldsymbol{z}_{\boldsymbol{u}} = \boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v}$, then Eq.(6) reduces to as Eq.(10).

$$\begin{aligned} \max_{\boldsymbol{u}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_{\boldsymbol{u}} \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1, \|\boldsymbol{u}\|_0 \le k_{\boldsymbol{u}} \end{aligned} \tag{10}$$

Based on the Theorem 1, we obtain the update rule of $\boldsymbol{u}$ as Eq.(11).

$$\boldsymbol{u} \leftarrow \frac{\Pi(\boldsymbol{z}_{\boldsymbol{u}}, k_{\boldsymbol{u}})}{\|\Pi(\boldsymbol{z}_{\boldsymbol{u}}, k_{\boldsymbol{u}})\|_2} \tag{11}$$

ii) Optimizing $\boldsymbol{v}$ with fixed $\boldsymbol{u}$ and $\boldsymbol{w}$. Fix $\boldsymbol{u}$ and $\boldsymbol{w}$ in Eq.(6), let $\boldsymbol{z}_{\boldsymbol{v}} = \boldsymbol{Y}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{X}\boldsymbol{u}$, then Eq.(6) reduces to as Eq.(12).

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & \boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_{\boldsymbol{v}} \\ \text{s.t.} \quad & \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1, \|\boldsymbol{v}\|_0 \le k_{\boldsymbol{v}} \end{aligned} \tag{12}$$

Similarly, we obtain the update rule of $\boldsymbol{v}$ as Eq.(13).

$$\boldsymbol{v} \leftarrow \frac{\Pi(\boldsymbol{z}_{\boldsymbol{v}}, k_{\boldsymbol{v}})}{\|\Pi(\boldsymbol{z}_{\boldsymbol{v}}, k_{\boldsymbol{v}})\|_2} \tag{13}$$

iii) Optimizing $\boldsymbol{w}$ with fixed $\boldsymbol{u}$ and $\boldsymbol{v}$. Fix $\boldsymbol{u}$ and $\boldsymbol{v}$ in Eq.(6), then Eq.(6) reduces to as Eq.(14).

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1, \|\boldsymbol{w}\|_0 \le k_{\boldsymbol{u}} \end{aligned} \tag{14}$$

Let $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{u}$, $\boldsymbol{t}_2 = \boldsymbol{Y}\boldsymbol{u}$ and $\boldsymbol{z}_{\boldsymbol{w}} = \boldsymbol{t}_1 \odot \boldsymbol{t}_2$ where '$\odot$' denotes point multiplication which is equivalent to '.*' in Matlab, then we have $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} = \boldsymbol{t}_1\mathrm{diag}(\boldsymbol{w})\boldsymbol{t}_2 = (\boldsymbol{t}_1 \odot \boldsymbol{w})^{\mathrm{T}}\boldsymbol{t}_2 = \boldsymbol{w}^{\mathrm{T}}(\boldsymbol{t}_1 \odot \boldsymbol{t}_2) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_{\boldsymbol{w}}$. Thus, problem (14) reduces to as Eq.(15).

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_{\boldsymbol{w}} \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1, \|\boldsymbol{w}\|_0 \le k_{\boldsymbol{w}} \end{aligned} \tag{15}$$

Similarly, we obtain the update rule of $\boldsymbol{w}$ as Eq.(16).

$$\boldsymbol{w} \leftarrow \frac{\Pi(\boldsymbol{z}_{\boldsymbol{w}}, k_{\boldsymbol{w}})}{\|\Pi(\boldsymbol{z}_{\boldsymbol{w}}, k_{\boldsymbol{w}})\|_2} \tag{16}$$

Finally, combining Eqs.(11), (13) and (16), we propose the following alternating iterative algorithm to solve problem (6) as Algorithm 1.

Terminating condition We can set different stop conditions to control the iterations. For example, the update length of $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ are smaller than a given threshold (*i.e.*, $\|\boldsymbol{u}^k - \boldsymbol{u}^{k+1}\|_2^2 < 10^{-6}$, $\|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_2^2 < 10^{-6}$ and $\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_2^2 < 10^{-6}$), or the maximum number of iterations is a given number (*e.g.*, 1000), or the change of objective value is less than a give threshold.

Algorithm 1 L_0-SWCCA

Require: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$, $k_{\boldsymbol{u}}$, $k_{\boldsymbol{v}}$ and $k_{\boldsymbol{w}}$;
Ensure: $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$;
  initial $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$;
  repeat
    Update $\boldsymbol{u}$ according to Eq.(11)
    Update $\boldsymbol{v}$ according to Eq.(13)
    Update $\boldsymbol{w}$ according to Eq.(16)
until convergence of $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$.

Computation complexity The complexity of matrix multiplication with one $n \times p$ matrix and another one $n \times q$ is $\mathcal{O}(npq)$. To reduce the computational complexity of $\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v}$, we note that $\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} = \boldsymbol{X}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v}) = \boldsymbol{X}^{\mathrm{T}}[(\boldsymbol{Y}\boldsymbol{v}) \odot \boldsymbol{w}]$. Let $\boldsymbol{t}_1 = \boldsymbol{Y}\boldsymbol{v}$, $\boldsymbol{t}_2 = \boldsymbol{t}_1 \odot \boldsymbol{w}$ and $\boldsymbol{t}_3 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{t}_2$. Thus, the complexity of $\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v}$ is $\mathcal{O}(nq + n + np)$. Similarly, we can see that the complexity of $\boldsymbol{Y}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{X}\boldsymbol{u}$ is $\mathcal{O}(np + n + nq)$, and the complexity of $(\boldsymbol{X}\boldsymbol{u}) \odot (\boldsymbol{Y}\boldsymbol{v})$ is $\mathcal{O}(nq + np + n)$. In Algorithm 1, we need to obtain the largest k absolute values of a given vector $\boldsymbol{z}$ of size $p \times 1$ [*i.e.*, $\Pi(\boldsymbol{z}, k)$]. We adopt a linear time selection algorithm called Quick select (QS) algorithm to compute $\Pi(\boldsymbol{z}, k)$, which applies a divide and conquer strategy, and the average time complexity of QS algorithm is $\mathcal{O}(p)$. Thus, the entire time complexity of Algorithm 1 is $\mathcal{O}(Tnp + Tnq)$, where T is the number of iterations for convergence. In general, T is a small number.

Convergence analysis Similar with Theorem 1 in Ref.[26], we can prove that Algorithm 1 converges globally to a critical point (we omit the proof here).

IV. Experiments

1. Synthetic data 1

Here we generate the first synthetic data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $n = 50$, $p = 100$ and $q = 80$ using the following two steps:

Step 1: Generate two canonical vectors $\boldsymbol{u}$, $\boldsymbol{v}$ and a weighted vector $\boldsymbol{w}$ as Eq.(17).

$$\begin{aligned} \boldsymbol{u} &= [r(1, 30), r(0, 70)]^{\mathrm{T}} \\ \boldsymbol{v} &= [N(20), r(0, 20), N(10), r(0, 30)]^{\mathrm{T}} \\ \boldsymbol{w} &= [r(1, 30), r(0, 20)]^{\mathrm{T}} \end{aligned} \tag{17}$$

where $r(a, n)$ denotes a row vector of size n, whose elements are equal to a, $N(m)$ denotes a row vector of size m, whose elements are randomly sampled from a standard normal distribution.

Step 2: Generate two input matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ as Eq.(18).

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{w}\boldsymbol{u}^{\mathrm{T}} + \boldsymbol{\epsilon}_x \\ \boldsymbol{Y} &= \boldsymbol{w}\boldsymbol{v}^{\mathrm{T}} + \boldsymbol{\epsilon}_y \end{aligned} \tag{18}$$

where the elements of $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ are randomly sampled from a standard normal distribution.

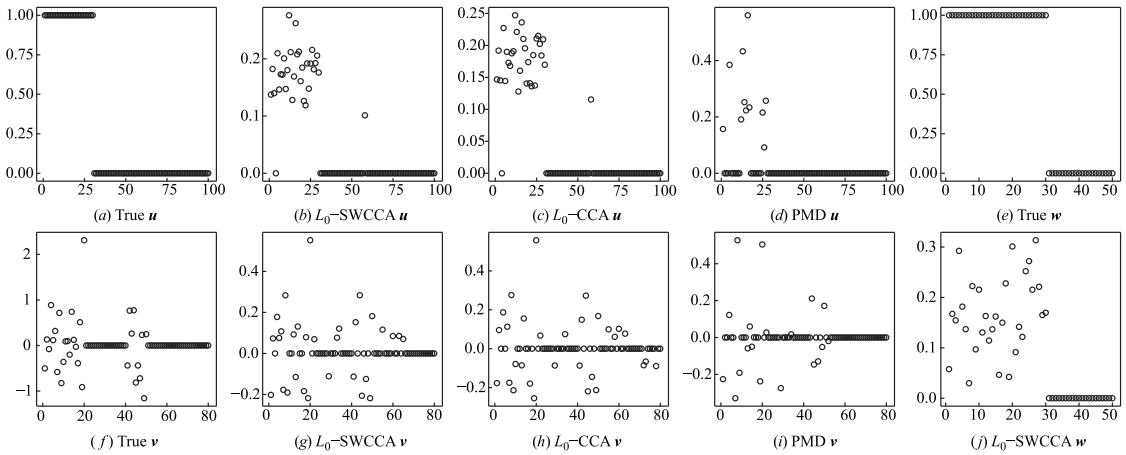

Fig. 2. Results of the synthetic data 1. (a), (f) and (e) denote true $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$; (g), (d) and (j) denote estimated $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ by L_0-SWCCA; (c) and (h) denote estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by L_0-SCCA; (d) and (i) denote stimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by PMD

We evaluate the performance of L_0-SWCCA with the above synthetic data and compare its performance with the typical sparse CCA, including L_0-SCCA[13] and PMD[6] with L_1-penalty. For comparison, we set parameters $k_u = 30$, $k_v = 30$ and $k_w = 30$ for L_0-SWCCA; $k_u = 30$, $k_v = 30$ for L_0-SCCA; $c_1 = \dfrac{30}{100}\sqrt{p}$ and $c_2 = \dfrac{30}{80}\sqrt{q}$ for PMD. Note that $c_1 = c\sqrt{p}$ and $c_2 = c\sqrt{q}$ where $c \in (0, 1)$ for PMD are to approximately control the sparse proportion of the canonical vectors ($\boldsymbol{u}$ and $\boldsymbol{v}$).

The true and estimated patterns for $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ in the synthetic data 1 are shown in Fig.2 and the heatmaps of reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ by different methods are shown in Fig.3. Compared to PMD, L_0-SWCCA and L_0-SCCA does fairly well for identifying the local non-zero pattern of the underlying factors (*i.e.*, $\boldsymbol{u}$ and $\boldsymbol{v}$). However, the two traditional SCCA methods (L_0-SCCA and PMD) do not recognize the difference between samples and remove the noisy samples (Fig.3(c) and (d)). Interestingly, L_0-SWCCA not only discovers the true patterns for $\boldsymbol{u}$, $\boldsymbol{v}$ (Fig.2(b) and (g)), but also identifies the true non-zero characteristics of samples ($\boldsymbol{w}$) (Fig.2(e), and Fig.3(j)). Furthermore, to assess our approach is indeed able to find a greater correlation level between two input matrices, we define the correlation criterion as Eq.(19).

$$\rho = \mathrm{cor}((\boldsymbol{X}\boldsymbol{u}) \odot \boldsymbol{w}, (\boldsymbol{Y}\boldsymbol{v}) \odot \boldsymbol{w}) \tag{19}$$

where $\mathrm{cor}(\cdot)$ is a function to calculate the correlation coefficient of the two vectors. For comparison, we set $\boldsymbol{w} = [1, \cdots, 1]^{\mathrm{T}}$ for L_0-SCCA and PMD to compute the correlation criterion. L_0-SWCCA gets the largest $\rho = 0.96$ compared to L_0-SCCA with $\rho = 0.80$ and PMD with $\rho = 0.87$ in the above synthetic data. All results show that L_0-SWCCA is more effective to capture the latent patterns of canonical vectors than other methods.

2. Synthetic data 2

Here we apply another way to generate synthetic data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$ with $n = 50$, $p = 100$, and $q = 80$. The following three steps are used to generate the second synthetic data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$.

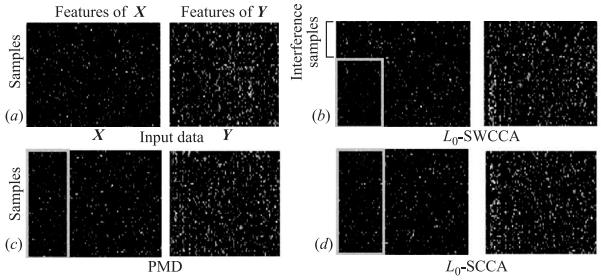

Fig. 3. (a) Heatmaps of the original input matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the synthetic data 1; (b) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ by L_0-SWCCA; (c) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by PMD; (d) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by L_0-SCCA. Note that because the true signal of $\boldsymbol{v}$ in Eq.(17) is relatively weak, resulting in the pattern of $\boldsymbol{Y}$ is not very clear

Step 1: We first generate two zero matrices as Eq.(20).

$$\begin{aligned} \boldsymbol{X} &\leftarrow matrix(0, nrow = n, ncol = p) \\ \boldsymbol{Y} &\leftarrow matrix(0, nrow = n, ncol = q) \end{aligned} \tag{20}$$

Step 2: We then update two sub-matrices in the $\boldsymbol{X}$ and $\boldsymbol{Y}$.

$$\begin{aligned} \boldsymbol{X}[1:30, 1:50] &\leftarrow 1 \\ \boldsymbol{Y}[1:30, 1:40] &\leftarrow -1 \end{aligned} \tag{21}$$

Step 3: We add the Gaussian noise in $\boldsymbol{X}$ and $\boldsymbol{Y}$.

$$\begin{aligned} \boldsymbol{X} &\leftarrow \boldsymbol{X} + \boldsymbol{\epsilon}_x \\ \boldsymbol{Y} &\leftarrow \boldsymbol{Y} + \boldsymbol{\epsilon}_y \end{aligned} \tag{22}$$

For simplicity and comparison, we can set true $\boldsymbol{u} = [r(1, 50),\ r(0, 50)]^{\mathrm{T}}$, true $\boldsymbol{v} = [r(-1, 40),\ r(0, 40)]^{\mathrm{T}}$ and true $\boldsymbol{w} = [r(1, 30), r(0, 20)]^{\mathrm{T}}$ to characterize the patterns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (Fig.4(a), (f) and (e)). Similarly, we also apply L_0-SWCCA, L_0-SCCA[13] and PMD[6] to the synthetic data 2. For comparison, we set parameters $k_u = 50$,

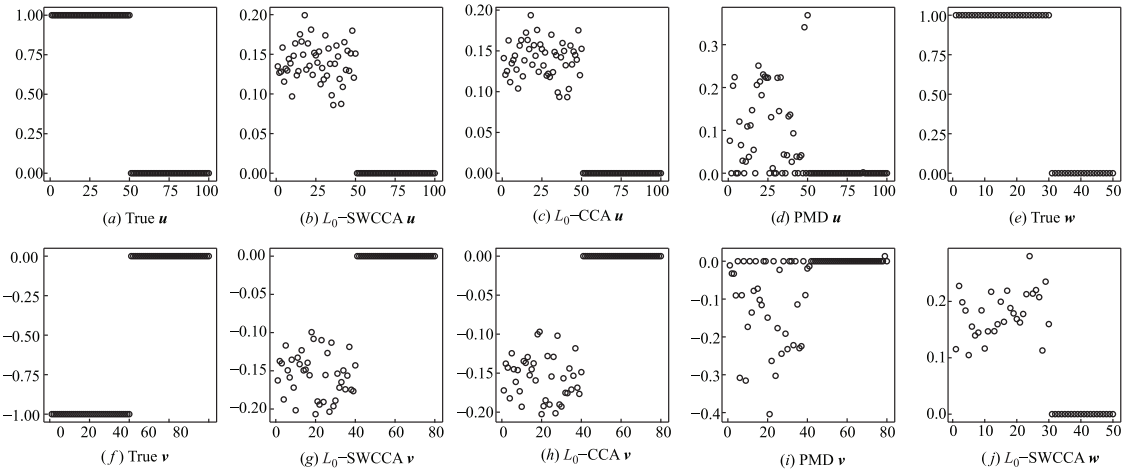

Fig. 4. Results of the synthetic data 2. (a), (f) and (e) denote true $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$; (b), (g) and (j) denote estimated $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ by L_0-SWCCA; (c) and (h) denote estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by L_0-SCCA; (d) and (i) denote estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by PMD

$k_v = 40$ and $k_w = 30$ for L_0-SWCCA; $k_u = 50$, $k_v = 40$ for L_0-SCCA; $c_1 = \dfrac{50}{100}\sqrt{p}$ and $c_2 = \dfrac{40}{80}\sqrt{q}$ for PMD.

The true and estimated patterns for $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ are shown in Fig.4 and the corresponding to heatmaps of reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ by different methods are shown in Fig.5. L_0-SWCCA and L_0-SCCA are superior to PMD about identifying the latent patterns of canonical vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (Fig.4(b), (c), (d), (g), (h) and (i)). However L_0-SCCA and PMD fail to remove interference samples (Fig.5(c) and (d)). Compared to L_0-SCCA and PMD, L_0-SWCCA can clearly identify the true non-zero characteristics of samples (Fig.4(e) and Fig.5(j)). Similarly, we also compute the correlation criterion based on Eq.(19). We find that L_0-SWCCA gets the largest correlation $\rho = 0.97$ compared to L_0-SCCA with $\rho = 0.93$ and PMD with $\rho = 0.95$. All results show that our method is more effective to capture the latent patterns of canonical vectors than other ones.

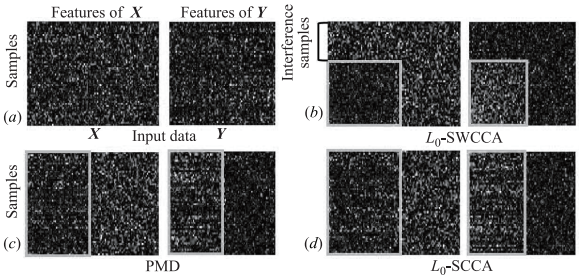

Fig. 5. (a) Heatmaps of the original input matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the synthetic data 2; (b) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ by L_0-SWCCA; (c) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by PMD; (d) Reordered $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the estimated $\boldsymbol{u}$ and $\boldsymbol{v}$ by L_0-SCCA

3. Breast cancer data

We first consider a breast cancer dataset[6,27] consisting of gene expression and DNA copy number variation data across 89 cancer samples. Specifically, the gene expression data $\boldsymbol{X}$ and the DNA copy number data $\boldsymbol{Y}$ are of size $n \times p$ and $n \times q$ with $n = 89$, $p = 19672$ and $q = 2149$. We apply SWCCA and related ones to this data to identify a gene set whose expression is strongly correlated with copy number changes of some genomic regions.

In PMD[6], we set $c_1 = c\sqrt{p}$ and $c_2 = c\sqrt{q}$, where $c \in (0, 1)$ is to approximately control the sparse ratio of canonical vectors. We ensure that the canonical vectors ($\boldsymbol{u}$ and $\boldsymbol{v}$) extracted by the three methods (PMD, L_0-SCCA, and L_0-SWCCA) have the same sparsity level for comparison. We first apply PMD in the breast cancer data to obtain two sparse canonical vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ for each given $c \in (0, 1)$. Then, we compute the number of nonzero elements in the above extracted $\boldsymbol{u}$ and $\boldsymbol{v}$, denoted as $N_{\boldsymbol{u}}$ and $N_{\boldsymbol{v}}$. Finally, we set $k_{\boldsymbol{u}} = N_{\boldsymbol{u}}$, $k_{\boldsymbol{v}} = N_{\boldsymbol{v}}$ in L_0-SCCA and L_0-SWCCA, and set $k_{\boldsymbol{w}} = 53 \approx 0.6 \times 89$ in L_0-SWCCA to identify the sample loading $\boldsymbol{w}$ with sparse ratio 60%.

We adopt two criteria: correlation level defined in Eq.(19) and objective value defined in Eq.(5) for comparison. Here we consider different c values ($i.e.$, $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7$) to control the different sparse ratio of canonical vectors. We find that, compared to PMD and L_0-SCCA, L_0-SWCCA does obtain higher correlation level and objective value for all cases (Table 1). Since the 'breast cancer data' did not collect any clinical information of patients, it is very difficult to study the specific characteristics of these selected samples. To this end, we also apply our method to another biological data with more detailed clinical information.

Table 1. Results on Correlation level (CL) and Objective value (OV) for different c values.

CL	c=0.1	c=0.2	c=0.3	c=0.4	c=0.5	c=0.6	c=0.7
PMD	0.77	0.81	0.85	0.86	0.86	0.85	0.83
L_0-SCCA	0.88	0.84	0.86	0.85	0.84	0.82	0.81
L_0-SWCCA	0.99	0.98	1.00	1.00	0.99	0.97	0.99
OV	c=0.1	c=0.2	c=0.3	c=0.4	c=0.5	c=0.6	c=0.7
PMD	253	768	1390	2020	2589	3056	3392
L_0-SCCA	371	1066	1824	2467	3014	3360	3520
L_0-SWCCA	1570	3046	4960	6576	7860	10375	10692

4. TCGA BLCA data

Recently, it is a hot topic to study microRNA (miRNA) and gene regulatory relationship from matched miRNA and gene expression data[25,28]. Here, we apply SWCCA onto the Bladder urothelial carcinoma (BLCA) miRNA and gene expression data across 405 patients from TCGA (*https://cancergenome.nih.gov/*) to identify a subtype-specific miRNA-gene co-correlation module. To remove some noise miRNAs and genes, we first adapt standard deviation method to extract 200 largest variance miRNAs and 5000 largest variance genes for further analysis. Finally, we obtain a miRNA expression matrix $\boldsymbol{X} \in \mathbb{R}^{405\times 200}$, which is standardized for each miRNA, and a gene expression matrix $\boldsymbol{Y} \in \mathbb{R}^{405\times 5000}$, which is standardized for ecah gene. We apply L_0-SWCCA onto BLCA data with $k_{\boldsymbol{u}} = 10$, $k_{\boldsymbol{v}} = 200$ and $k_{\boldsymbol{w}} = 203$ to identify a miRNA set with 10 miRNAs and a gene set with 200 genes and a sample set with 203 patients. We also apply PMD with $c_1 = (10/200)\sqrt{p}$, $c_2 = (200/5000)\sqrt{q}$ and L_0-SCCA with $k_{\boldsymbol{u}} = 10$ and $k_{\boldsymbol{v}} = 200$ onto BLCA data for comparison.

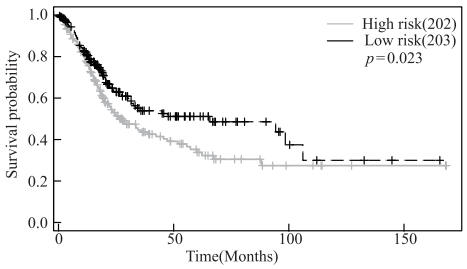

Fig. 6. Kaplan-Meier survival analysis between the selected patients and the remaining patents based on the $\boldsymbol{w}$ estimated by L_0-SWCCA. P-value is calculated by log-rank test

Similarly, L_0-SWCCA obtains the largest Correlation level and Objective value than others ones [(CL, OV): (0.98, 1210) for L_0-SWCCA, (0.84, 346) for PMD, (0.86,469) for L_0-SCCA], respectively. More importantly, we also analyze characteristics of these selected patients by L_0-SWCCA. We find that it is significantly different with respect to patient survival time between the selected 203 patients and the remaining 202 patients with p-value = 0.023 (Fig.6). These results imply that L_0-SWCCA can be used to discover BLCA subtype-specific miNRA-gene co-correlation modules.

Furthermore, we also assess whether these identified genes by L_0-SWCCA are biologically relevant with BLCA. DAVID (*https://david.ncifcrf.gov/*) is used to perform the Gene ontology (GO) Biological processes (BPs) and KEGG pathways enrichment analysis. Several significantly enriched GO BPs and pathways relating to BLCA are discovered including *GO:0008544:epidermis development* (B-H adjusted p-value = 1.1E-12), *hsa00591:Linoleic acid metabolism* (B-H adjusted p-value = 4.8E-3), *hsa00590:Arachidonic acid metabolism* (B-H adjusted p-value = 3.6E-3) and *hsa00601:Glycosphingolipid biosynthesis-lacto and neolacto series* (B-H adjusted p-value = 2.6E-2). Finally, we also examine whether the identified miRNAs by L_0-SWCCA are associated with BLCA. Interestingly, in the identified 10 miRNAs by L_0-SWCCA, we find that there are six miRNAs (including *hsa-miR-200a-3p, hsa-miR-200b-5p, hsa-miR-200b-3p, hsa-miR-200a-5p, hsa-miR-200c-3p* and *hsa-miR-200c-5p*) belonging to miR-200 family. Notably, several studies[29,30] have also reported miR-200 family plays key roles in BLCA. All these results imply that the identified miNRA-gene module by L_0-SWCCA may help us to find new therapeutic strategy for BLCA.

V. Extensions

1. SWCCA with generalized penalties

We first consider a general regularized SWCCA framework as Eq.(23).

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v} \\ & -\mathcal{R}_{\boldsymbol{u}}(\boldsymbol{u}) - \mathcal{R}_{\boldsymbol{v}}(\boldsymbol{v}) - \mathcal{R}_{\boldsymbol{w}}(\boldsymbol{w}) \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1 \end{aligned} \tag{23}$$

where $\mathcal{R}_{\boldsymbol{u}}(\cdot)$, $\mathcal{R}_{\boldsymbol{v}}(\cdot)$, $\mathcal{R}_{\boldsymbol{w}}(\cdot)$ are three regularized functions. For different prior knowledge, we can use different sparsity inducing penalties.

1) LASSO regularized SWCCA. If $\mathcal{R}_{\boldsymbol{u}}(\boldsymbol{u}) = \lambda_{\boldsymbol{u}}\|\boldsymbol{u}\|_1$, $\mathcal{R}_{\boldsymbol{v}}(\boldsymbol{v}) = \lambda_{\boldsymbol{v}}\|\boldsymbol{v}\|_1$ and $\mathcal{R}_{\boldsymbol{w}}(\boldsymbol{w}) = \lambda_{\boldsymbol{w}}\|\boldsymbol{w}\|_1$. We obtain a L_1 (Lasso) regularized SWCCA (L_1-SWCCA). Similar to solve Eq.(6), we only need to solve the following problem to solve L_1-SWCCA as Eq.(24).

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} + \lambda_{\boldsymbol{u}}\|\boldsymbol{u}\|_1 \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1 \end{aligned} \tag{24}$$

where $\boldsymbol{z} = \boldsymbol{X}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w})\boldsymbol{Y}\boldsymbol{v}$. We first replace the constraint $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1$ with $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} \le 1$ and obtain the following the problem as Eq.(25).

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} + \lambda_{\boldsymbol{u}}\|\boldsymbol{u}\|_1 \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} \le 1 \end{aligned} \tag{25}$$

It is easy to see that problem (25) is equivalent to (24). Thus, we can obtain its Lagrangian form as Eq.(26).

$$\mathcal{L}(\boldsymbol{u}, \lambda_{\boldsymbol{u}}, \eta_{\boldsymbol{u}}) = -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} + \lambda_{\boldsymbol{u}}\|\boldsymbol{u}\|_1 + \eta_{\boldsymbol{u}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} - 1) \tag{26}$$

Thus, we can use a coordinate descent method[31] to minimize Eq.(26) and obtain the following update rule of $\boldsymbol{u}$ as Eq.(27).

$$\boldsymbol{u} = \frac{\mathcal{S}_{\lambda_{\boldsymbol{u}}}(\boldsymbol{z})}{\|\mathcal{S}_{\lambda_{\boldsymbol{u}}}(\boldsymbol{z})\|_2} \tag{27}$$

where $\mathcal{S}_{\lambda_{\boldsymbol{u}}}(\cdot)$ is a soft thresholding operator and $\mathcal{S}_{\lambda_{\boldsymbol{u}}}(z_i) = \mathrm{sign}(|z_i| - \lambda_{\boldsymbol{u}})_+$. Based on the above, an alternating iterative strategy can be used to solve L_1-SWCCA.

2) Group LASSO regularized SWCCA. If $\mathcal{R}_{\boldsymbol{u}}(\boldsymbol{u}) = \lambda_{\boldsymbol{u}} \sum_l \|\boldsymbol{u}^{(l)}\|_2$, $\mathcal{R}_{\boldsymbol{v}}(\boldsymbol{v}) = \lambda_{\boldsymbol{v}} \sum_l \|\boldsymbol{v}^{(l)}\|_2$ and $\mathcal{R}_{\boldsymbol{w}}(\boldsymbol{w}) = \lambda_{\boldsymbol{w}} \sum_l \|\boldsymbol{w}^{(l)}\|_2$. Problem (24) reduces to $L_{2,1}$-regularized SWCCA ($L_{2,1}$-SWCCA). Similarly, we should solve the following projection problem as Eq.(28).

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} + \lambda_{\boldsymbol{u}} \sum_l \|\boldsymbol{u}^{(l)}\|_2 \\ \text{s.t.} \quad & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} \le 1 \end{aligned} \tag{28}$$

Thus, we obtain its Lagrangian form as Eq.(29).

$$\mathcal{L}(\boldsymbol{u}, \lambda_{\boldsymbol{u}}, \eta_{\boldsymbol{u}}) = -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z} + \lambda_{\boldsymbol{u}} \sum_l \|\boldsymbol{u}^{(l)}\|_2 + \eta_{\boldsymbol{u}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} - 1) \tag{29}$$

where $\boldsymbol{u}^{(l)}$ is the lth group of $\boldsymbol{u}$. We adopt a block-coordinate descent method[31] to solve it and obtain the learning rule of $\boldsymbol{u}^{(l)}$ $(l = 1, \cdots, L)$ as Eq.(30).

$$\boldsymbol{u}^{(l)} = \begin{cases} \dfrac{1}{2\eta_{\boldsymbol{u}}} \boldsymbol{z}^{(l)} \left(1 - \dfrac{\lambda_{\boldsymbol{u}}}{\|\boldsymbol{z}^{(l)}\|_2}\right), & \text{if } \|\boldsymbol{z}^{(l)}\|_2 > \lambda_{\boldsymbol{u}} \\ \boldsymbol{0}, & \text{otherwise} \end{cases} \tag{30}$$

By cyclically applying the above updates, we can minimize Eq.(29). Thus, an alternating iterative strategy can be used to solve $L_{2,1}$-SWCCA.

2. Multi-view sparse weighted CCA

In various scientific fields, multiple view data (more than two views) can be available from multiple sources or diverse feature subsets. For example, multiple high-throughput molecular profiling data by omics technologies can be produced for the same individuals in bioinformatics[25,26]. Integrating these data together can significantly increase the power of pattern discovery and individual classification. Here we extend SWCCA to Multi-view SWCCA (MSWCCA) model for multi-view data analysis (Fig.7) as follows:

$$\begin{aligned} \max_{\boldsymbol{u}_i, \boldsymbol{w}} \quad & \boldsymbol{w}^{\mathrm{T}} \Big[\bigodot_{i=1}^{M} (\boldsymbol{X}_i \boldsymbol{u}_i) \Big] - \sum_{i=1}^{M} \mathcal{R}_{\boldsymbol{u}_i}(\boldsymbol{u}_i) - \mathcal{R}_{\boldsymbol{w}}(\boldsymbol{w}) \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = 1, \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_i = 1 \text{ for } i = 1, \cdots, M \end{aligned}$$

where $\bigodot_{i=1}^{M} (\boldsymbol{X}_i \boldsymbol{u}_i) = (\boldsymbol{X}_1 \boldsymbol{u}_1) \odot (\boldsymbol{X}_2 \boldsymbol{u}_2) \cdots \odot (\boldsymbol{X}_M \boldsymbol{u}_M)$. When $M = 2$, we can see that $\boldsymbol{w}^{\mathrm{T}}[(\boldsymbol{X}_1 \boldsymbol{u}_1) \odot (\boldsymbol{X}_2 \boldsymbol{u}_2)] = \boldsymbol{u}_1^{\mathrm{T}} \boldsymbol{X}_1^{\mathrm{T}} \mathrm{diag}(\boldsymbol{w}) \boldsymbol{X}_2 \boldsymbol{u}_1$ and it reduces to SWCCA. So we can solve MSWCCA in a similar manner with SWCCA.

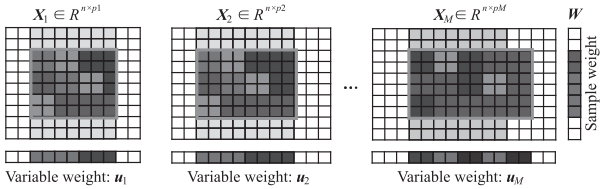

Fig. 7. Illustration of the multi-view sparse weighted CCA designed for integrating multiple data matrices

VI. Conclusions

In this paper, we propose a sparse weighted CCA framework. Compared to SCCA, SWCCA can reveal that the selected variables are only strongly related to a subset of samples. We develop an efficient alternating iterative algorithm to solve the L_0-regularized SWCCA. Our tests using both simulation and biological data show that SWCCA can obtain more reasonable patterns compared to the typical SCCA. Moreover, the key idea of SWCCA is easy to be adapted by other penalties like LASSO and Group LASSO. Lastly, we extend SWCCA to MSWCCA for multi-view situation with multiple data matrices.

References

[1] A. Klami, S. Virtanen and S. Kaski,"Bayesian canonical correlation analysis", *J. Mach. Learn. Res.*, Vol.14, pp.965–1003, 2013.

[2] L. Sun, S. Ji and J. Ye, "A least squares formulation for canonical correlation analysis", *International Conference on Machine Learning (ICML)*, Helsinki, Finland, pp.1024–1031, 2008.

[3] X. Yang, W. Liu, D. Tao, *et al.*, "Canonical correlation analysis networks for two-view image recognition", *Information Sciences*, Vol.385-386, pp.338–352, 2017.

[4] J. Cai, Y. Tang and J. Wang, "Kernel canonical correlation analysis via gradient descent", *Neurocomputing*, Vol.182, pp.322–331, 2016.

[5] C. Wang, J. Liu, W. Min, *et al.*, "A novel sparse penalty for singular value decomposition", *Chinese Journal of Electronics*, Vol.26, No.2, pp.306–312, 2017.

[6] D.M. Witten, R.J. Tibshirani and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis", *Biostatistics*, Vol.10, No.3, pp.515–534, 2009.

[7] S. Mizutani, E. Pauwels, V. Stoven, *et al.*, "Relating drug-protein interaction network with drug side effects", *Bioinformatics*, Vol.28, No.18, pp.i522–i528, 2012.

[8] K.A. Lé Cao, P.G. Martin, C. Robert-Grani, *et al.*, "Sparse canonical methods for biological data integration: Application to a cross-platform study", *BMC Bioinformatics*, Vol.10, Article ID 34, 17 pages, 2009.

[9] J. Fang, D. Lin, S.C. Schulz, *et al.*, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules", *Bioinformatics*, Vol.32, No.22, pp.3480–3488, 2016.

[10] Y. Kosuke, Y. Junichiro and D. Kenji, "Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data", *BMC Bioinformatics*, Vol.18, Article ID 108, 11 pages, 2017.

[11] E. Parkhomenko, D. Tritchler and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration", *Stat. Appl. Genet. Mol. Biol.*, Vol.8, No.1, pp.1–34, 2009.

[12] D.M. Witten and R.J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data", *Stat. Appl. Genet. Mol. Biol.*, Vol.10, No.3, pp.515–534, 2009.

[13] M. Asteris, A. Kyrillidis, O. Koyejo, *et al.*, "A simple and provable algorithm for sparse diagonal CCA", *International Conference on Machine Learning* (*ICML*), New York, NY, USA, pp.1148–1157, 2016.

[14] D. Chu, L.Z. Liao, M.K. Ng, *et al.*, "Sparse canonical correlation analysis: New formulation and algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.12, pp.3050–3065, 2013.

[15] D.R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis", *Mach. Learn.*, Vol.83, No.3, pp.331–353, 2011.

[16] C. Gao, Z. Ma, Z. Ren, *et al.*, "Minimax estimation in sparse canonical correlation analysis", *The Annals of Statistics*, Vol.43, No.5, pp.2168–2197, 2015.

[17] D. Lin, J. Zhang, J. Li, *et al.*, "Group sparse canonical correlation analysis for genomic data integration", *BMC Bioinformatics*, Vol.14, Article ID 245, 16 pages, 2013.

[18] S. Virtanen, A. Klami, and S. Kaski, "Bayesian CCA via group sparsity", *International Conference on Machine Learning* (*ICML*), Bellevue, WA, USA, pp.457–464, 2011.

[19] J. Chen, F.D. Bushman, J.D. Lewis, *et al.*, "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis", *Biostatistics*, Vol.14, No.2, pp.244–258, 2013.

[20] X. Chen, H. Liu, and J.G. Carbonell, "Structured sparse canonical correlation analysis", *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), La Palma, Canary Islands, pp.199–207, 2012.

[21] L. Du, H. Huang, J. Yan, *et al.*, "Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method", *Bioinformatics*, Vol.32, No.10, pp.1544–1551, 2016.

[22] L. Jacob, G. Obozinski and J.P. Vert, "Group lasso with overlap and graph lasso", *International Conference on Machine Learning* (*ICML*), Montreal, Canada, pp.433–440, 2009.

[23] X. Dai, T. Li, Z. Bai, *et al.*, "Breast cancer intrinsic subtype classification, clinical use and future trends", *American Journal of Cancer Research*, Vol.5, No.10, pp.2929–2943, 2015.

[24] J. Chen and S. Zhang, "Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data", *Bioinformatics*, Vol.32, No.11, pp.1724–1732, 2016.

[25] W. Min, J. Liu, F. Luo, *et al.*, "A novel two-stage method for identifying microRNA-gene regulatory modules in breast cancer", *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Washington, D.C., USA, pp.151–156, 2015.

[26] J. Sun, J. Lu, T. Xu *et al.*, "Multi-view sparse co-clustering via proximal alternating linearized minimization", *International Conference on Machine Learning* (*ICML*), Lille, France, pp.757–766, 2015

[27] K. Chin, S. Devries, J. Fridlyand, *et al.*, "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies", *Cancer Cell*, Vol.10, No.6, pp.529–541, 2006.

[28] S. Zhang, Q. Li and X.J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules", *Bioinformatics*, Vol.27, No.13, pp.i401–i409, 2011.

[29] Wiklund E.D. Bramsen, J.B. Bramsen, *et al.*, "Coordinated epigenetic repression of the miR-200 family and miR-205 in invasive bladder cancer", *International Journal of Cancer*, Vol.27, No.6, pp.1327–1334, 2011.

[30] Y. Cheng, X. Zhang, P. Li, *et al.*, "Mir-200c promotes bladder cancer cell migration and invasion by directly targeting recK", *OncoTargets and Therapy*, doi:10.2147/OTT.S101067, Vol.9, pp.5091–5099, 2016.

[31] P. Tseng, "Convergence of a block coordinate descent method for non-differentiable minimization", *Journal of Optimization Theory and Applications*, Vol.109, No.3, pp.475–494, 2001.

MIN Wenwen was born in 1988. He is currently a Ph.D. candidate of computer science in Wuhan University. His research interests include sparsity-constrained optimization, structured sparsity learning models and their applications in computational biology. (Email: minwenwen07@foxmail.com)

LIU Juan (corresponding author) was born in 1970. She received the Ph.D. degree in computer science from Wuhan University and now serves as a professor and Ph.D. supervisor in Wuhan University. Her research interests include data mining, nature language process and bioinformatics. (Email: liujuan@whu.edu.cn)

ZHANG Shihua was born in 1980. He is currently an associate professor of the Institute of Applied Mathematics, Academy of Mathematics and Systems Science, at CAS. His interests are within Bioinformatics, Computational Biology and Network Science, particularly in Cancer Genomics and Network Biology. (Email: zsh@amss.ac.cn)