



Structured Sparse Subspace Clustering with Within-Cluster Grouping

Huazhu Chen, Weiwei Wang*, Xiangchu Feng

School of Mathematics and Statistics, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:

Received 12 April 2017

Revised 17 February 2018

Accepted 20 May 2018

Available online 22 May 2018

Keywords:

Subspace clustering

Grouping-effect-within-clusters

Affinity matrix learning

ABSTRACT

Many high-dimensional data in computer vision essentially lie in multiple low-dimensional subspaces. Recently developed subspace clustering methods have shown good effectiveness in recovering the underlying low-dimensional subspace structure of high-dimensional data. The state-of-the-art methods show that sparseness and grouping effect of the affinity matrix are important for subspace clustering. The Structured Sparse Subspace Clustering (SSSC) model is a unified optimization framework for learning both the self-representation of the data and their subspace segmentation. But the SSSC only considers structured sparseness property of the affinity matrix. In this work, we define a concept of grouping-effect-within-cluster (GEWC) to group data from the same subspace together. Based on GEWC, we design a new regularization term coupling the self-representation matrix and the segmentation matrix. The new regularization term interactively enforces both to have the expected properties: the segmentation matrix enforces the self-representation coefficient vectors to have large cosine similarity, or GEWC, whenever the data points are drawn from the same subspace and they have the same cluster labels. On the other hand, the self-representation matrix enforces data to have the same cluster labels whenever their self-representation coefficient vectors have large cosine similarity. Incorporating the new penalty into the SSSC model, we present a new unified minimization framework for affinity learning and subspace clustering. The new model considers not only structured sparseness but also GEWC. Experimental results on several commonly used datasets demonstrate that our method outperforms other state-of-the-art methods in revealing the subspace structure of high-dimensional data.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

High-dimensional data are ubiquitous in many computer vision problems and they essentially lie in multiple low-dimensional subspaces, with each subspace corresponding to one cluster. For example, under the assumption of Lambertian reflectance, Basri and Jacobs [1] show that the face images of a subject obtained under a wide variety of lighting conditions lie in a 9-dimensional linear subspace. Recovering the underlying low-dimensional structure in high-dimensional dataset helps to reduce the computational cost [2], memory requirements of algorithms, and noise or outlier possibly existing in the high-dimensional data. This leads to the subspace clustering problem, which is formally defined as follows [3]:

Definition 1. (Subspace Clustering) Let $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in R^{n \times N}$ be a set of data points, where n is the feature dimensionality and N is the number of data points. Assume that the data are drawn from a union of K subspaces $\{S_i\}_{i=1}^K$ of unknown dimensionality,

$\{\mathbf{r}_i\}_{i=1}^K$, respectively. Subspace clustering aims to segment the data into the underlying subspace from which they are drawn.

1.1. Notation

We first introduce some notations used in this work.

For a matrix $Z = (Z_{ij}) = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N) \in R^{N \times N}$ where $\mathbf{z}_j$ is the j th column of the matrix Z , $\|Z\|_1 = \sum_{i=1}^N \sum_{j=1}^N |Z_{ij}|$ and $\|Z\|_F = \sqrt{\sum_{i=1}^N \sum_{j=1}^N |Z_{ij}|^2}$ are the ℓ_1 -norm and Frobenius norm of the matrix

Z , respectively. $\|Z\|_* = \sum_i \sigma_i$ is the trace norm, where σ_i is the singular value of the matrix Z . $\text{tr}(Z) = \sum_i Z_{ii}$ is the trace of the matrix

Z . $\|\mathbf{z}_j\|_2 = \sqrt{\sum_{i=1}^N Z_{ij}^2}$ is the ℓ_2 -norm of the vector $\mathbf{z}_j$. $\text{diag}(Z) \in R^{N \times N}$ is a diagonal matrix whose diagonal entries are Z_{ii} ($i = 1, \dots, N$). $\text{Diag}(\mathbf{z}_j) \in R^{N \times N}$ is also a diagonal matrix whose diagonal entries are Z_{ij} ($i = 1, \dots, N$). $\mathbf{1} \in R^N$ denotes the vector of all 1's. $|Z|$ means the elementwise absolute value of Z .

Define $Q = (Q_{ij}) = [\mathbf{q}^1; \mathbf{q}^2; \dots; \mathbf{q}^N] \in R^{N \times K}$, where $\mathbf{q}^i$, the i th row of the matrix Q , is the cluster indicator of the data point $\mathbf{x}_i$: $Q_{ij} = 1$ if $\mathbf{x}_i$ lies in the j th cluster; otherwise, $Q_{ij} = 0$. We assume

* Corresponding author.

E-mail addresses: chenhuazhu2001@126.com (H. Chen), wwwang@mail.xidian.edu.cn (W. Wang), xfeng@mail.xidian.edu.cn (X. Feng).

that each data point lies in exactly one subspace, hence a valid segmentation matrix Q satisfies $Q\mathbf{1} = \mathbf{1}$. Obviously, $\mathbf{q}^i = \mathbf{q}^j$ if and only if the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace. K is the predefined number of clusters, so we have that $\text{rank}(Q) = K$. We define the collection of segmentation matrices as

$$\mathcal{Q} = \{Q \in \{0, 1\}^{N \times K} : Q\mathbf{1} = \mathbf{1} \text{ and } \text{rank}(Q) = K\}$$

Based on Q , we define $\Theta = (\Theta_{ij}) \in \mathbb{R}^{N \times N}$, where $\Theta_{ij} = \frac{1}{2} \|\mathbf{q}^i - \mathbf{q}^j\|^2$. Θ is a binary matrix: $\Theta_{ij} = 0$ if and only if the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace; otherwise, $\Theta_{ij} = 1$.

1.2. Related work

During the past two decades, a number of subspace clustering methods have been proposed. These methods can be roughly grouped into four categories: iterative methods [4–8], algebraic methods [9–12], statistical methods [13–18] and Spectral clustering based methods [19–40] (see [3] for detail). Iterative methods alternatively assign the data points to different subspaces and fit each subspace with the data points. The methods generally need to set the dimensionality of each subspace in advance and the final results depend on the initialization. A typical representative of the algebraic methods is the factorization based method, which searches for an initial segmentation by thresholding the entries of a similarity matrix (constructed from the decomposition of the data matrix). These methods have strict theoretical guarantees when the subspace is independent. However, they are not robust to data noise or outliers. Statistical methods often assume that the data inside each subspace follows certain kind of distribution, such as Gaussian distribution. Beyond this assumption, the Expectation Maximization (EM) algorithm is used to alternate between data clustering and subspace estimation. The main drawback of these methods is their dependency on estimating the exact subspace models.

Spectral clustering based algorithms are popular most as they are easy to implement, and insensitive to initialization and data corruptions. In Spectral clustering based methods, how to learn a good affinity matrix $A = (A_{ij})$, where A_{ij} measures the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$, is one of the main challenges. Most previous methods learn the affinity matrix from the self-representation of the data points. They find a self-representation matrix Z from the data matrix X by solving the following minimization problem:

$$\min_Z \Omega(Z) + \lambda \Phi(E) \quad \text{s.t. } X = XZ + E, Z \in \mathbb{C} \quad (1)$$

where λ is a tradeoff parameter. $\Omega(Z)$ and $\mathbb{C}$ are the regularization term and the constraint set, which impose some expected properties on Z . $\Phi(E)$ is a function penalizing the representation error, corruptions or outliers in the data points. $\|E\|_F^2$ is usually used for Gaussian noise and $\|E\|_1$ used for outlying entries. The optimal solution Z^* of model (1) is then used to compute the affinity matrix. A commonly used formula is $A = (|Z^*| + |Z^{*T}|)/2$. The final clustering result is obtained by applying a spectral clustering algorithm to the affinity matrix A .

The primary difference between different methods lies in the choice of the regularization term of Z . For example, in the Sparse Subspace Clustering (SSC) [24,25], $\|Z\|_1$ is used as a convex surrogate of $\|Z\|_0$ to promote sparsity of Z . In the Low-Rank Represent (LRR) [26] $\|Z\|_*$ is used to seek a jointly low-rank representation of all data. In the Least Squares Regression (LSR) [29], the squared Frobenius norm $\|Z\|_F^2$ is used to yield a block diagonal representation matrix Z when the subspaces are independent. In the Correlation Adaptive Subspace Segmentation (CASS) [30], the trace lasso $\sum_j \|X \text{Diag}(\mathbf{z}_j)\|_*$ is used for each column of Z to gain a tradeoff effect between ℓ_1 and ℓ_2 norms. In the Smooth Representation Clus-

tering(SMR) [39], $\text{tr}(Z^T LZ)$ is used to enforce spatially close data to have close representation coefficients.

While the above methods have been incredibly successful in many applications, there exists a major disadvantage: the relationship between the self-representation and the data segmentation is not fully exploited. The Structured Sparse Subspace Clustering (SSSC) [40] formulates the learning of the self-representation and the data segmentation into the following unified optimization framework:

$$\min_{Z, E, Q} \|Z\|_1 + \alpha \|\Theta \odot Z\|_1 + \lambda \Phi(E) \quad \text{s.t. } X = XZ + E, \text{diag}(Z) = 0 \quad (2)$$

where the operator $\odot$ denotes the Hadamard product (i.e., element-wise product). $\alpha > 0$ and $\lambda > 0$ are tradeoff parameters.

The first term is the sparsity penalty used by SSC. The role of the second term $\|\Theta \odot Z\|_1 = \sum_i \sum_j \frac{1}{2} \|\mathbf{q}^i - \mathbf{q}^j\|^2 |Z_{ij}|$ in model (2) is

two-fold. As for the self-representation coefficients matrix Z , given the segmentation matrix Q , minimizing $\frac{1}{2} \|\mathbf{q}^i - \mathbf{q}^j\|^2 |Z_{ij}|$ tends to enforce Z_{ij} vanish whenever $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from different subspaces and thus $\Theta_{ij} = 1$. So the role of the second term for Z is that it further enhances a zero affinity between $\mathbf{x}_i$ and $\mathbf{x}_j$ if they are drawn from different subspaces. As for the segmentation matrix Q , the second term plays the role of enforcing $\mathbf{q}^i = \mathbf{q}^j$ whenever $Z_{ij} \neq 0$.

Although the experimental results in [40] show this unified framework outperforms SSC, it has some shortcomings. First, the second term enforces $\mathbf{q}^i = \mathbf{q}^j$ whenever $Z_{ij} \neq 0$ under an implicit assumption that $Z_{ij} \neq 0$ holds if the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace. However, this property is not enforced in the model. Moreover, the first term may cause $Z_{ij} = 0$ even if the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace, thus eliminating the role of Z_{ij} in enforcing $\mathbf{q}^i = \mathbf{q}^j$ in the second term. Second, the model (2) only considers the sparseness of the self-representation coefficients while the grouping effect is ignored. The methods LRR, LSR, CASS and SMR show that the grouping effect (defined in 2) of the self-representation coefficients are also an expected property, which encourages clustering highly correlated data together.

In this work, we first define a new grouping effect, called grouping-effect-within-cluster (GEWC). Instead of grouping spatially close data together (the grouping effect in Definition 2), GEWC tends to group data from the same subspace together, which is exactly one of the goals of subspace clustering. Combining the GEWC and the segmentation matrix Q , we impose a new penalty. The role of the new penalty is also two-fold. Given the segmentation matrix Q , it enforces $\mathbf{z}_i$ highly correlated with $\mathbf{z}_j$ (the GEWC we define in definition 3) whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace and thus $\Theta_{ij} = 0$. On the other hand, given the self-representation matrix Z , the new penalty enforces $\mathbf{q}^i = \mathbf{q}^j$ whenever $\mathbf{z}_i$ is highly correlated with $\mathbf{z}_j$. Incorporating the new penalty into model (2), we present a new unified minimization framework for affinity learning and subspace clustering. Our contributions include:

We define the concept of GEWC. Instead of grouping spatially close data together, GEWC groups data from the same subspace together, which is one of the goals of subspace clustering.

We impose a new regularization term which couples the self-representation matrix Z and the segmentation matrix Q , interactively enforces each other to have the expected properties: the segmentation matrix Q enforces the self-representation coefficient vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ to have large cosine similarity, or GEWC, whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace and they have the same cluster labels. On the other hand, the self-representation matrix Z enforces the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ to have

the same cluster labels $\mathbf{q}^i = \mathbf{q}^j$ whenever their self-representation coefficient vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ have large cosine similarity.

Incorporating the new penalty into the model given in [40], we present a new unified minimization framework for affinity learning and subspace clustering. The new model considers not only structured sparseness but also GEWC, thus can better reveal the subspace structure of high-dimensional data.

Experimental results on several commonly used datasets demonstrate that our method outperforms other state-of-the-art methods.

2. A unified optimization framework for Structured Sparse Subspace Clustering with GEWC

In SMR [39], the authors define the following grouping effect :

Definition 2. (Grouping effect [39]) Given a set of data points $X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{n \times N}$, where n is the feature dimensionality and N is the number of data points. A self-representation coefficient matrix $Z = [\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{N \times N}$ has grouping effect if $\|\mathbf{z}_i - \mathbf{z}_j\|_2 \rightarrow 0$ whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ satisfy $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \rightarrow 0, \forall i \neq j$.

The works in LSR, CASS and SMR confirm that the grouping effect tends to group highly correlated data into the same cluster and experiments therein show that the grouping effect improves the clustering results on several commonly used datasets [41,42]. However, the grouping effect in Definition 2 has some limitations. It essentially enforces spatially close points to have close representation coefficients. For subspace clustering, the data points are assumed to be drawn from a union of several low-dimensional subspaces. Therefore, data points from different subspaces (e.g., near the intersection of two subspaces) could be spatially very close; and conversely, data points from the same subspace could be far apart. So, defining the grouping effect based on spatial distance is unreasonable. It is natural to consider the angle or the cosine similarity between the data points. A large cosine similarity means high possibility that the data points are drawn from the same subspace. Ideally, if the cosine similarity between the data points obtains the maximum value 1, the data points must lie in the same subspace. So we enforce the cosine similarity between the representation coefficients of the data points from the same subspace to be as large as possible. Based on this analysis, we define the GEWC as follows:

Definition 3. (Grouping-effect-within-cluster) Given a set of data points $X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{n \times N}$, where n is the feature dimensionality and N is the number of data points, a subspace self-representation matrix $Z = [\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{N \times N}$ has GEWC if $\frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2} \rightarrow 1$ whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ ($i \neq j$) lie in the same subspace, where $\frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$ is the cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$.

If we know the exact segmentation matrix Q and the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ ($i \neq j$) are drawn from the same subspace, we have $\Theta_{ij} = 0$. GEWC can be achieved by solving the following problem:

$$\max_Z \sum_i \sum_j (\mathbf{1}^T - \Theta)_{ij} \frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2} \quad (3)$$

If $\|\mathbf{z}_i\|_2 = 1$ ($\forall i = 1, \dots, N$), problem (3) is equivalent to the following minimization problem

$$\min_Z \sum_i \sum_j (\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j) \quad (4)$$

Incorporating problem (4)–(2), we obtain the following model

$$\begin{aligned} \min_{Z, E, Q} & \|Z\|_1 + \alpha \|\Theta \odot Z\|_1 + \lambda \sum_i \sum_j (\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j) + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = \mathbf{0}, \|\mathbf{z}_i\|_2 = 1, Q \in \mathbb{Q} \ (\forall i = 1, \dots, N) \end{aligned} \quad (5)$$

where $\alpha > 0$, $\lambda > 0$ and $\beta > 0$ are trade off parameters. When the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same subspace and $\Theta_{ij} = 0$, minimizing the term $\sum_i \sum_j (\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j)$ enforces $\mathbf{z}_i^T \mathbf{z}_j \approx 1$, thus the self-representation coefficients have GEWC. Conversely, if $\mathbf{z}_i^T \mathbf{z}_j \approx 1$, this term enforces $\Theta_{ij} = 0$, or $Q(i, :) = Q(j, :)$, which means the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same label.

Note that, $\|Z\|_1 + \alpha \|\Theta \odot Z\|_1 = \|(\mathbf{1}^T + \alpha \Theta) \odot Z\|_1$, problem (5) can be written as follows

$$\begin{aligned} \min_{Z, E, Q} & \|(\mathbf{1}^T + \alpha \Theta) \odot Z\|_1 + \lambda \sum_i \sum_j (\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j) + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = \mathbf{0}, \|\mathbf{z}_i\|_2 = 1, Q \in \mathbb{Q} \ (\forall i = 1, \dots, N) \end{aligned} \quad (6)$$

where the term $\Phi(E)$ depends upon the prior knowledge about the pattern of noise or corruptions. In model (6), as for the self-representation coefficients Z , the first and the second terms complement each other. Specifically, the first term enforces the self-representation coefficients to be sparse, in particular, a zero affinity $A_{ij} = \frac{|Z_{ij}| + |Z_{ji}|}{2}$ whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from different subspaces; the second term enforces a large cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$ whenever the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same subspace. As for the segmentation matrix Q , the second term enforces the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same cluster labels $Q(i, :) = Q(j, :)$, if they are drawn from the same subspace. In all, the model (6) considers both structured sparseness and GEWC of the self-representation coefficients, so we call it SSSEWC in brief.

Fig. 1 shows the GEWC of the self-representation coefficients obtained by SSC, SSSC and our method. We use all faces of the first, fourth and seventh subjects in the Extended Yale B dataset as the feature data, each subject corresponding to one subspace. Some original images of these data are shown in Fig. 2. For visualization, the dimensionality of the feature data and the self-representation coefficient vectors are reduced to two by using the principle component analysis (PCA). The red, green and blue dots respectively correspond to the first, fourth and seventh subjects: (a) the feature data, (b–d) self-representation coefficients computed by SSC [24,25], SSSC [40], and our method. One can see that, with both structured sparseness and GEWC, the self-representation coefficient vectors obtained by our method show better cluster discriminative property: the self-representation coefficient vectors corresponding to the three subjects lie around three distinct lines, exhibiting large cosine similarities within clusters and small cosine similarities between clusters. The self-representation coefficient vectors obtained by SSC and SSSC have heavy overlap between subjects: they show large cosine similarities within clusters as well as large cosine similarities between clusters. The better cluster discriminative property of our method facilitates the succeeding spectral clustering as shown in the experiment section.

3. An alternating minimization algorithm for our model

In this section, we present an efficient algorithm for our model (6) by alternatively solving the following two subproblems:

1. Fix Q , find Z and E by solving a structured sparse and GEWC representation problem.
2. Fix Z and E , find Q by spectral clustering.

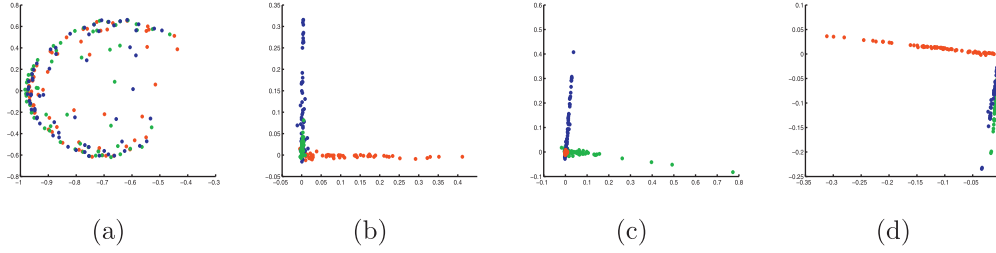


Fig. 1. Grouping effect within cluster of representation based methods. Face images of the first, fourth and seventh subjects in the Extended Yale B dataset as feature data, displayed after reducing the dimensionalities to two by PCA, (a) the original image, (b) self-representation coefficient vectors computed by SSC, (c) self-representation coefficient vectors computed by SSGWC, (d) self-representation coefficient vectors computed by SSSC, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

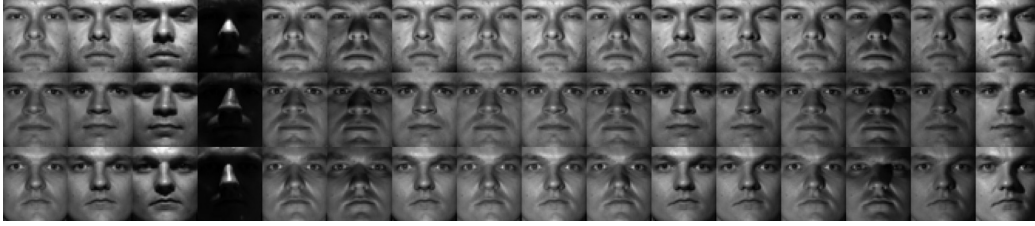


Fig. 2. Sample images from the first, fourth and seventh subjects of the Extended Yale B dataset.

3.1. Solution of the representation

Given the segmentation matrix Q (or the structure matrix Θ), we solve the following problem for Z and E :

$$\begin{aligned} \min_{Z, E} & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \sum_i \sum_j (\mathbf{1}\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j) + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = \mathbf{0}, \|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N) \end{aligned} \quad (7)$$

Using the constraints: $\|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N)$, one can easily find

$$\min_Z \sum_i \sum_j (\mathbf{1}\mathbf{1}^T - \Theta)_{ij} (-\mathbf{z}_i^T \mathbf{z}_j)$$

is equivalent to

$$\min_Z \sum_i \sum_j (\mathbf{1}\mathbf{1}^T - \Theta)_{ij} \frac{1}{2} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

So problem (7) can be converted to

$$\begin{aligned} \min_{Z, E} & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \sum_i \sum_j (\mathbf{1}\mathbf{1}^T - \Theta)_{ij} \frac{1}{2} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = \mathbf{0}, \|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N) \end{aligned} \quad (8)$$

Considering that

$$\sum_i \sum_j (\mathbf{1}\mathbf{1}^T - \Theta)_{ij} \frac{1}{2} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 = \text{tr}(Z^T LZ) \quad (9)$$

where $L = D - (\mathbf{1}\mathbf{1}^T - \Theta)$, and D is a diagonal matrix with diagonal entries $D_{jj} = \sum_i (\mathbf{1}\mathbf{1}^T - \Theta)_{ij}$. Problem (8) can be written as

$$\begin{aligned} \min_{Z, E} & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \text{tr}(Z^T LZ) + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = \mathbf{0}, \|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N) \end{aligned} \quad (10)$$

which is equivalent to the following problem

$$\begin{aligned} \min_{Z, E, C} & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \text{tr}(C^T LC) + \beta \Phi(E) \\ \text{s.t. } & X = XC + E, C = Z - \text{diag}(Z), \|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N) \end{aligned} \quad (11)$$

The problem (11) can be solved by using the Alternating Direction Method of Multipliers (ADMM) [43,44]. The Augmented Lagrange function of (11) is

$$\begin{aligned} L(Z, C, E, Y_1, Y_2) = & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \text{tr}(C^T LC) + \beta \Phi(E) \\ & + \text{tr}[Y_1^T (X - XC - E)] + \text{tr}[Y_2^T (C - Z + \text{diag}(Z))] \\ & + \frac{\mu}{2} (\|X - XC - E\|_F^2 + \|C - Z + \text{diag}(Z)\|_F^2) \end{aligned} \quad (12)$$

where Y_1 and Y_2 are Lagrange multipliers and $\mu > 0$ is a penalty parameter. Since $L(Z, C, E, Y_1, Y_2)$ is separable, it can be optimized by updating each of Z, C, E, Y_1 and Y_2 alternately with others being fixed. The constraints $\|\mathbf{z}_i\|_2 = 1 \ (\forall i = 1, \dots, N)$ are processed by normalization after each iteration. The solutions of the subproblems are as follows:

a) Z -subproblem: Fixing C, E, Y_1 and Y_2 , we update Z by solving the following problem:

$$Z^{t+1} = \arg \min_{\|\mathbf{z}_i\|=1} \frac{1}{\mu^t} \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \frac{1}{2} \|Z - \text{diag}(Z) - U^t\|_F^2 \quad (13)$$

where $U^t = C^t + \frac{1}{\mu^t} Y_2^t$. The problem (13) has the following solution

$$\mathbf{z}_i^{t+1} = \frac{\tilde{\mathbf{z}}_i}{\|\tilde{\mathbf{z}}_i\|_2} \quad (14)$$

where $\tilde{Z} = (\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \dots, \tilde{\mathbf{z}}_N) = \bar{Z} - \text{diag}(\bar{Z})$ and $\bar{Z}$ has the following closed-form:

$$\bar{Z}_{ij} = \text{sgn}(U_{ij}^t) \max(|U_{ij}^t| - \frac{(\mathbf{1}\mathbf{1}^T + \alpha\Theta)_{ij}}{\mu^t}, 0) \quad (15)$$

b) C -subproblem: C is updated by solving the following problem with Z, E, Y_1 and Y_2 being fixed.

$$\begin{aligned} C^{t+1} = \arg \min_C & \lambda \text{tr}(C^T LC) + \frac{\mu^t}{2} (\|XC + E^t - X - \frac{Y_1^t}{\mu^t}\|_F^2 \\ & + \|C - Z^{t+1} + \text{diag}(Z^{t+1}) + \frac{Y_2^t}{\mu^t}\|_F^2) \end{aligned} \quad (16)$$

whose solution is given by

$$C^{t+1} = A^{-1}B \quad (17)$$

where $A = 2\lambda L + \mu^t X^T X + \mu^t I$ and $B = \mu^t (X^T (X - E^t + \frac{Y_1^t}{\mu^t}) + Z^{t+1} - \text{diag}(Z^{t+1}) - \frac{Y_2^t}{\mu^t})$.

c) E -subproblem: Fixing other variables, we update E as follows:

$$E^{t+1} = \arg \min_E \frac{\beta}{\mu^t} \Phi(E) + \frac{1}{2} \|E - (X - XC^{t+1} + \frac{Y_1^t}{\mu^t})\|_F^2 \quad (18)$$

If $\Phi(E) = \|E\|_1$, then

$$E_{ij}^{t+1} = \text{sgn}((X - XC^{t+1} + \frac{Y_1^t}{\mu^t})_{ij}) \max(|(X - XC^{t+1} + \frac{Y_1^t}{\mu^t})_{ij}| - \frac{\beta}{\mu^t}, 0) \quad (19)$$

d) Y_1 and Y_2 -subproblem: The update for the Lagrange multipliers Y_1 and Y_2 is simple gradient ascent step.

$$\begin{aligned} Y_1^{t+1} &= Y_1^t + \mu^t (X - XC^{t+1} - E^{t+1}) \\ Y_2^{t+1} &= Y_2^t + \mu^t (C^{t+1} - Z^{t+1} + \text{diag}(Z^{t+1})) \end{aligned} \quad (20)$$

For clarity, we summarize the ADMM algorithm procedure in Algorithm 1, where t is the number of iteration.

3.2. Spectral clustering

Given the self-representation coefficient matrix Z and noise matrix E , problem (6) becomes the following problem:

$$\arg \min_Q \alpha \|\Theta \odot Z\|_1 + \lambda \sum_i \sum_j \Theta_{ij} \cdot \mathbf{z}_i^T \mathbf{z}_j \quad (21)$$

s.t. $Q \in \mathbb{Q}$

We use the method presented in [40] to solve problem (22). Note that,

$$\begin{aligned} & \alpha \|\Theta \odot Z\|_1 + \lambda \sum_i \sum_j \Theta_{ij} \cdot \mathbf{z}_i^T \mathbf{z}_j \\ &= \alpha \sum_i \sum_j \Theta_{ij} \cdot |z_{ij}| + \lambda \sum_i \sum_j \Theta_{ij} \cdot (\mathbf{z}_i^T \mathbf{z}_j) \\ &= \sum_i \sum_j (\alpha |z_{ij}| + \lambda \mathbf{z}_i^T \mathbf{z}_j) \frac{1}{2} \|Q(i, :) - Q(j, :)\|^2 \\ &= \frac{1}{2} \sum_i \sum_j A_{ij} \|Q(i, :) - Q(j, :)\|^2 \\ &= \text{trace}(Q^T \hat{L} Q) \end{aligned}$$

where $A_{ij} = \frac{1}{2}(\alpha(|z_{ij}| + |z_{ji}|) + \lambda(\mathbf{z}_i^T \mathbf{z}_j + \mathbf{z}_j^T \mathbf{z}_i))$, measures the similarity of points $\mathbf{x}_i$ and $\mathbf{x}_j$, $\hat{L} = \hat{D} - A$ is a graph Laplacian, and $\hat{D}$ is a diagonal matrix whose diagonal entries are $\hat{D}_{jj} = \sum_i A_{ij}$. We relax the problem (21) as follows:

$$\min_Q \text{trace}(Q^T \hat{L} Q) \quad \text{s.t.} \quad Q^T Q = I, Q \in \mathbb{R}^{N \times K} \quad (22)$$

The solution to (22) can be found efficiently by the spectral clustering [45]. In particular, the spectral clustering algorithm starts with finding the K columns of Q as the K eigenvectors corresponding to the smallest eigenvalues of the graph Laplacian matrix $\hat{L}$. It then uses the k-means algorithm to segment the rows of Q into K clusters and this clustering results are used to produce a binary matrix $Q \in \{0, 1\}^{N \times K}$ such that $Q\mathbf{1} = \mathbf{1}$.

3.3. Summary of the proposed algorithm

We summarize the algorithm of solving the proposed problem (6) in Algorithm 2. The algorithm alternatives between solving for the self-representation coefficient matrix Z and the error matrix E by fixing the segmentation Q using Algorithm 1, and solving for Q by fixing Z and E using spectral clustering.

Stopping criterion: The stopping criterion of Algorithm 2 is

$$\|\Theta_{T+1} - \Theta_T\|_\infty < 1$$

where $T = 1, 2, \dots$ is the number of outer iterations.

Initialization and parameters setting: In the algorithm 2, we initialize $\Theta^0 = \mathbf{0}$, where $\mathbf{0}$ is a matrix of all zeros. This means all data points are assign into one subspace at the start. The variables Z and E are also initialized as matrices of all zeros. We fix the iteration index $T = 20$ in all experiments.

Complexity analysis: Thanks to the sparseness of the affinity matrix, the main computational burden of our method is owing to the solution of problem (7) in Algorithm 2. The soft thresholding used for solving Z and E has complexity $O(N^2) + O(nN)$. The complexity for solving C is $O(\max(N^3, nN^2))$. Therefore the computational cost of Algorithm 2 is $O(tT(\max(N^3, nN^2)))$. The memory requirement for our method is $O(\max(nN, N^2))$. The computational complexity of our method is similar to that of the SSSC, however, the clustering performance of our method is better than the SSSC.

4. Experiments

To validate the clustering performance of our method, we test it on five commonly used benchmark datasets: Extended Yale B [46], COIL20 [47], USPS [42], ORL [48] and Hopkins 155 dataset [49]. The SSGEWC method is mainly used for image dataset. We compare our method with some state-of-the-art methods: SSC [25], LRR [26], LSR [29], CASS [30], SMR [39] and SSSC [40]. Since Spectral Clustering (SC) method is used in SSGEWC, SC method with the cosine similarity is also compared with our method. The parameters for each method are tuned to achieve the best performance by hand. The segmentation error is used to evaluate the subspace segmentation performance. As used by most state-of-the-art methods, the clustering error rate is used to evaluate the methods in comparison:

$$\text{error} = \frac{N_{\text{error}}}{N_{\text{total}}} \quad (23)$$

where N_{error} denotes the number of misclassified points, and N_{total} denotes the total number of points. The reported clustering error rates of other methods are obtained by running the codes provided by the authors. Note that all of the methods mentioned above are implemented in the environment of Matlab 8.3 on a machine with Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.60GHz.

4.1. Experiments on Extended Yale B dataset

The Extended Yale B [46] is a human face dataset. It consists of 38 subjects, with 64 frontal face images per subject acquired under various pose and lighting conditions. Some sample images of the first, fourth and seventh subjects in the Extend Yale B dataset are showed in Fig. 2. In this experiments, we follow the protocol introduced in [24,25,40]: (a) each image is resized to 48×42 pixels and rearranged into a 2,016-dimensional vector which is then taken as a sample; (b) the 38 subjects are divided into 4 groups, containing subjects 1–10, 11–20, 21–30, and 31–38, respectively. For each of the first three groups we consider all choices of $K \in \{2, 3, 5, 8, 10\}$ subjects and for the last group we consider all choices of $K \in \{2, 3, 5, 8\}$. The representation error matrix E is measured by ℓ_1 -norm.

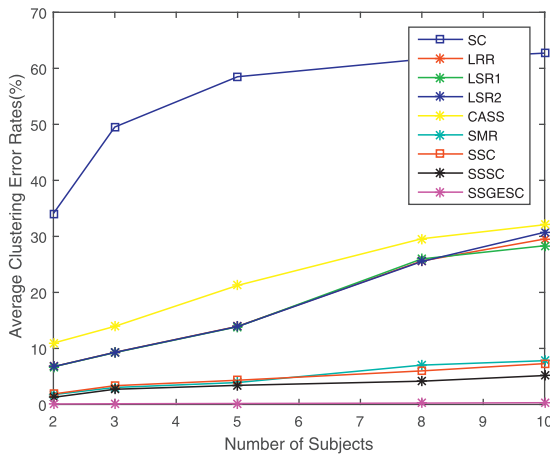
Our model (6) involves three parameters. Experiment on the first group in case of $K = 10$ show that our method performs well with $\alpha \in [0.06, 0.12]$, $\beta \in [0.14, 0.22]$ and $\lambda \in [0.01, 0.1]$. We choose $\alpha = 0.1$, $\beta = 0.16$ and $\lambda = 0.05$ which lead to the best clustering accuracy on the first 10 subjects. Then this setting of parameters are used for all experiments on this dataset.

We chose all choices of data points for each subject number and each group in the experiment. For example, for the case of $K = 2$, the first three groups has totally 45×3 choices of data points and the last group has 28 choices. We perform clustering on each choice. Finally, the average, median and standard deviation of the clustering error rates of all choices are reported in Table 1. For

Table 1

Clustering error rates (%) on Extended Yale B dataset. The best results are in bold font.

methods		SC	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
2 subjects	Ave.	34.08	6.74	6.72	6.74	10.95	1.75	1.87	1.27	0.08
	Med.	40.63	7.03	7.03	7.03	6.25	0.78	0.00	0.00	0.00
	Std.	14.59	6.72	4.16	4.22	12.22	3.36	6.39	5.54	0.33
3 subjects	Ave.	49.52	9.30	9.25	9.29	13.94	3.03	3.35	2.71	0.12
	Med.	49.74	9.90	9.90	9.90	7.81	2.08	0.78	0.52	0.00
	Std.	9.98	3.63	3.64	3.64	14.22	5.57	7.02	6.80	0.31
5 subjects	Ave.	58.50	13.94	13.87	13.91	21.25	3.91	4.32	3.41	0.17
	Med.	58.13	14.38	14.22	14.38	18.91	2.50	2.81	4.88	0.00
	Std.	3.79	3.36	3.40	3.40	13.70	5.39	4.60	1.25	0.27
8 subjects	Ave.	61.62	25.61	25.98	25.52	29.58	7.02	5.99	4.15	0.24
	Med.	61.13	24.80	25.10	24.80	29.20	3.71	4.13	2.93	0.20
	Std.	3.86	5.80	5.48	5.47	5.66	6.41	4.49	3.22	0.24
10 subjects	Ave.	62.71	29.53	28.33	30.73	32.08	7.81	7.29	5.16	0.31
	Med.	62.66	30.00	30.00	33.59	35.31	7.03	5.47	4.22	0.16
	Std.	2.42	4.32	5.65	3.29	11.59	5.90	4.28	4.30	0.27
Ave. of Ave.		53.29	17.02	16.83	17.24	21.56	4.70	4.56	3.34	0.18

**Fig. 3.** Clustering performance on the Yale B dataset: average clustering error rates versus number of subjects.

the case of $K \in \{3, 5, 8\}$, we do experiments in a similar way. For $K = 10$, we perform experiment only on the first three groups. We also show the plots of the average clustering error rates versus the number of subjects in Fig. 3. From Table 1 and Fig. 3, we can see that SSGEWC performs best in terms of the average clustering error rates among all methods in comparison. In addition, the small deviations of each subject indicate that SSGEWC is stable most to outlying entries. Moreover, we can see that our method is stable most to the increasing numbers of subjects from the slowly rising average clustering error rates. More specifically, with structure sparsity, SSSC obtains average clustering error rate 3.34%. With the additional GEWC, our method obtains average clustering error rate 0.18%, which is much less than that of SSSC. This indicates that our GEWC term can effectively complement the structure sparseness in subspace clustering. Without constraint of sparsity, SMR and CASS obtain average clustering error rates 4.70% and 21.56%, which are 4.52% and 21.38% more than that of our method. This also indicates that both sparsity and the GEWC are important for subspace clustering.

To show that our GEWC is more effective than the grouping effect defined in Definition 2, we replace our GEWC term in model (6) with the grouping effect term of SMR and get the following model:

$$\begin{aligned} \min_{Z, E, Q} & \|(\mathbf{1}\mathbf{1}^T + \alpha\Theta) \odot Z\|_1 + \lambda \sum_i \sum_j W_{ij} \|z_i - z_j\|_2^2 + \beta \|E\|_1 \\ \text{s.t.} & X = XZ + E, \text{diag}(Z) = \mathbf{0}, Q \in \mathcal{Q} \end{aligned} \quad (24)$$

Table 2

Clustering error rates (%) on Extended Yale B dataset.

Method	SSGEWC			Model (24)		
	Ave.	Med.	Std.	Ave.	Med.	Std.
2 subjects	0.08	0.00	0.33	0.10	0.00	0.35
3 subjects	0.12	0.00	0.31	0.17	0.00	0.51
5 subjects	0.17	0.00	0.27	0.22	0.00	0.47
8 subjects	0.24	0.20	0.24	0.36	0.20	0.41
10 subjects	0.31	0.16	0.27	0.73	0.78	0.40

where $W = (W_{ij})$ is the weight matrix measuring the spatial closeness of the data points. We use the definition of W in SMR. Table 2 compares the average, median and standard deviation clustering error rates obtained by our model (6) and model (24). The smaller error rates of our method indicate that our GEWC is more efficient than the spatial distance based grouping effect in Definition 2.

To better perceive the different effects of the SSC, SSSC and SSGEWC, we show in Fig. 4 the self-representation coefficient matrix Z , the affinity matrix A , the eigenvectors corresponding to the three smallest eigenvalues of the Laplacian matrix based on A , and the matrix Θ obtained (from top to bottom) on the data matrix consisting the first, fourth and seventh subjects. The affinity matrix used by SSC, SSSC is $A_{ij} = \frac{1}{2}(|Z_{ij}| + |Z_{ji}|)$; The affinity matrix used by our method is $A_{ij} = \frac{1}{2}(\alpha(|Z_{ij}| + |Z_{ji}|) + \lambda(z_j^T z_j + z_i^T z_i))$. The three eigenvectors form a matrix Q of size 192×3 . Fig. 4(g–i) show the matrix Q obtained by SSC, SSSC and SSGEWC, where the red, green and blue points correspond to the 1–64, 64–128, 129–192 rows of the matrix Q , respectively. There are 192 3-dimensional row vectors in Q , which are then used as input to the k -means algorithm to obtain the clustering results.

The diagonal blocks of the self-representation matrix and the affinity matrix measure the similarity between the data within the cluster and the entries off the diagonal blocks correspond to the data outside the cluster. Ideally, the entries off the diagonal blocks matrix Z and A should be zero. From Fig. 4(c,f), we can see that there is fewer nonzero entries off the diagonal blocks of the matrices Z and A obtained by SSGEWC, and the entries within the diagonal blocks dominate the matrices in amplitude. Moreover, the row vectors in Q obtained by our method concentrate in three subjects and there is little overlap between subjects. These good properties of our method lead to the exact clustering result Θ shown in Fig. 4(l). In comparison, the matrices Z and A obtained by SSC and SSSC have much more nonzero entries off the diagonal blocks. The row vectors in Q obtained by SSC and SSSC are scattered and there

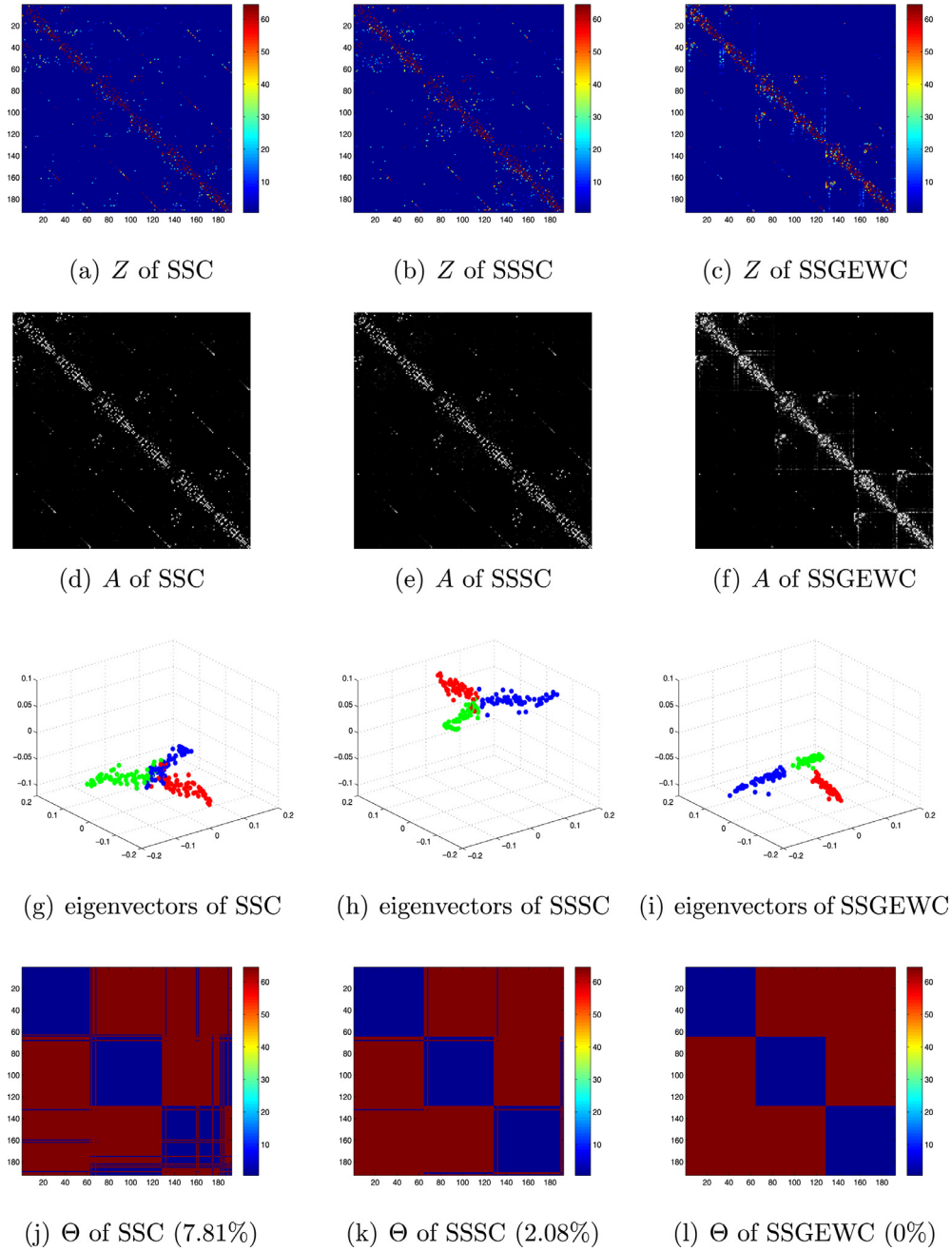


Fig. 4. Visualization of the self-representation coefficient matrices Z , the affinity matrices A , the eigenvectors corresponding to the three smallest eigenvalues of A , and the matrix Θ about of the first, fourth and seventh subjects of the Extended Yale B dataset obtained by SSC, SSSC and our method. The percentage numbers are the corresponding clustering error rates. For ease of visualization, we compute the entry-wise absolute value and rescale each entry by a factor of 800. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are heavy overlap. This lead to wrong clustering as shown by the matrix Θ (Fig. 4(j–k)).

4.2. Experiments on COIL20 dataset

The COIL20 [47] dataset consists of 72×20 grayscale images of 20 objects viewed from varying angles. Fig. 5 shows some sample images. To reduce the computational cost and the memory requirements, we use the first 50 images of each object, and each image is resized to 32×32 pixels and reduce 1024 dimension to 60. The squared Frobenius norm is used to measure the representation error matrix E .

In our experiment, the 20 objects are divided into two groups, where the first group corresponds to the first ten subjects and the second group corresponds to other objects. For each group we consider all choices of $K \in \{6, 7, 8, 9, 10, 20\}$. When $K = 20$, we consider all subjects.

Experiments on the first group show that the proposed method performs well and stable with $\alpha \in [0.05, 0.1]$, $\beta \in [27, 33]$ and $\lambda \in [0.06, 0.1]$. We choose $\alpha = 0.07$, $\beta = 30$ and $\lambda = 0.08$, which lead to the best average clustering accuracy of all choices of eight subjects in the first group. Then we use these parameters for all experiments on this dataset.

To show the performance of our method, we perform clustering on all choices of each case of subject numbers. We report the aver-

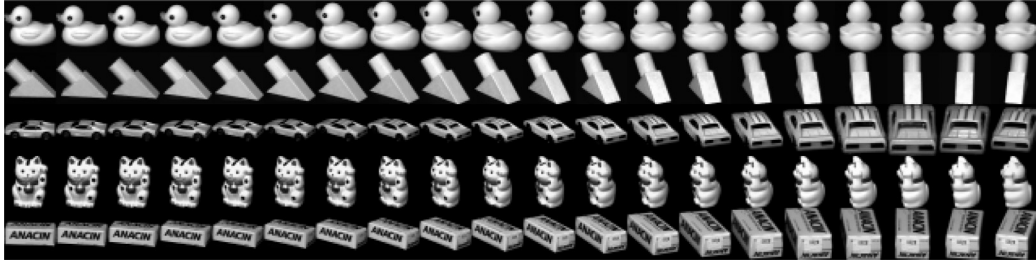


Fig. 5. Sample images from the COIL20 dataset.

Table 3

Clustering error rates (%) on COIL20 dataset. The best results are in bold font.

methods		SC	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
6 subjects	Ave.	11.95	3.46	25.01	23.73	11.02	16.19	10.76	8.68	7.30
	Med.	16.67	22.50	25.00	23.00	14.67	19.67	11.00	7.00	5.50
	Std.	10.05	10.08	12.77	9.70	8.74	10.30	7.63	7.96	7.08
7 subjects	Ave.	12.85	30.89	26.13	24.64	14.96	19.69	13.62	11.20	8.97
	Med.	14.29	27.86	24.57	22.71	14.29	20.57	14.29	14.29	7.71
	Std.	10.19	10.49	12.97	9.67	7.75	7.77	5.86	6.01	5.10
8 subjects	Ave.	13.52	39.17	27.76	25.69	15.95	21.89	14.70	12.93	10.11
	Med.	12.25	37.62	27.75	22.88	16.00	21.00	13.62	13.50	12.50
	Std.	10.52	9.40	13.75	10.32	6.66	6.47	4.95	5.00	4.59
9 subjects	Ave.	13.97	43.34	27.75	26.90	18.37	23.30	15.70	14.22	11.21
	Med.	15.00	42.88	27.75	24.11	15.78	22.00	15.44	13.00	12.00
	Std.	11.35	8.38	13.76	11.01	8.57	6.86	4.77	3.30	3.02
10 subjects	Ave.	13.80	49.00	28.80	27.00	19.40	21.30	18.30	16.80	11.30
	Med.	13.80	49.00	28.80	27.00	19.40	21.30	18.30	16.80	11.30
	Std.	15.98	16.12	16.12	16.40	9.05	3.25	9.62	6.36	1.83
20 subjects	Ave.	13.47	60.13	31.40	33.30	21.30	43.00	26.60	23.40	14.90
	Med.	—	—	—	—	—	—	—	—	—
	Std.	—	—	—	—	—	—	—	—	—
Ave. of Ave.		13.26	40.76	27.71	27.09	16.83	24.22	16.56	14.54	10.63

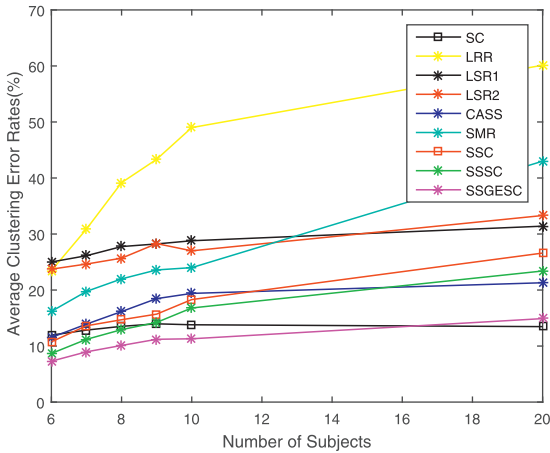


Fig. 6. Clustering performance on the COIL20 dataset: average clustering error rates versus number of subjects.

age, median and standard deviation of the clustering error rates of all choices in Table 3. We also show the plots of the average clustering error rates versus the number of subjects in Fig. 6. It is obvious that, among all methods in comparison, our method achieves best performance except SC in case of $K = 20$. Moreover, for each subject number, the smaller deviation shows that our method is more stable than other methods to choices of data points. The gradually rising plot of the average clustering error rates versus the number of subjects indicates that our method is also stable to the increasing numbers of clusters. Specifically, the average of average clustering error obtained by our method is 10.63%, which is respectively 2.63%, 5.93% and 3.91% less than SC, SSC and SSSC.

4.3. Experiments on USPS dataset

The USPS [42] is a handwritten digits dataset. It consists of 9298 images of 10 subjects, corresponding to 10 handwritten digits, 0–9. Fig. 7 shows some sample images, and each image has 16×16 pixels. We use the first 100 images of each digit for experiments. We also use the standard PCA to project the data into 40-dimensional vectors and the squared Frobenius norm is used to measure the error matrix E . In this experiment, the number K of the underlying subspaces is 10.

Experimental results show that the proposed method performs well with $\alpha \in [0.01, 0.07]$, $\beta \in [1.7, 2.3]$ and $\lambda \in [0.01, 0.05]$ from the experiment results. We choose $\alpha = 0.03$, $\beta = 2$ and $\lambda = 0.01$ which lead to the best clustering accuracy. Table 4 lists the clustering error rates. It can be seen that SSGEWC outperforms other methods on the USPS dataset.

4.4. Experiments on ORL dataset

The ORL [48] is a face recognition dataset. It contains 40 distinct subjects and each subject has 10 different images. We use all the 400 images and resize each image into 28×23 . Fig. 8 shows some samples. All data points are used in this experiment and the number K of the underlying subspaces is 40. We adopt the ℓ_1 -norm for the representation error term E . In this experiment, the above process is repeated 20 test runs and the best clustering performance is given as the final result. Experiments show that the proposed method performs well with $\alpha \in [0.4, 0.6]$, $\beta \in [0.3, 0.5]$ and $\lambda \in [0.4, 0.6]$, and the values of α and λ are larger than β . We choose $\alpha = 0.5$, $\beta = 0.4$ and $\lambda = 0.5$ which lead to the “best” clustering performance. The clustering error rates are reported in Table 5. We can see that SSGEWC outperforms other methods on the ORL dataset from Table 5.



Fig. 7. Sample images from the USPS dataset.

Table 4

Clustering error rates (%) and computational time costs on USPS dataset. The best results are in bold font.

methods	SC	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
10 subjects	17.71	26.90	42.90	25.30	18.00	11.10	10.10	9.30	8.40

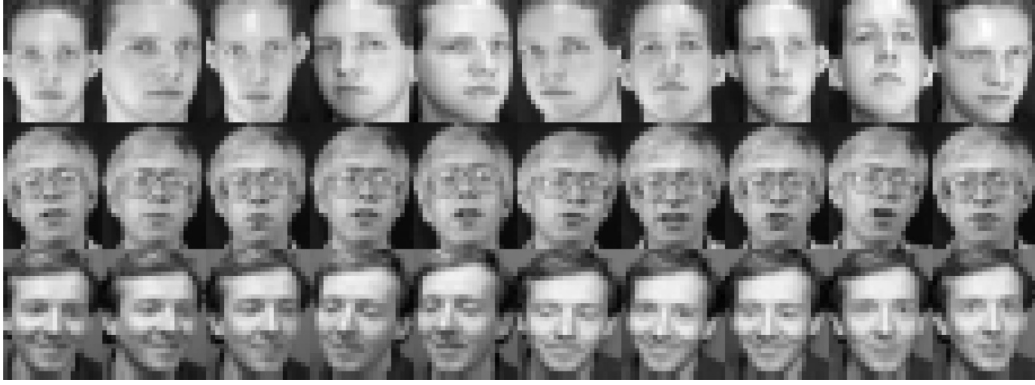


Fig. 8. Sample images from the ORL dataset.

Table 5

Clustering error rates (%) and computational time costs on ORL dataset. The best results are in bold font.

methods	SR	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
40 subjects	17.25	22.08	14.75	14.75	21.75	13.00	15.75	13.00	12.00

Algorithm 1 Solving problem (11) by ADMM.

Input: Data matrix X , Θ^0 , α , λ , β and K

Initialize: $\Theta = \Theta^0$, $C^0 = Z^0 = \mathbf{0}$, $E^0 = \mathbf{0}$, $Y_1^0 = Y_2^0 = \mathbf{0}$, $\mu^0 = 0.1$, $\mu_{\max} = 10^{10}$, $t = 0$, $\rho = 1.1$, $\varepsilon = 10^{-5}$

While not converge do

1: update Z^{t+1} , C^{t+1} , E^{t+1} .

2: update Y_1^{t+1} and Y_2^{t+1} .

3: update $\mu^{t+1} = \min(\mu_{\max}, \rho\mu^t)$.

4: check the convergence conditions

$$\|X - XC^{t+1} - E^{t+1}\|_{\infty} < \varepsilon \text{ and } \|C^{t+1} - Z^{t+1} + \text{diag}(Z^{t+1})\|_{\infty} < \varepsilon$$

5: $t = t + 1$

iiii **end while**

Output: Z^{t+1} and E^{t+1}

Algorithm 2 Solving problem (6).

Input: Data matrix X , α , λ , β and the predefined number of clusters K

Initialize: $\Theta^0 = \mathbf{0}$

While not converge do

1: Fix Q , solve problem (7) via Algorithm 1 to obtain Z and E ;

2: Fix Z and E , solve problem (22) via spectral clustering to obtain Q ;

3: Compute $\Theta_{ij} = \frac{1}{2} \|\mathbf{q}^i - \mathbf{q}^j\|^2$;

iiii **end while**

Output: Segmentation matrix Q

4.5. Experiments on Hopkins 155 dataset

The Hopkins 155 [49] is a motion segmentation dataset, which consists of 155 video sequences with 2 or 3 motions (a motion corresponding to a subspace) in each video [49]. Some samples are shown in Fig. 9. We use the squared Frobenius norm to measure the matrix E , with the affine constraint. The all methods without other preprocessing. SSGEWC performs well with $\alpha \in [0.01, 0.05]$, $\beta \in [\frac{6000}{\min_j \max_{i: i \neq j} x_i^T x_j}, \frac{8000}{\min_j \max_{i: i \neq j} x_i^T x_j}]$ and $\lambda \in [0.01, 0.05]$ on 2

motions. Here, the expression $\frac{\beta}{\min_j \max_{i: i \neq j} x_i^T x_j}$ was originally given in SSC [25]. Experiments on 2 motions show that SSGEWC performs approximately best with $\alpha = 0.02$, $\beta = \frac{7000}{\min_j \max_{i: i \neq j} x_i^T x_j}$ and $\lambda = 0.03$. Then these parameters are also used for 3 motions. Experimental results presented in Table 6 indicate that SSGEWC performs better than other existing methods except SMR. Although the total error rate of SSGEWC is higher than SMR, the error rate of the 2 motions of SSGEWC is less than SMR.

The above experiments show that, our GEWC is more effective than the grouping effect based on spatial distance for subspace clustering in most cases. With both structure sparseness and

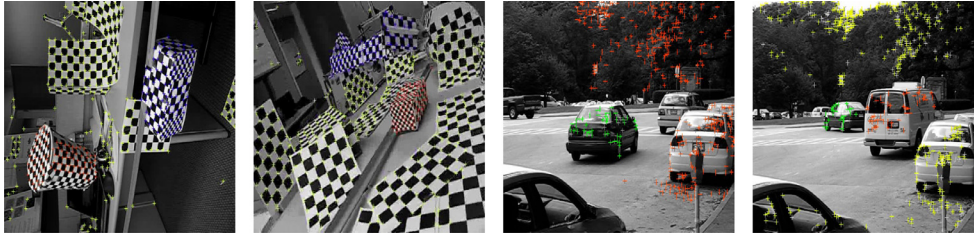


Fig. 9. Sample images from the Hopkins 155 dataset.

Table 6

Clustering error rates (%) on the Hopkins 155 dataset. The best results are in bold font.

methods		SC	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
2 motions	Ave.	10.96	3.76	2.20	2.22	4.88	1.68	1.95	1.68	1.57
	ERR(%)	Med.	7.50	0.00	0.00	0.00	0.28	0.00	0.00	0.00
	Std.	12.46	10.31	5.73	5.73	8.27	3.92	7.19	6.06	4.45
3 motions	Ave.	17.12	6.49	7.18	7.18	12.42	3.44	4.94	5.26	4.12
	ERR(%)	Med.	16.06	1.20	2.40	7.40	1.11	0.89	0.81	0.81
	Std.	9.52	12.32	8.96	8.86	11.97	6.55	9.91	10.35	5.93
Total	Ave.	12.35	4.33	3.31	3.34	6.58	2.08	2.63	2.49	2.14
	ERR(%)	Med.	10.30	0.00	0.22	0.21	0.17	0.00	0.00	0.00
	Std.	12.11	10.82	6.72	6.86	9.72	4.68	7.95	7.37	4.92

Table 7

The runtime costs (in seconds) on the first ten subjects of Extended Yale B dataset in case of $K = 10$, the first ten subjects of COIL20 dataset in case of $K = 10$, the USPS dataset, the ORL dataset and the first 2 motions of Hopkins 155 dataset.

methods		SC	LRR	LSR1	LSR2	CASS	SMR	SSC	SSSC	SSGEWC
Extended Yale B		0.98	68.71	1.30	0.68	175	2.68	20.18	166.81	101.16
COIL20		0.53	1.48	0.79	0.77	1798.34	2.32	7.13	20.19	9.81
USPS		1.46	3.48	2.79	1.77	4778.98	10.73	43.42	51.21	130.22
ORL		1.71	10.23	2.97	2.69	4551.55	11.07	12.84	25.24	17.32
Hopkins 155		0.07	0.10	0.08	0.07	84.32	0.22	0.46	0.52	0.61

GEWC, our method outperforms other state-of-the-art methods in revealing the subspace structure of high-dimensional data.

4.6. Runtime costs

Table 7 presents the real runtime (in seconds) on the first ten subjects of Extended Yale B dataset in case of $K = 10$, the first ten subjects of COIL20 dataset in case of $K = 10$, the USPS dataset, the ORL dataset and the first 2 motions of Hopkins 155 dataset. The two-stage methods except CASS take much less runtime than the unified methods SSSC and our SSGEWC. It is because that the two-stage methods take most runtime in learning the affinity and they infer the cluster labels by running the spectral clustering only once, while the unified methods update both the affinity and the labels iteratively. The CASS method takes much more time than others because it involves SVD and uses more iterations for convergence. SSSC and our method SSGEWC have the same computational complexity, but the runtimes of our method on most datasets are much less because our method takes less iteration for convergence. In all, our method improves the clustering performance at the cost of a moderate computational burden increase.

5. Conclusion

In this work, we define the concept of GEWC and incorporating a new regularization term into the SSSC model. The new regularization term couples the self-representation matrix and the segmentation matrix so that they interactively enforces each other to have the expected properties: the segmentation matrix enforces the self-representation coefficient vectors to have large cosine similarity, or GEWC, whenever the data points are drawn from the

same subspace and they have the same cluster labels. On the other hand, the self-representation matrix enforces data to have the same cluster labels whenever their self-representation coefficient vectors have large cosine similarity. The new unified minimization framework for affinity learning and subspace clustering considers not only structured sparseness but also GEWC. Experimental results on several commonly used datasets demonstrate that our method outperforms other state-of-the-art methods in revealing the subspace structure of high-dimensional data.

Acknowledgments

The authors would like to thank the anonymous reviewers for their considerations and suggestions. We would also give thanks to the [National Natural Science Foundation of China](#) under grants no. 61472303, 61271294, and the Fundamental Research Funds for the Central Universities under grant no. NSIY21 for supporting our research works.

References

- [1] R. Basri, D. Jacobs, Lambertian reflectance and linear subspaces, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2) (2003) 218–233.
- [2] P. Zhu, W. Zhu, Q. Hu, C. Zhang, W. Zuo, Subspace clustering guided unsupervised feature selection, *Pattern Recognit.* 67 (2017). 364–37.
- [3] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (3) (2011) 52–68.
- [4] P.S. Bradley, O.L. Mangasarian, k-plane clustering, *J. Global Optim.* 16 (1) (2000) 23–32.
- [5] P. Tseng, Nearest q-flat to m points, *J. Optim. Theory Appl.* 105 (1) (2000) 249–252.
- [6] T. Zhang, A. Szlam, G. Lerman, Median k-flats for hybrid linear modeling with many outliers, in: *IEEE International Conference on Computer Vision Workshops*, 2009, pp. 234–241.
- [7] P. Agarwal, N. Mustafa, k-means projective clustering, in: *ACM Symposium on Principles of Database Systems*, 2004, pp. 155–165.

- [8] r.O. Rodrigues, L. Torok, P. Liatsis, J. Viterbo, A. Couci, K-MS: a novel clustering algorithm based on morphological reconstruction, *Pattern Recognit.* 66 (2017) 392–403.
- [9] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), *IEEE Trans. Pattern Anal. Mach.Intell.* 27 (12) (2005) 1–15.
- [10] Y. Ma, A.Y. Yang, H. Derksen, R. Fossom, Estimation of subspace arrangements with applications in modeling and segmenting mixed data, *SIAM Rev.* 50 (3) (2008) 413–458.
- [11] K. Huang, Y. Ma, R. Vidal, Minimum effective dimension for mixtures of subspaces: a robust GPCA, algorithm and its applications, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 631–638.
- [12] M.C. Tsakiris, R. Vidal, Algebraic clustering of affine subspaces, *Adv. Pure Math.* 05 (2) (2015) 62–70.
- [13] T. Boult, L. Brown, Factorization-based segmentation of motions, in: *IEEE Workshop on Motion Understanding*, 1991, pp. 179–186.
- [14] A. Leonardis, H. Bischof, J. Mayer, Multiple eigenspaces, *Pattern Recognit.* 35 (11) (2002) 2613–2627.
- [15] C. Archambeau, N. Delannay, M. Verleysen, Mixtures of robust probabilistic principal component analyzers, *Neurocomputing* 71 (7) (2008) 1274–1282.
- [16] A. Gruber, Y. Weiss, Multibody factorization with uncertainty and missing data using the EM algorithm, in: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 10, 2004, pp. 707–714.
- [17] Y. Ma, H. Derksen, W. Hong, J. Wright, Segmentation of multivariate mixed data via lossy coding and compression, *IEEE Trans. Pattern Anal. Mach.Intell.* 29 (9) (2007) 1546–1562.
- [18] A.Y. Yang, S. Rao, Y. Ma, Robust statistical estimation and segmentation of multiple subspaces, in: *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2006, pp. 99–99.
- [19] J. Yan, M. Pollefeys, A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate, in: *European Conference on Computer Vision*, 2006, pp. 94–106.
- [20] A. Goh, R. Vidal, Segmenting motions of different types by unsupervised manifold clustering, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.
- [21] Z. Fan, J. Zhou, Y. Wu, Multibody grouping by inference of multiple subspaces from high-dimensional data using oriented-frames, *IEEE Trans. Pattern Anal. Mach.Intell.* 28 (1) (2006) 91–105.
- [22] G. Chen, G. Lerman, Spectral curvature clustering (SCC), *Int. J. Comput. Vision* 81 (3) (2009) 317–330.
- [23] T. Zhang, A. Szlam, Y. Wang, G. Lerman, Hybrid linear modeling via local best-fit flats, *Int. J. Comput. Vision* 100 (3) (2012) 217–240.
- [24] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.
- [25] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach.Intell.* 35 (11) (2013) 2765–2781.
- [26] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Ma, Robust recovery of subspace structures by low-rank representation, *IEEE Trans. Pattern Anal. Mach.Intell.* 35 (1) (2013) 171–184.
- [27] P. Favaro, R. Vidal, A. Ravichandran, A closed form solution to robust subspace estimation and clustering, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1801–1807.
- [28] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognit. Lett.* 43 (1) (2014) 47–61.
- [29] C.Y. Lu, H. Min, Z.Q. Zhao, L. Zhu, D.S. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, in: *European Conference on Computer Vision*, 2012.
- [30] C. Lu, Z. Lin, S. Yan, Correlation adaptive subspace segmentation by trace lasso, in: *IEEE International Conference on Computer Vision*, 2013.
- [31] C. Lu, S. Yan, Z. Lin, Correntropy induced l2 graph for robust subspace clustering, in: *IEEE International Conference on Computer Vision*, 2013.
- [32] Y.X. Wang, H. Xu, C. Leng, Provable subspace clustering: when LRR meets SSC, *Neural Information Processing Systems*, 2013.
- [33] L. Fei, Y. Xu, X. Fang, J. Yang, Low rank representation with adaptive distance penalty for semi-supervised subspace classification, *Pattern Recognit.* 67 (2017) 252–262.
- [34] R. Heckel, H. Bolcskei, Robust subspace clustering via thresholding, *Inf. Theory IEEE Trans.* 61 (11) (2015) 6320–6342.
- [35] Y. Zhang, Z. Sun, R. He, T. Tan, Robust subspace clustering via half-quadratic minimization, in: *IEEE International Conference on Computer Vision*, 2013, pp. 3096–3103.
- [36] D. Park, C. Caramanis, S. Sanghavi, Greedy subspace clustering, *Neural Information Processing Systems*, 2014.
- [37] B. Li, Y. Zhang, Z. Lin, H. Lu, Subspace clustering by mixture of gaussian regression, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [38] Q. Li, Z. Sun, Z. Lin, R. He, T. Tan, Transformation invariant subspace clustering, *Pattern Recognit.* 59 (2016) 142–155.
- [39] H. Hu, Z. Lin, J. Feng, J. Zhou, Smooth Representation Clustering, in: *Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2014, pp. 3834–3841.
- [40] C.G. Li, R. Vidal, Structured sparse subspace clustering: a unified optimization framework, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 277–286.
- [41] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach.Intell.* 19 (7) (1996) 711–720.
- [42] J.J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Anal. Mach.Intell.* 16 (5) (1994) 550–554.
- [43] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (1) (2010) 1–122.
- [44] Z. Lin, M. Chen, L. Wu, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2011, Arxiv:1009.5055v2.
- [45] U. von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* 17 (4) (2007) 395–416.
- [46] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, *IEEE Trans. Pattern Anal. Mach.Intell.* 23 (6) (2001) 643–660.
- [47] M. Zheng, J. Bu, C. Chen, et al., Graph regularized sparse coding for image representation, *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* 20 (5) (2011) 1327–1336.
- [48] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, applications of computer vision, 1994, in: *Proceedings of the Second IEEE Workshop on. IEEE Xplore*, 1994, pp. 138–142.
- [49] R. Tron, R. Vidal, A benchmark for the comparison of 3-d motion segmentation algorithms, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

Huazhu Chen received the B.S. and M.S. degrees from Henan University, Kaifeng, China, in 2005 and 2008, respectively. Currently she is pursuing her Ph.D. degree at the School of Mathematics and Statistics, Xidian University. Her research interests include subspace clustering, sparse representation, low-rank representation and their applications in image processing.

Weiwei Wang received the B.S., M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 1993, 1998 and 2001, respectively. She is currently a Professor with the School of Mathematics and Statistics, Xidian University. Her research interests include matrix factorization, subspace clustering, sparse representation, low-rank representation and their applications in image processing.

Xiangchu Feng received the B.S. degree in Computational Mathematics from the Xi'an JiaoTong University, Xi'an, China, in 1984, and the M.S. and Ph.D. degrees in Applied Mathematics from Xidian University, Xi'an, in 1989 and 1999, respectively. He is currently a Professor with the School of Mathematics and Statistics, Xidian University. His research interests include numerical analysis, wavelets, and partial differential equations for image processing.