

# Efficient subspace clustering based on self-representation and grouping effect

Shichao Zhang<sup>1</sup> · Yonggang Li<sup>1</sup> · Debo Cheng<sup>1</sup> · Zhenyun Deng<sup>1</sup> · Lifeng Yang<sup>1</sup>

Received: 30 December 2015 / Accepted: 13 May 2016 / Published online: 13 June 2016  
© The Natural Computing Applications Forum 2016

**Abstract** Traditional subspace clustering methods [such as sparse subspace clustering (SSC), least squares representation (LSR) and smooth representation clustering] either considered the grouping effect or the sparsity to group original data into clusters. This paper demonstrates the necessary of both the grouping effect and the sparsity for conducting subspace clustering, by proposing a new Self-Representation and *Subspace Clustering based on Grouping Effect (SRGE)* method. Specifically, first of all, a row sparse  $\ell_{2,1}$ -norm regularizer is utilized to represent each sample by other samples. Then, the grouping effect of the data is designed to ensure that the coefficient of close samples is similar, aiming at generating a diagonal block self-representation coefficient matrix. Finally, an affinity matrix is obtained for conducting spectral clustering. The proposed method can be regarded as a trade-off between SSC and LSR. The experimental results of segmentation on real datasets showed that the proposed method significantly outperformed the state-of-the-art methods in terms of all kinds of evaluation metrics.

**Keywords** Subspace clustering · Sparse · Self-representation · Affinity matrix

## 1 Introduction

In numerous aspects of machine learning, data mining and computer vision [1–3], data are usually high dimensional [4–6]. Moreover, a set of high-dimensional data is often drawn from multiple low-dimensional subspaces [7], such as the face images, the point trajectories of moving objects [8] and the texture features of pixel on an image [9–11]. Recently, subspace clustering [12] processes this kind of data by following their underlying subspaces to attract increasing attentions. A number of subspace clustering methods have thus been proposed. Roughly, according to the principle of representing the subspaces, the previous subspace clustering methods can be grouped into three categories, such as algebraic methods [13, 14], statistical methods [11] and spectral clustering-based methods [15–20].

The early studies of subspace clustering are mostly based on algebraic methods or statistical methods. Although exquisite formulations of algebraic methods for spectral clustering are used, their performance drops quickly in the datasets with noise or partially coupled subspaces, such as generalized principal component analysis (GPCA) [14]. By contrast, the statistical methods such as expectation maximization (EM) regard subspace clustering as a mixed data inference problem so that prevalent methods stem from general statistical learning domain can be used. Though many new techniques have been introduced to promote the criterion [e.g., agglomerative lossy compression (ALC)], the performance of statistics methods is restricted, due to its dependency on reckoning precise subspace models.

The representative spectral clustering methods include sparse subspace clustering (SSC), least squares representation (LSR) and smooth representation clustering (SRC) [15, 16, 18, 19]. The key step of these spectral clustering

---

✉ Shichao Zhang  
zhangsc@mailbox.gxnu.edu.cn

<sup>1</sup> Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, Guangxi, China

methods is to construct an affinity matrix by utilizing the global or the local information of samples. Unlike the previous works that computing affinity matrix from existing algebraic or statistical methods, the recent spectral clustering methods were put forward under the self-representation concept, i.e., representing each sample by a linear combination of other samples. More specifically, SSC could obtain a block-diagonal and also sparse affinity matrix when the subspaces are independent [21]. However, if the data from the same subspace are extremely relevant, SSC has to select only one at random. In this way, the spectral clustering method cannot firmly yield the proper groups, as the affinity matrix may be “too sparse” [22]. Therefore, though SSC could find a sparse affinity matrix, it ignores the relevant structures of the data in the same subspace and thus its performance is unsatisfied. By substituting the sparse representation with a low-rank representation and also considering the correlation structure of the data, LRR is designed to group the related data from the same subspace into together so that achieving a block-diagonal affinity matrix while the datasets are without interference and the subspaces are independent. However, the datasets in real life are always with outliers or noises, and thus, the subspaces are overlapping. While dealing with such data, LRR will output a dense and far from block-diagonal solution due to that the nuclear norm lacks of the ability to choose subspaces. Therefore, LRR could cluster the correlated data into together, resulting in the dense affinity matrix rather than a sparse one. The most recent works for subspace clustering are least squares regression (LSR) and SRC. Both of them encourage grouping effect [18, 19] to promote the accuracy of clustering models but lack of sparsity. In sum, the previous spectral clustering methods utilize the characterization of self-representation, such as sparse representation and low-rank representation (LRR), to improve their performance, but they still have more or less problems. This paper thus focuses on conducting spectral clustering for subspace clustering [23, 24].

In order to effectively cluster the data drawn from real life, a good subspace clustering model should take both grouping effect and sparsity into account. To do this, we propose a new method based on self-representation for subspace clustering by utilizing grouping effect and also an  $\ell_{2,1}$ -norm regularizer inducing row sparsity, named *Subspace Clustering based on Grouping Effect (SRGE)*. On the one hand, affinity matrix achieved by SRGE is sparse (i.e., block-diagonal), with less connection between clusters. On the other hand, it not be too sparse, i.e., the nonzero connection within cluster is sufficient enough for grouping correlated data that drawn from the same subspace [25]. In this way, the model can both group the correlated data

drawn from the same subspace (i.e., LSR and LRR) and reduce the connections between clusters (i.e., SSC).

In order to overcome the disadvantages of the algorithms mentioned above, we introduce self-representation-based  $\ell_{2,1}$ -norm to achieve a proper sparsity for the affinity matrix and also utilize the grouping effect to make sure the similar samples to be clustered into together rather than merely depending on the sparsity or low rank.

The contributions of this paper are summarized as follows:

1. With grouping effect in our subspace clustering model, it can be self-adaptive for different types of data since it takes the correlations of the data into account. The grouping effect makes sure SRGE groups the highly correlated data together. Moreover, if the subspaces are independent and the objective function satisfies the enforce block grouping effect (EBGE) conditions as well, the optimal solution of SRGE is block-diagonal and also has grouping effect. The term  $\text{tr}(\tilde{\mathbf{L}}\mathbf{Z}^T)$  is a special case that satisfies the EBGE conditions.
2. The SRGE takes the correlation of samples into account by the self-representation of samples, i.e., representing each sample by other samples rather than data pairs. Moreover, the robust loss function (i.e.,  $\|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_{2,1}$ ) is a balance between  $\ell_1$ -norm and F-norm and has been verified robust to noises [26].
3. Finally, as the objective function of our model is convex but not smooth, we utilize ADMM [27] to solve it efficiently.

The rest of this paper is organized as follows: In Sect. 2, we provide a brief review of the previous subspace clustering methods. Then, we propose the new grouping effect-based self-representation for subspace clustering in Sect. 3. The experimental results are presented in Sect. 4. Finally, we state the conclusions and our future work in Sect. 5.

## 2 Related work

### 2.1 Self-representation

Given a dataset  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbf{R}^{d \times n}$ , where  $\mathbf{x}_i \in \mathbf{R}^d$  is a data point, self-representation is to represent each sample  $\mathbf{x}_i$  by the other samples, i.e.,  $\mathbf{x}_i = \sum_{j \neq i} \mathbf{Z}_i \mathbf{x}_j$ . Usually, data drawn from real life contain noises or outliers; then, we have  $\mathbf{x}_i = \sum_{j \neq i} \mathbf{Z}_i \mathbf{x}_j + \mathbf{e}$ , where  $\mathbf{e}$  and  $\mathbf{Z}$  are the representation error and the self-representation matrix, respectively. It is robust to the outliers because the self-representation coefficients depend on all the other samples rather than data pairs.

The recent spectral clustering methods search the self-representation matrix by solving the following representation-based model:

$$\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{A}(\mathbf{X})\mathbf{Z}\|_1 + \lambda Q(\mathbf{Z}) \quad \text{s.t.} \quad \mathbf{Z} \in \mathbf{C} \quad (1)$$

where  $\mathbf{X} \in \mathbf{R}^{d \times n}$  is the dataset with  $n$  samples and  $d$  dimensional features.  $\mathbf{A}(\mathbf{X})$  is the dictionary matrix.  $Q(\mathbf{Z})$  and  $\mathbf{C}$  are the regularization and constraint set on  $\mathbf{Z}$ , respectively.  $\lambda > 0$  is a weigh parameter. In this paper, we also utilize this self-representation-based model and set  $\mathbf{A}(\mathbf{X}) = \mathbf{X}$ .

## 2.2 Subspace clustering

The subspace clustering is defined as follows [18]:

**Definition 1** (*Subspace clustering*) Given a group of data vectors  $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_k] \in \mathbf{R}^{d \times n}$  drawn from a union of  $k$  subspaces  $\{\mathcal{S}_i\}_{i=1}^k$  ( $i = 1, \dots, k$ ), let  $X_i$  be a collection of  $n_i$  data vectors that drawn from the subspace  $\mathcal{S}_i$  and  $n = \sum_{i=1}^k n_i$ . The task of subspace clustering is to cluster these data according to the underlying subspaces that they are drawn from.

The detailed steps of spectral clustering are shown as follows: Firstly, construct the affinity matrix  $\mathbf{J}$  (ideally, it should be block-diagonal and its entries of the between-cluster points are zero, i.e., sparse) according to the underlying subspaces. Secondly, compute the first  $k$  eigenvalues and eigenvectors of the affinity matrix  $\mathbf{J}$  so as to construct an eigenvector space. Finally, cluster the eigenvectors by  $k$ -means.

For the noise-free data and the mutually independent subspaces, we could find that the early methods such as the literature [28] can well solve the subspace clustering problem. However, in most of actual applications, the data usually contain a variety of noises or it lie on the cross sections of multiple dependent subspaces. Therefore, the data with different labels may be grouped into the same cluster. This is absolutely incorrect. To eliminate the influence of these errors (i.e., noises) which play an important role on subspace clustering, various subspace clustering methods have been proposed [14–16, 18, 25]. In the following discussion, we will have a review of the related works of the recent spectral clustering-based subspace clustering methods.

Elhamifar and Vidal [15] proposed SSC. They utilized a combination of other samples to represent each sample. Moreover, in this model, representation coefficients were enforced to be sparse. After that, Liu et al. [18] put forward the LRR algorithm for subspace clustering. LRR aims to enforce the affinity matrix to be low rank. In this way, the global information of the samples can be caught. LRR is

robust for the reason that the rank will be high if there are noises in the data. In recent years, Lu et al. [16] put forward a least square regression-based algorithm for the construction of the affinity matrix. It deems that the performance of the subspace clustering could be enhanced by grouping effect. After that, Hu et al. [19] proposed SRC. In this method, the grouping effect of representation coefficients was used to construct the affinity matrix. The SRC algorithm with enforcing grouping effect has a robust performance on subspace clustering. Furthermore, Peng et al. [29] proposed thresholding ridge regression for subspace clustering by eliminating the effect of errors coming from linear projection spaces.

By utilizing  $\ell_1$ -norm (minimization), SSC inspires sparse block-diagonal for clustering, while it is short of grouping effect. On the contrary, LRR utilizes rank minimization, while SRC and LSR make use of  $\ell_2$ -norm. All of these three methods take advantage of strong grouping effect, but they are lack of sparsity. Moreover, they are sensitive to noises and outliers. Therefore, although they have obtained significant success in subspace clustering, none of them could create a pinpoint block-diagonal representation matrix for realistic data.

## 3 Method

### 3.1 Notations

In this work, we utilize bold italic capital letters and lowercase symbols to denote the matrices and vectors, respectively. The trace of a matrix  $\mathbf{A}$  which is square is  $\text{tr}(\mathbf{A})$ .  $\mathbf{A}^T$  means the transpose of  $\mathbf{A}$ , and  $\mathbf{A}^{-1}$  means the inverse of  $\mathbf{A}$ , respectively.  $[\mathbf{A}]_j$  denotes the  $j$ -th column of the matrix  $\mathbf{A}$ . We denote  $v$  converges to  $v_0$  with  $v \rightarrow v_0$ . Several norms of vector and matrix are utilized.  $\|\mathbf{v}\|_p$  means the  $\ell_p$ -norm of the vector  $\mathbf{v}$ .  $\|\mathbf{A}\|_1$ ,  $\|\mathbf{A}\|_F$ ,  $\|\mathbf{A}\|_{2,1}$  denote the  $\ell_1$ -norm ( $\|\mathbf{A}\|_1 = \sum_{j=1}^n \sum_{i=1}^n |A_{ij}|$ ), F-norm,  $\ell_{2,1}$ -norm ( $\sum_j \left\| [\mathbf{A}]_j \right\|_2$ , i.e., sum of the  $\ell_2$ -norms of the column vectors) of  $\mathbf{A}$ , respectively. Rank ( $\mathbf{Z}$ ) means the rank of  $\mathbf{Z}$ , and moreover,  $\|\mathbf{Z}\|_*$  means the sum of all the singular values of  $\mathbf{Z}$ .

### 3.2 Grouping effect

LSR has grouping effect which is only for its specific model and cannot be put into the other models. Aiming to address this problem, in this section, we reanalyze the grouping effect which is reconstruction based and then put forward the enforced block grouping effect conditions (EBGE) which helps us to apply the grouping effect flexibly.

Grouping effect was first demonstrably stated by Lu et al. [16, 25]: If the samples are close to each other, their representation coefficients are also close to each other [19]. The grouping effect makes sure the data to be clustered according to their underlying subspaces. In this way, the data with extremely similar representation coefficients and normally from same subspace could be grouped together. It is defined as follows:

**Definition 2** (*Grouping effect*) Given a set of data  $X = [x_1, x_2, \dots, x_n] \in \mathbf{R}^{d \times n}$ , a self-representation matrix  $Z = [z_1, z_2, \dots, z_n] \in \mathbf{R}^{n \times n}$  has grouping effect if and only if  $\|x_i - x_j\|_2 \rightarrow 0 \Rightarrow \|z_i - z_j\|_2 \rightarrow 0, \forall i \neq j$ .

SSC, LRR and LSR utilize different criterions (independent, noise free or orthogonal) to control the affinities of the data points. The ultimate goal of them is to obtain a block-diagonal affinity matrix for true clustering. While what kind of criteria will be able to get such an affinity matrix? At first, we consider a very simple case by utilizing a basis of subspace as dictionary.

**Theorem 1** (Block-diagonal) *Given the subspace  $\{S_i\}_{i=1}^k$  ( $i = 1, \dots, k$ ) and assuming they are independent,  $B_i$  is a matrix and the columns of it are composed by a basis of the subspaces  $S_i$ ,  $B = [B_1, \dots, B_k]$ , and  $U_i$  is a matrix composed by some column vectors from  $S_i$ ,  $U = [U_1, \dots, U_k]$ . The solution  $Z^*$  of the following equation*

$$U = BZ \quad (2)$$

is unique and block-diagonal.

*Proof* To prove Theorem 1, we just need to prove: For any data point  $y$  from  $S_i$ , there is a unique decomposition of  $y$ :  $y = B_1 z_1 + \dots + B_i z_i + \dots + B_k z_k$ , where  $B_i z_i \in S_i$  ( $i = 1, \dots, k$ ). Since the subspaces are independent, then we could find that  $B_i z_i = y$  and  $B_j z_j = 0$  for all  $j \neq i$ . For  $B_j$  is the basis of the subspace  $S_j$ , thus  $z_i \neq 0$  and  $z_i$  is unique, and  $z_j = 0$  for all  $j \neq i$ .

In Theorem 1, we can know that if the subspaces are independent, it is easy to cluster data accurately by solving Eq. (2), while this model is not suitable for the data with noise which damage the structure of the subspace.

Therefore, we rewrite the model Eq. (1) into a more general form:

$$\min_z f(Z) + \lambda Q(Z) \quad \text{s.t.} \quad Z \in C = \{Z | X = XZ\} \quad (3)$$

If  $f(Z)$  and  $Q(Z)$  satisfy the EBGE conditions, then the solution to problem Eq. (3) is block-diagonal and also has grouping effect.

*Enforced block grouping effect conditions (EBGE)* We assume  $f(Z)$  and  $Q(Z)$  are two functions defined on  $C$  ( $C \neq \emptyset$ ) that is a set of matrices, and moreover,

$Z = \begin{bmatrix} A & U \\ V & B \end{bmatrix} \in C$ . In  $Z$ ,  $A$  and  $B$  are square matrices,  $U$  and  $V$  are of suitable dimension and  $A, B \in C$ .  $Z^M = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \in C$  denotes the block-diagonal matrix of  $Z$ .

The EBGE conditions of problem Eq. (3) are:

1. For any permutation matrix  $P$ , we have  $f(Z) = f(ZP)$ ,  $P, ZP \in C$ .
2.  $f(Z) \geq f(Z^M)$ , the equality is tenable if and only if  $Z = Z^M$ .
3.  $Q(Z) = Q(ZP)$ , and  $Z \in C$  if and only if  $ZP \in C$ , for any permutation matrix  $P$ .
4.  $Z \in C$  if and only if  $P^T Z P \in C$ , and  $Q(Z) = Q(P^T Z P)$ , for any permutation matrix  $P$ .

**Theorem 2** *If  $f(Z)$  and  $Q(Z)$  satisfy the EBGE conditions, the optimal solution  $Z^*$  to problem Eq. (3) is block-diagonal and also has grouping effect:*

$$Z^* = \begin{bmatrix} Z_1^* & 0 & \dots & 0 \\ 0 & Z_2^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_k^* \end{bmatrix}$$

where  $Z_i^* \in \mathbf{R}^{n_i \times n_i}$  match with  $X_i$ , and  $n_i$  is the dimensionality of the subspace  $S_i$  correspondingly, for each  $i$ .

*Proof*

1. The optimal solution of Eq. (3) is block-diagonal. Since  $f(Z) = f(ZP)$  for all permutations  $P$ , thus the objective function is invariant to all permutations. We utilize  $Z^A$  to denote the optimal solution of problem Eq. (3) and decompose it into two parts  $Z^A = Z^F + Z^*$ , where

$$Z^A = \begin{bmatrix} Z_1^* & * & \dots & * \\ * & Z_2^* & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & Z_k^* \end{bmatrix},$$

$$Z^F = \begin{bmatrix} 0 & * & \dots & * \\ * & 0 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 0 \end{bmatrix}.$$

We assume  $[X]_j = [XZ^A]_j \in S_i$ ; thus,  $[XZ^*]_j \in S_i$  and  $[XZ^F]_j \in S_i$  ( $i \neq l$ ). While  $[XZ^F]_j = [XZ^A]_j - [XZ^*]_j$ , and since the subspaces are independent,  $S_l \cap S_i = \emptyset$ , so we have  $[XZ^F]_j = 0$ . After that we have  $XZ^F = \mathbf{0}$  and  $XZ^A = X$ , which  $Z^A$  is doable for the problem Eq. (3). For the EBGE conditions (2),  $f(Z^A) \geq f(Z^*)$  is

tenable, while  $Z^A$  is the optimal solution:  $f(Z^A) \leq f(Z^*)$ . So  $f(Z^A) = f(Z^*)$ , the equality is tenable if and only if  $Z^A = Z^*$ . Therefore, the optimal solution of Eq. (3) is block-diagonal.

- The optimal solution of Eq. (3) has grouping effect. Through Definition 2, we have  $\|X_2 - X_1\|_F \rightarrow 0 \Rightarrow \|Z_2^* - Z_1^*\|_F \rightarrow 0$ . If the EBGE condition (3) is satisfied, it is effortless to verify  $\|z_i - z_j\|_2 \rightarrow 0$  as  $Z_2^*$  and  $Z_1^*$  only differ in the  $i$ -th and the  $j$ -th columns. Inspired by the SRC [19], we can easily know that  $\text{tr}(ZZ^T)$  and  $\text{tr}(Z^T LZ)$  satisfy the EBGE conditions. In this paper, we utilize  $\text{tr}(ZZ^T)$  as the regularization of our objective function.

### 3.3 SRGE Algorithm

As stated in Sect. 1, in this section we propose the SRGE method.

The SRGE algorithm proposed in this paper is very efficient. Firstly, SRGE represents each sample by utilizing the correlations between samples. Secondly,  $\ell_{2,1}$ -norm and trace-norm are introduced as the row sparse constraint and grouping effect constraint, respectively. Finally, the block-diagonal self-representation matrix  $Z$  is generated, and then the affinity matrix  $J$  which is finally put into the spectral model to cluster data is obtained.

According to the definition of the sample self-representation, we need to find a column vector  $z_i \in R^{n \times 1}$ , so that  $x_i$  can be represented by  $Xz_i$ . Normally, the noises and outliers from the data in our real life often bring errors into the representation, i.e.,  $x_i = Xz_i + e$ , where the  $e$  is the representation error. The main purpose of this paper is to find the optimal representation matrix  $Z^*$ , which makes the error between  $X$  and  $XZ$  as small as possible.

As the loss function of SRC is  $\|X - XZ\|_F^2$  and the solution of it is not sparse, and moreover, the influence of noises and outliers of the original data are not well disposed. Therefore, we introduce the  $\|X - XZ\|_{2,1}$  as the loss function of the problem Eq. (1). Because  $\|X - XZ\|_{2,1}$  is not squared, the outliers will become less important than the other samples. The regularization term of LSR is as follows:

$$Q(Z) = \text{tr}(ZZ^T) = \frac{1}{n} \|Z^T e\|_2^2 + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \|z_i - z_j\|_2^2 \quad (4)$$

where  $e$  is the all ones vector. It means to assign equal weights to all pairs of representations regardless of whether the representations are close to others or not. It does have grouping effect but in the implicitly way. In order to explicitly integrate grouping effect into our representation

model, we introduce the following term as the regularization:

$$Q(Z) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \|z_i - z_j\|_2^2 = \text{tr}(ZZ^T) \quad (5)$$

where  $W = (w_{ij})$  is the weight matrix that reflects the closeness of the point  $x_i$  and  $x_j$ .  $L = D - W$  is the Laplacian matrix, in which  $D$  is the diagonal matrix, i.e.,  $D_{ii} = \sum_{j=1}^n w_{ij}$ . The most common way to generate  $W$  is utilizing the  $k$  nearest neighbor ( $k$ -nn) graph by heat kernel [30, 31] or 0–1 weights [32, 33]. In our experiments, we utilize the 0–1 weights which is good enough to construct the  $k$ -nn graph (according to [19], the default value of  $k$  is 4). As we all known, there are many other subspace clustering methods to construct the complex  $k$ -nn graph [3], but that is not the main focus of this paper.

In order to prevent the instability of the samples, we use  $\tilde{L} = L + \xi I$  as the substitution of  $\tilde{L}$  (a default value of  $\xi$  is 0.001); then, the problem Eq. (1) is transformed into the following objective function:

$$\min_Z J(Z) = \|X - XZ\|_{2,1} + \lambda \text{tr}(Z\tilde{L}Z^T) \quad (6)$$

The above objective function firstly utilizes trace-norm [34] to ensure each sample is represented by samples that has the similar representation coefficients (i.e., the closest samples). And then it takes row sparse  $\ell_{2,1}$ -norm as the loss function to reduce the influence of noises and the outliers. Both of the above two norms help to guarantee the model is robust.

Traditional methods utilize  $J_1 = (|Z^*| + |Z^{*T}|)/2$  to obtain the affinity matrix and then use spectral clustering algorithm to cluster data. In this way, the clustering results of these methods will have following characters: The samples in the same subspace are of high similarity, there are great differences between the samples that drawn from different subspaces, and all the subspaces are block-diagonal. These characteristics help spectral clustering-based subspace clustering methods a lot in achieving a good clustering performance.

### 3.4 Subspace clustering by SRGE

The same as SSC, LSR and LRR, our algorithm is also a spectral clustering-based method. After obtaining the optimal self-representation coefficient matrix  $Z^*$ , we compute the affinity matrix by:

$$J_1 = (|Z^*| + |Z^{*T}|)/2 \quad (7)$$

and then utilize the spectral clustering algorithm [23] to generate the final clustering result.

$\mathbf{J}_1$  is efficient, however, its effectiveness mainly due to the block-diagonal character of  $\mathbf{Z}^*$ . Aiming to utilize the grouping effect that is useful to obtain a better subspace clustering result, we compute the affinity matrix by:

$$\mathbf{J}_2 = \frac{(\mathbf{z}_i^{*T} \mathbf{z}_j^*)^2}{\|\mathbf{x}_i\|_2^2 \|\mathbf{x}_j\|_2^2} \quad (8)$$

The new affinity is the inner product of representation coefficient vector normalized by norms of the original features. This normalization decreases the biases generated by the amplitudes of the original attribute. It is very familiar in the motion segmentation and handwritten numeral recognition problems.

The details of subspace clustering by SRGE are summarized in Algorithm 1.

---

**Algorithm 1:** Subspace Clustering via SRGE

---

**Input:** data matrix  $\mathbf{X} \in \mathbf{R}^{n \times n}$ , regularization parameter  $\lambda$

**Output:** clustering error

1. Construct the  $k$ - $m$  graph  $\mathbf{W} \in \mathbf{R}^{n \times n}$  by 0-1 weights and then obtain  $\tilde{\mathbf{L}}$  which is the deformation of Laplacian.
  2. Solve the problem Eq. (3) by the ADMM algorithm, *i.e.*, algorithm 2, to construct the optimal representation coefficient matrix  $\mathbf{Z}^* \in \mathbf{R}^{n \times n}$ .
  3. Define the affinity matrix  $\mathbf{J}$  by Eq. (7) or Eq. (8).
  4. Use spectral clustering algorithm to cluster the data into  $m$  (known from the priori knowledge) clusters.
  5. Calculate the clustering error by Kuhn–Munkres algorithm [22].
- 

### 3.5 Optimization algorithm

We learn that the objective function Eq. (6) is convex but not smooth. It is not feasible to be solved directly. In this work, we propose an efficient optimization algorithm to solve the problem Eq. (6).

At first, differentiating Eq. (6) concerning each column of  $\mathbf{Z}$  and setting it to zero, the optimal solution  $\mathbf{Z}^*$  could be obtained from the following equation:

$$\mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Z} + \mathbf{Z} (\alpha \tilde{\mathbf{L}}) + (-\mathbf{X}^T \mathbf{D} \mathbf{X}) = 0 \quad (9)$$

Equation (9) is a normative Sylvester equation [35], and it has a unique solution:

$$\mathbf{Z} = \text{lyap}(\mathbf{X}^T \mathbf{D} \mathbf{X}, (\lambda \tilde{\mathbf{L}}), (-\mathbf{X}^T \mathbf{D} \mathbf{X})) \quad (10)$$

where  $\mathbf{X}$  and  $\tilde{\mathbf{L}}$  are known,  $\mathbf{D}$  is a block-diagonal matrix, let  $\mathbf{U} = \mathbf{X} - \mathbf{X} \mathbf{Z} = [\mathbf{u}_1, \dots, \mathbf{u}_n]^T$ , then the diagonal elements of  $\mathbf{D}$  are  $d_{ii} = \frac{1}{\|\mathbf{u}_i\|_2}$ . As  $\mathbf{D}$  is unknown and depends on  $\mathbf{Z}$ , we utilize the ADMM algorithm to solve it in the iterative way.

At first, we decompose Eq. (6) into  $N$  subproblem as follows:

$$\arg \min_{z_i} \|\mathbf{x}_i - \mathbf{x}_i z_i\|_{2,1} + \lambda \text{tr}(\mathbf{z}_i \tilde{\mathbf{L}} \mathbf{z}_i^T) \quad (11)$$

where  $z_i$  is the column subsectors of  $\mathbf{Z}$  and  $\text{vec}(\mathbf{Z}) = [z_1, z_2, \dots, z_n]^T$  is satisfied.

For the reason that Eq. (6) utilizes constrained trace-norm as the regularization, the elements of  $\mathbf{Z}$  cannot be solved independently; thus, it is extraordinary and inefficient to utilize the soft threshold  $\lambda$  in the equation. However, it can be used in the process of optimization efficiently; at this time, it is called as multiplier alternating direction method. Under the help of dummy variable, the general form of this method can be rewritten as follows:

$$\arg \min_{\mathbf{Z}, \mathbf{V}} \|\mathbf{X} - \mathbf{X} \mathbf{Z}\|_{2,1} + \lambda \text{tr}(\mathbf{V} \tilde{\mathbf{L}} \mathbf{V}^T) + \rho \|\mathbf{Z} - \mathbf{V}\|_F^2 \quad \text{s.t.} \quad \mathbf{Z} = \mathbf{V} \quad (12)$$

The objective function is complex, so we utilize the extending form of Lagrange function to be a substitute of Eq. (12):

$$\begin{aligned} \mathbf{L}(\mathbf{Z}, \mathbf{C}, \mathbf{\Lambda}) = & \|\mathbf{X} - \mathbf{X} \mathbf{Z}\|_{2,1} + \lambda \text{tr}(\mathbf{V} \tilde{\mathbf{L}} \mathbf{V}^T) \\ & + \frac{\rho}{2} \|\mathbf{Z} - \mathbf{V}\|_F^2 + \text{vec}(\mathbf{\Lambda}) \text{vec}(\mathbf{Z} - \mathbf{V}) \end{aligned} \quad (13)$$

The basic idea of ADMM includes the following iteration steps:

1.  $\mathbf{Z}^{(k+1)} = \arg \min_{\mathbf{Z}} \mathbf{L}(\mathbf{Z}, \mathbf{V}^{(k)}, \mathbf{\Lambda}^{(k)})$
2.  $\mathbf{V}^{(k+1)} = \arg \min_{\mathbf{V}} \mathbf{L}(\mathbf{Z}^{(k+1)}, \mathbf{V}, \mathbf{\Lambda})$
3.  $\mathbf{\Lambda}^{(k+1)} \leftarrow \mathbf{\Lambda}^{(k)} + \rho(\mathbf{Z} + \mathbf{V})$

Given  $\mathbf{A}$  and  $\rho$ , the key to this method is to obtain an optimal solution of Eq. (12). Problem Eq. (6) can be decomposed into two subproblems by using ADMM algorithm.

The first subproblem: If we just minimize the  $\mathbf{Z}$  in Eq. (12), *i.e.*, when the trace-norm penalizes  $\text{tr}(\mathbf{V} \tilde{\mathbf{L}} \mathbf{V}^T)$ , the problem will be transformed into a very simple least square regression problem. The second subproblem: If we just minimize the  $\mathbf{V}$  in Eq. (12), *i.e.*, when  $\|\mathbf{X} - \mathbf{X} \mathbf{Z}\|_{2,1}$  disappeared,  $\mathbf{V}$  can be solved independently. In the above ways, which decompose the problem into two subproblems, the threshold value was utilized efficiently. As the current estimation of  $\mathbf{Z}$  and  $\mathbf{V}$  is combined with the third step of ADMM, the current estimation of the Lagrange multiplier matrix  $\mathbf{\Lambda}$  can be updated. The penalize parameter  $\rho$  plays a special and important role: A flawed estimate  $\mathbf{A}$  is allowed to solve  $\mathbf{Z}$  and  $\mathbf{V}$ .

## 4 Experiments

In this section, we evaluate our SRGE method on four applications of subspace clustering: motion segmentation, handwritten digit clustering, psychology balance clustering

and face clustering. We compare our method with the state-of-the-art representation reconstruction-based methods such as SSC, LRR, LSR and SRC.

### 4.1 Experimental datasets and evaluation criteria

The datasets utilized in our experiment are: Hopkins 155 [36], USPS [37], Jaffe [38] and Balance [39]. All of them are the most common benchmark datasets for judging subspace clustering methods. The best results for all methods are reported.

---

**Algorithm 2.** ADMM algorithm

---

**Input:** dataset  $X$ , penalty parameter  $\rho$

**Output:**  $Z^* \in R^{n \times n}$

1. Initialization:  $Z^0, V^0, A^0$

2. Repeat  $V^{(k+1)} \leftarrow V^{(k)}$ , update  $Z$  by putting  $Z^{(k)}$  into

$$Z^{(k+1)} = \text{lyap}(X^T D^{(k)} X, (\lambda \tilde{L}), (-X^T D^{(k)} X)), \text{ where}$$

$$V^{(k)} = X - XZ^{(k)} = [u_1^{(k)}, \dots, u_n^{(k)}]^T, \quad D^{(k)} \text{ is a block-diagonal matrix and}$$

$$\text{the diagonal elements are } d_i^{(k)} = \frac{1}{\|u_i^{(k)}\|_2}.$$

4.  $A^{(k+1)} \leftarrow A^{(k)} + \rho(Z + V)$

5.  $k = k + 1$

6. Output the optimal solution  $Z^*$  until  $k = 30$ .

---

Hopkins 155 [36], as a motion segmentation dataset, consists of 155 video sequences. Each sequence has two or three motions (correspondingly, a motion is a subspace), and every sequence is a sole clustering task; therefore, there are 155 subspace clustering tasks totally. Unlike other algorithms, our method does not have any dimensionality reduction. Then, we compare our method with SSC, LRR, LSR and SRC. During the comparison, we utilize the same parameter for all sequences, which is the same as the other methods. All the methods are performed on each sequence. And the mean, maximum, minimum and the standard deviation of the clustering error are reported.

USPS [37] is also widely used in subspace clustering. It contains 9298 handwritten digit images. Each of the images has 256 ( $16 \times 16$ ) pixels. We utilize the first 100 images in our experiments.

Jaffe [38] is an international standard face dataset and consists of 10 female positive facial expression images (213 samples), with each image having  $32 \times 32$  pixels.

Balance [39] is the experiment results of psychology. It contains 625 samples that with 4-dimensionality features.

The same as other methods, we utilize clustering error (CE) as a measure of the accuracy [40]. In the condition of the optimal permutation, CE could obtain the minimum

error by matching the ground truth and the clustering result. The definition of it is:

$$CE = 1 - \frac{1}{N} \sum_{i=1}^N \delta(p_i, \text{map}(q_i)) \tag{14}$$

where  $q_i$  and  $p_i$  mean the output label and the ground truth of the  $i$ -th sample, respectively. In above function,  $\delta(x, y) = 1$ , if and only if  $x = y$ , otherwise  $\delta(x, y) = 0$ . The best mapping function,  $\text{map}(q_i)$ , permutes clustering labels to match the ground truth labels and also can be efficiently calculated by the Kuhn–Munkres algorithm [41–45].

### 4.2 Experimental results and analysis

In Table 1, we display the clustering error of five methods on Hopkins 155 database by utilizing the common affinity measure Eq. (7). It reveals that SRGE achieves a clustering error of 3.35 %, while the best result of other methods is 3.92 % achieved by SSC. The improvement of SRGE on this database is limited because the reported error is the mean of 155 clustering errors. Among these errors, most of them are zero, and even if there are high improvements, the mean result is limited by others. Another reason of the limitation is that the correlations of data are very strong, i.e., the dimensions of each subspace are only two or three [16]. In order to distinguish the performance of each method, we represent the best results with boldface in each table.

Tables 2, 4 and 5 show the clustering errors on USPS, Jaffe and Balance, respectively. For fair comparison, we utilize the same affinity measure Eq. (7) in all algorithms. It can be summarized that the performance improvement by our method over others is prominent, especially on Jaffe database, in which the clustering error is 0.94 %. Moreover, as we stated in Sect. 4, the affinity measure Eq. (8) is much better than Eq. (7) in developing the grouping effect. Therefore, in each method, we utilize the new affinity measure  $J_2$  to cluster data on USPS database and the results are displayed in Table 3. By comparing Tables 2 and 3, we can learn that the results of all methods with  $J_2$  are

**Table 1** The clustering results (%) of each method by  $J_1$  on Hopkins 155

| Method   | SSC         | LRR   | LSR         | SRC   | SRGE         |
|----------|-------------|-------|-------------|-------|--------------|
| Max      | 46.97       | 47.64 | 39.71       | 46.70 | <b>38.86</b> |
| Mean     | 3.92        | 5.14  | 4.21        | 4.24  | <b>3.35</b>  |
| Min      | 0           | 0     | 0           | 0     | <b>0</b>     |
| Median   | 0           | 0.53  | 0.52        | 0.29  | <b>0</b>     |
| STD      | <b>7.61</b> | 10.04 | 8.60        | 9.80  | 7.7          |
| Time (s) | 2.50        | 2.03  | <b>0.12</b> | 0.40  | 0.39         |

**Table 2** The clustering results (%) of each method by  $J_1$  on USPS

| Method | LRR   | LSR   | SSC   | SRC   | SRGE         |
|--------|-------|-------|-------|-------|--------------|
| CE (%) | 22.60 | 26.10 | 43.10 | 12.70 | <b>12.10</b> |

**Table 3** The clustering results (%) of each method by  $J_2$  on USPS

| Method | LRR   | LSR   | SSC   | SRC   | SRGE         |
|--------|-------|-------|-------|-------|--------------|
| CE (%) | 17.50 | 18.40 | 42.20 | 11.30 | <b>11.00</b> |

**Table 4** The clustering results (%) of each method by  $J_1$  on Jaffe

| Method | LRR   | LSR   | SSC   | SRC  | SRGE        |
|--------|-------|-------|-------|------|-------------|
| CE (%) | 47.33 | 37.91 | 13.15 | 4.69 | <b>0.94</b> |

**Table 5** The clustering results (%) of each method by  $J_1$  on Balance

| Method | LRR   | LSR   | SSC   | SRC   | SRGE         |
|--------|-------|-------|-------|-------|--------------|
| CE (%) | 46.40 | 38.24 | 43.68 | 41.76 | <b>35.20</b> |

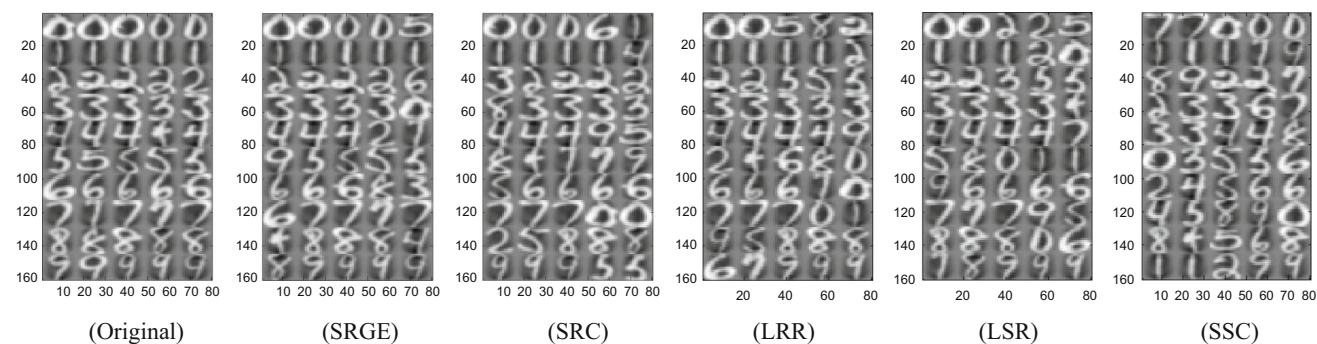
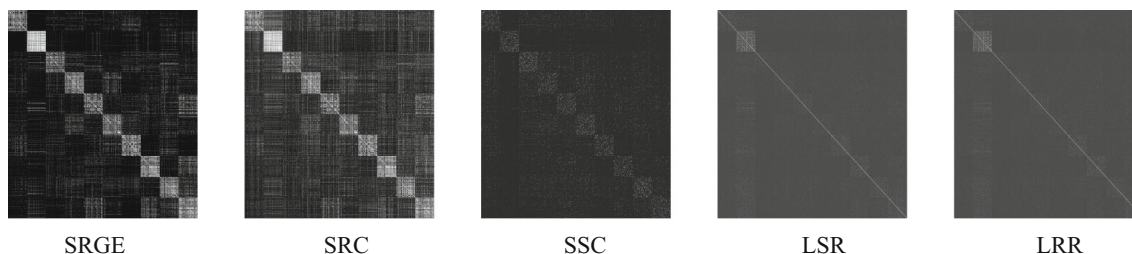
improved significantly than the one with  $J_1$ . Moreover, the result of SRGE obviously outperforms the others.

In Fig. 1, we diagrammatize the clustering results and the original state of the images from USPS by using the affinity measure Eq. (7). As the labels of our experiments are random selected, we rearrange them and use them to generate Fig. 1 eventually. Moreover, we illustrate the affinity matrices

which achieved by utilizing Eq. (8) in Fig. 2. Obviously, the grouping effect from SRGE is better than the one from others. It also indicates that the block-diagonal matrix obtained by SRGE is much more excellent than others.

## 5 Conclusions and future work

In this work, we propose a new subspace clustering method, namely SRGE. We utilize the self-similarity of samples and the trace-norm which has the property of grouping effect to construct the self-representation coefficient matrix. Afterward, by using spectral clustering method, we obtain the clustering result. In addition, effectiveness and robustness of SRGE mainly come from the self-representation of samples and the grouping effect. The former represents each sample by all the other samples rather than data pairs, while the latter makes sure the similar data have the same representation coefficients and be assigned to the same group eventually. Owing to our proposed objective function is convex but not smooth, we use ADMM algorithm to solve it. Moreover, we propose a new affinity measure based on grouping effect. Finally, the experimental results on benchmark datasets indicate that SRGE is much more efficient than the state-of-the-art subspace clustering methods. In the future work, we will use the grouping effect in large-scale subspace clustering and semi-surprised learning.

**Fig. 1** The clustering results of each method on USPS obtained from the first five images of each number**Fig. 2** On USPS database, affinity matrices of SRGE, SRC, SSC, LSR and LRR are generated by utilizing the affinity measure Eq. (8) (the grouping effect of SRGE is much more protruding than those of others)



**Acknowledgments** This work was supported in part by the China 973 Program under Grant 2013CB329404, in part by the National Natural Science Foundation of China under Grant 61450001, Grant 61263035 and Grant 61573270, in part by the Guangxi Natural Science Foundation under Grant 2012GXNSFGA060004 and Grant 2015GXNSFCB139011, in part by the China Postdoctoral Science Foundation under Grant 2015M57570837, in part by the Guangxi Higher Institutions' Program of Introducing 100 High-Level Overseas Talents, in part by the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing and in part by the Guangxi Bagui Scholar Teams for Innovation and Research Project.

## References

- Zhu X, Li X, Zhang S (2016) Block-row sparse multiview multi-label learning for image classification. *IEEE Trans Cybern* 46(2):450–461
- Zhu X, Suk HI, Wang L et al (2015) A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Human Immunol* 75(6):570–577
- Yang Y, Zha Z, Gao Y et al (2014) Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Trans Multimed* 16(6):1677–1689
- Zhu X, Huang Z, Shen HT et al (2013) Linear cross-modal hashing for effective multimedia search. In: *ACM MM*, pp 143–152
- Zhu X, Huang Z, Yang Y et al (2013) Self-taught dimensionality reduction on the high-dimensional small-sized data. *Pattern Recognit* 46(1):215–229
- Wang T, Qin Z, Zhang S et al (2012) Cost-sensitive classification with inadequate labeled data. *Inf Syst* 37(5):508–516
- Zhu X, Suk HI, Lee SW et al (2015) Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. *IEEE Trans Biomed Eng* 63:607–618
- Tomasi C, Kanade T (1992) Shape and motion from image streams under orthography: a factorization method. *Int J Comput Vis* 9(2):137–154
- Zhu X, Zhang L, Huang Z (2014) A sparse embedding and least variance encoding approach to hashing. *IEEE Trans Image Process* 23(9):3737–3750
- Zhu X, Zhang S, Jin Z et al (2011) Missing value estimation for mixed-attribute datasets. *IEEE Trans Knowl Data Eng* 23(1):110–121
- Ma Y, Derksen H, Hong W et al (2007) Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Trans Pattern Anal Mach Intell* 29(9):1546–1562
- Vidal R (2011) Subspace clustering. *IEEE Signal Process Mag* 28(2):52–68
- Costeira JP, Kanade T (2005) A multibody factorization method for independently moving objects. *Int J Comput Vis* 29(3):159–179
- Rene V, Yi M, Shankar S (2005) Generalized principal component analysis (GPCA). *IEEE Trans Pattern Anal Mach Intell* 27(12):1745–1959
- Elhamifar E, Vidal R (2009) Sparse subspace clustering. In: *CVPR*, pp 2790–2797
- Lu CY, Min H, Zhao ZQ et al (2012) Robust and efficient subspace segmentation via least squares regression. In: *ECCV*, pp 347–360
- Peng X, Zhang L, Yi Z (2013) Scalable sparse subspace clustering. In: *Computer vision and pattern recognition (CVPR)*, pp 430–437
- Liu G, Lin Z, Yan S et al (2013) Robust recovery of subspace structures by low-rank representation. *IEEE Trans Pattern Anal Mach Intell* 35(1):171–184
- Hu H, Lin Z, Feng J et al (2014) Smooth representation clustering. In: *CVPR*, pp 3834–3841
- Zhu X, Li X, Zhang S et al (2016) Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE Trans Neural Netw Learn Syst* 1–13
- Donoho DL (2006) For most large underdetermined systems of linear equations the minimal  $\ell_1$ -norm solutions is also the sparsest solution. *Commun Pure Appl Math* 59(6): 797–829
- Zhu X, Huang Z, Cui J et al (2013) Video-to-shot tag propagation by graph sparse group lasso. *IEEE Trans Multimed* 15(3):633–646
- von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17(4):395–416
- Yang Y, Yang Y, Shen H et al (2013) Discriminative nonnegative spectral clustering with out-of-sample extension. *IEEE Trans Knowl Data Eng* 25(8):1760–1771
- Lu C, Lin Z, Yan S (2013) Correlation adaptive subspace segmentation by trace lasso. In: *ICCV*, pp 1345–1352
- Zhu X, Suk HI, Shen D (2014) A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *Neuroimaging* 100:91–105
- Zhang S, Zhang C, Yang Q (2003) Data preparation for data mining. *Appl Artif Intell* 17(5–6):375–381
- Cai D, He XF, Han JW (2005) Document clustering using locality preserving indexing. *IEEE Trans Knowl Data Eng* 17(12):1624–1637
- Peng X, Yi Z, Tang H (2015) Robust subspace clustering via thresholding ridge regression. In: *AAAI conference on artificial intelligence (AAAI)*, pp 3827–3833
- Zhang S (2012) Nearest neighbor selection for iteratively kNN imputation. *J Syst Softw* 85(11):2541–2552
- Zhu X, Huang Z, Shen H, Cheng J et al (2012) Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognit* 45(8):3003–3016
- Zhang S, Qin Z, Ling C et al (2005) “Missing is useful”: Missing values in cost-sensitive decision trees. *IEEE Trans On Knowl and Data Eng* 17(12):1689–1693
- Zhang S, Zhang C, Yan X (2003) Post-mining: maintenance of association rules by weighting. *Inf Syst* 28(7):691–707
- Grave E, Obozinski G, Bach F (2011) Trace lasso: a trace norm regularization for correlated designs. In: *NIPS*, pp 2187–2195
- Bartels R, Stewart G (1972) Solution of the matrix equation  $AX + XB = C$ . *Commun ACM* 15(9):820–826
- Qin Y, Zhang S, Zhu X et al (2007) Semi-parametric optimization for missing data imputation. *Appl Intell* 27(1):79–88
- Hull JJ (1994) A database for handwritten text recognition research. *IEEE Trans Pattern Anal Mach Intell* 16(5):550–554
- Wu X, Zhang C, Zhang S (2005) Database classification for multi-database mining. *Inf Syst* 30:71–88
- Siegler RS (1976) Three aspects of cognitive development. *Cognit Psychol* 28:481–502
- Lancaster P (1970) Explicit solutions of linear matrix equations. *SIAM Rev* 12(4):544–566
- Kuhn H (1955) The Hungarian method for the assignment problem. *Naval Res Logist Q* 2(1–2):83–97
- Elhamifar E, Vidal R (2012) Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans Pattern Anal Mach Intell* 35(11):2765–2781
- Feng J, Lin Z, Xu H et al (2014) Robust subspace segmentation with block-diagonal prior. In: *CVPR*, pp 3818–3825
- Wu X, Zhang C, Zhang S (2004) Efficient mining of both positive and negative association rules. *ACM Trans On Inf Syst* 22(3):381–405
- Liu H, Ma Z, Zhang S et al (2015) Penalized partial least square discriminant analysis with  $l_1$  for multi-label data. *Pattern Recognit* 48(5):1724–1733