

Discriminative and coherent subspace clustering

Huazhu Chen, Weiwei Wang*, Xiangchu Feng, Ruiqiang He

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

ARTICLE INFO

Article history:

Received 9 June 2017

Revised 6 November 2017

Accepted 2 January 2018

Available online 2 February 2018

Communicated by Dr. Haijun Zhang

Keywords:

Subspace clustering

Discrimination

Coherence

Affinity

Label

ABSTRACT

The ubiquitous large, complex and high dimensional datasets in computer vision and machine learning generate the problem of subspace clustering, which aims to partition the data into several low dimensional subspaces. Most state-of-the-art methods divide the problem into two stages: first learn the affinity from the data and then infer the cluster labels based on the affinity. The Structured Sparse Subspace Clustering (SSSC) model combines the affinity learning and the label inferring into one unified framework and empirically outperforms the two-stage methods. However, the SSSC method does not fully utilize the affinity and the labels to guide each other. In this work, we present a new regularity which combines the labels and the affinity to enforce the coherence of the affinity for data points from the same cluster and the discrimination of the labels for data points from different clusters. Based on this, we give a new unified optimization framework for subspace clustering. It enforces the coherence and discrimination of the affinity matrix as well as the labels, thus we call it Discriminative and Coherent Subspace Clustering (DCSC). Extended experiments on commonly used datasets demonstrate that our method performs better than some two stage state-of-the-art methods and the unified method SSSC in revealing the subspace structure of high-dimensional data.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

In the past few years, technology advances have made data collection easier and faster, resulting in large, multimodal and high-dimensional datasets. How to effectively compress, store, transmit and process massive amounts of such complex high-dimensional data has become a necessary and urgent task. Many existing methods [1] have exploited the observation that the high-dimensional data usually lie in a union of several low-dimensional subspaces or affine spaces. For instance, the face images of a subject obtained under a wide variety of lighting conditions can be accurately approximated with a 9-dimensional linear subspace [2]. This has motivated the problem of subspace clustering, which aims to partition the data points into several low dimensional subspaces and has found numerous applications in computer vision (e.g., image segmentation [3], motion segmentation [4] and face clustering [5]), image processing (e.g., image representation and compression [6]) and systems theory (e.g., hybrid system identification [7]).

Among the existing subspace clustering methods [8–44], the spectral clustering based methods [23–44] are becoming more popular because they are easy to be implemented, and insensi-

tive to initialization and data corruptions. Most spectral clustering based methods divide the problem into two separate stages. First an affinity matrix is learned from the data by using so called self-representation such as Sparse Subspace Clustering (SSC) [28,29], Low-rank Representation (LRR) [30] and some hybrid representation based on SSC or LRR. Then the labels are learned by a spectral clustering method such as Ncut [45]. Although the two-stage methods succeed in many applications, they have a major disadvantage: the relationship between the affinity matrix and the labels of the data is not fully exploited, thus they cannot guarantee an overall optimal performance.

By combining the two stages into one unified framework, the Structured Sparse Subspace Clustering (SSSC) [44] has shown that the overall performance of subspace clustering can be greatly improved. Actually, SSSC uses the self-representation coefficients and the labels to guide each other interactively so that both the affinity and the labels have some advantageous properties. Specifically, it uses the labels to enforce the affinity for data points from different clusters be sparse. Such a property of the affinity is called cluster discrimination property. On the other hand, the self-representation is used to guide the label inferring so that the data points from the same cluster could have same labels. We call this property of the labels coherence property.

Although this unified framework outperforms the two-stage methods, it has some shortcomings. It only enforces the sparseness/discrimination of the affinity matrix for data points from

* Corresponding author.

E-mail addresses: chenhuazhu2001@126.com (H. Chen), wwwang@mail.xidian.edu.cn (W. Wang), xcfeng@mail.xidian.edu.cn (X. Feng), ruiqianghe@sina.com (R. He).

different clusters and the coherence of labels for data points from the same cluster. It does not consider the coherence of the affinity for data points from the same cluster or the discrimination of the labels from different clusters. In all, the coupling of the affinity and the labels is not fully exploited.

In this work, we present a new regularity which combines the labels and the affinity to enforce coherence of the affinity and discrimination of labels. We combine it with the structure sparse regularity in SSSC to give a new unified optimization framework for subspace clustering. The main contributions of this work can be summarized as follows:

We present a label-guided regularity to ensure the coherence of the affinity for data points from the same cluster and the discrimination of the labels for data points from different clusters. By combining the label-guided regularity with the structure sparse regularity in SSSC [44], we give a new unified optimization framework for subspace clustering. It enforces the coherence and discrimination of the affinity matrix as well as the labels, thus we call it Discriminative and Coherent Subspace Clustering (DCSC). It can better recover the subspace structure underlying high dimensional datasets and provide more exact clustering results.

Experiments on several commonly used datasets show that our method outperforms other state-of-the-art subspace clustering methods including SSC [28,29], LRR [30], LRSC [32], LSR [33], CASS [34], TSC [39], NSN [41], SSSC [44], LatLRR [46], BDSSC [47], BDLRR [47] and OMP [48].

2. Related works

Let $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{n \times N}$ be a set of N (sufficiently many) sample points, with each column $\mathbf{x}_i$ being an n -dimensional feature vector, drawn from a union of K subspaces $\{S_c\}_{c=1}^K$ of unknown dimensions $\{r_c\}_{c=1}^K$, respectively. Subspace clustering aims to segment the data into the underlying subspace from which they are drawn.

For convenience, we define some notations used in this work before reviewing related works.

For a matrix $Z = (z_{ij}) = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N) \in \mathbb{R}^{N \times N}$ with $\mathbf{z}_j$ being the j th column, $\|Z\|_1 = \sum_{i,j} |z_{ij}|$ and $\|Z\|_F = \sqrt{\sum_{i,j} |z_{ij}|^2}$ are respectively the ℓ_1 -norm and the Frobenius norm of the matrix Z . $\|Z\|_*$ is the trace norm, i.e., the sum of the singular values of the matrix Z . $\text{diag}(Z) \in \mathbb{R}^{N \times N}$ is the diagonal matrix whose diagonal elements are z_{ii} ($i = 1, \dots, N$). $\text{Diag}(\mathbf{z})$ denotes a diagonal matrix whose i th diagonal element corresponds to the i th entry of the vector $\mathbf{z}$. $\mathbf{1} \in \mathbb{R}^N$ denotes the vector of all $\mathbf{1}$'s.

Define the cluster indicator matrix $Q = (q_{ij}) \in \mathbb{R}^{N \times K}$ by

$$q_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in S_j \\ 0, & \text{if } \mathbf{x}_i \notin S_j \end{cases} \quad (1)$$

Denote the i th row by $Q(i, :)$ and j th column by $Q(:, j)$, then the row $Q(i, :)$ is the cluster label of the data point $\mathbf{x}_i$ and the column $Q(:, j)$ indicates which points belong to cluster S_j . Assume that each data point lies in exactly one subspace or cluster, then each row of Q has only one entry equal to 1, thus a valid cluster indicator matrix Q satisfies $Q\mathbf{1} = \mathbf{1}$. In addition, it is expected that Q has only K different rows due to K subspaces. So Q also satisfies $\text{rank}(Q) = K$. We define the collection of cluster indicator matrices as

$$\mathcal{Q} = \{Q \in \{0, 1\}^{N \times K} : Q\mathbf{1} = \mathbf{1} \text{ and } \text{rank}(Q) = K\} \quad (2)$$

Let $Q^{(j)}$ be an $N \times N_j$ submatrix of $\text{Diag}(Q(:, j))$, consisting of the N_j nonzero columns of the $N \times N$ diagonal matrix $\text{Diag}(Q(:, j))$. Based on Q , we define $P = (p_{ij})$ by

$$p_{ij} = \frac{1}{2} \|Q(i, :) - Q(j, :)\|_2^2 = \begin{cases} 1, & \text{if } Q(i, :) = Q(j, :) \\ 0, & \text{if } Q(i, :) \neq Q(j, :) \end{cases} \quad (3)$$

which indicates whether the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same label, and thus whether they are drawn from the same cluster. So we call P the data connection matrix.

One of the main challenges in spectral clustering-based subspace clustering is how to learn a good affinity matrix $A = (A_{ij})$, where A_{ij} measures the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$. Recently, many works apply the self-representation to learn the affinity. These methods first find a self-representation matrix Z of the data matrix X by solving the following minimization problem:

$$\min_{Z, E} \Omega(Z) + \lambda \Phi(E) \quad \text{s.t. } X = XZ + E, Z \in \mathbb{C}, \quad (4)$$

where $\Omega(Z)$ and $\mathbb{C}$ are the regularity and constraint set, which impose some expected properties on Z . λ is a tradeoff parameter. $\Phi(E)$ is a function penalizing the representation error, corruptions or outliers in the data points. $\|E\|_F^2$ is usually used for Gaussian noise and $\|E\|_1$ is used for sparse entry-wise corruptions. The optimal solution Z^* of problem (4) is used to compute the affinity matrix. A commonly used formula is

$$A = (|Z^*| + |Z^{*T}|)/2 \quad (5)$$

which is further input into a spectral clustering algorithm to produce the final clustering result.

The primary difference between different methods lies in the choice of the regularization term of Z . For example, in the Sparse Subspace Clustering (SSC) [28,29], $\|Z\|_1$ is used as a convex surrogate of $\|Z\|_0$ to promote sparsity of Z . In the Low-Rank Represent (LRR) [30] $\|Z\|_*$ is used to seek a jointly low-rank representation of all data. SSC and LRR show empirical success in some high dimensional datasets. However, a large body of works has shown that SSC performs only optimally in representing data with low correlation and it has the instability problem: if the data from the same subspace are highly correlated or clustered, it will only select one of the several related data at random and ignore other correlated data. This makes it not good for grouping correlated data. The LRR aims at finding the lowest rank representations of all data jointly. It can capture the global structures and not sensitive to noise. However, LRR usually leads to dense representation and results in incorrect clustering. Besides, the number of subspaces and their dimensionality may not be small, thus the data matrix may be high-rank or even full-rank in practice. A number of variants of these algorithms have been proposed, including LatLRR [46], Spatial Weighted SSC [49], LatSSC [50], Kernel SSC [51], etc.

While the above methods have been incredibly successful in many applications, their major disadvantage is that the natural relationship between the coefficient matrix and the segmentation of the data is not explicitly captured. The Structured Sparse Subspace Clustering (SSSC) [44] is a unified optimization framework, which learns the label indicator matrix Q and the self-representation Z simultaneously by solving the following problem:

$$\min_{Z, E, Q} \|Z\|_1 + \alpha \|P \odot Z\|_1 + \lambda \Phi(E) \quad \text{s.t. } X = XZ + E, \quad \text{diag}(Z) = \mathbf{0}, Q \in \mathcal{Q}, \quad (6)$$

where the operator $\odot$ indicates the Hadamard product (i.e., element-wise product). $\alpha > 0$ and $\lambda > 0$ are tradeoff parameters. The unified framework SSSC empirically outperforms the two-stage methods.

3. Discriminative and coherent subspace clustering: a unified framework

3.1. Motivation

Ideally, the affinity and the labels should be coherent within clusters and discriminative between clusters. Specifically, for the

affinity, discrimination means that the affinity between data points from different clusters should be zero so that they have no connection and can be classified into different clusters; coherence means that the affinity between data points from the same cluster should be consistent so that they can be grouped together. For the labels, coherence means the data points from the same cluster should have the same label while discrimination means the data points from different clusters should have diverse labels.

The major disadvantage of the unified framework SSSC is that it does not fully exploit the labels and the affinity to enforce the above expected properties. Actually, in SSSC, only the second term $\|P \odot Z\|_1 = \sum_{ij} \frac{1}{2} \|Q(i, :) - Q(j, :)\|_2^2 |z_{ij}|$ combines the labels and the affinity. It has two-fold roles. Given the cluster indicator matrix Q , minimizing $\frac{1}{2} \|Q(i, :) - Q(j, :)\|_2^2 |z_{ij}|$ tends to enforce z_{ij} vanish, thus no connection between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ whenever $\frac{1}{2} \|Q(i, :) - Q(j, :)\|_2^2 = 1$, or the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have different labels and they could belong to different subspaces. So the role of the second term in affinity learning is that, it enhances the discrimination property of the affinity under the guidance of Q . On the other hand, given Z , the second term tends to enforce $Q(i, :) = Q(j, :)$ whenever $z_{ij} \neq 0$, which means that the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ will be assigned the same label if $z_{ij} \neq 0$, suggesting $\mathbf{x}_i$ and $\mathbf{x}_j$ could lie in the same subspace. So the second term also plays a role in enforcing the coherence of the labels with the guide of Z . In all, SSSC enforces the coherence of labels for data points from the same cluster and the discrimination of the affinity for data points from different clusters. However, SSSC does not consider the coherence of the affinity and the discrimination of the labels.

In this work, we present a new regularity to enforce the coherence of the affinity for data points from the same cluster as well as the discrimination of the labels for data points from different clusters. We combine it with the second term of SSSC to give a new unified optimization framework for subspace clustering. We call it Discriminative and Coherent Subspace Clustering (DCSC) because our model enforces both discriminative and coherent property of the affinity as well as the labels.

3.2. DCSC model

To enforce coherent affinity for data points from the same cluster and discriminative labels for data points from different clusters, we define a new regularity

$$\sum_{j=1}^K \|Z \text{Diag}(Q(:, j))\|_* \quad (7)$$

on the cluster indicator matrix Q and the self-representation matrix Z . This term plays two-fold roles. On one hand, given Z , $\|Z \text{Diag}(Q(:, j))\|_*$ is the trace lasso [34] norm of the column vector $Q(:, j)$. The trace lasso of a vector essentially interpolates between the ℓ_1 -norm and the ℓ_2 -norm of the vector depending on the correlation $Z^T Z$ among the coefficients in Z . In particular, when the coefficients are highly correlated ($Z^T Z$ is close to $\mathbf{1}\mathbf{1}^T$), it is close to the ℓ_2 -norm, while when the data are almost uncorrelated ($Z^T Z$ is close to I), it behaves like the ℓ_1 -norm. The ℓ_2 -norm minimization of a vector tends to make the entries of the vector be uniform while the ℓ_1 -norm minimization tends to make most entries vanish. So for uncorrelated coefficients, the trace lasso $\|Z \text{Diag}(Q(:, j))\|_*$ minimization tends to make the corresponding entries of $Q(:, j)$ vanish while for highly correlated coefficients, the trace lasso $\|Z \text{Diag}(Q(:, j))\|_*$ minimization tends to make the corresponding entries of $Q(:, j)$ uniform. This results in the discrimination of Q . On the other hand, given Q or its j th column $Q(:, j)$, in order to make the regularity (7) be more effective, we replace the regularity in Eq. (7) by

$$\sum_{j=1}^K \|ZQ^{(j)}\|_* \quad (8)$$

$ZQ^{(j)}$ is a submatrix of Z , consisting of the representation coefficients of the N_j points in subspace S_j . Minimizing $\sum_{j=1}^K \|ZQ^{(j)}\|_*$ tends to make the coefficients of data from the same subspace highly correlated, or coherent within clusters.

Combining the new regularity and the subspace structure sparsity in [44], we give the following Discriminative and Coherent Subspace Clustering (DCSC) model:

$$\begin{aligned} \min_{Z, E, Q} & \|Z\|_1 + \frac{\alpha}{2} \sum_{i,j=1}^N \|Q(i, :) - Q(j, :)\|_2^2 |z_{ij}| + \lambda \sum_{j=1}^K \|Z \text{Diag}(Q(:, j))\|_* \\ & + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0, Q \in \mathbb{Q} \end{aligned} \quad (9)$$

The second and the third term enforce the discrimination and coherence of the affinity and the labels under the guide of each other. Note that we also incorporate the ℓ_1 -norm of Z into the model to enforce sparseness/discrimination of the affinity. The term $\Phi(E)$ depends upon the prior knowledge about the pattern of noise or corruptions, and the parameter $\alpha > 0$, $\lambda > 0$ and $\beta > 0$ are tuned to balance the effect of the corresponding terms.

3.3. Minimization algorithm

In this section, we design an efficient algorithm to solve our model (9) by solving the following two subproblems alternatively:

1. Fixed Q , find Z and E by solving a representation problem.
2. Fixed Z and E , find Q by spectral clustering.

3.3.1. Fixed Q , find Z and E

When we fix the cluster indicator matrix Q and denote $ZQ^{(j)}$ by $Z^{(j)}$, the problem for Z and E becomes

$$\begin{aligned} \min_{Z, E} & \|(\mathbf{1}\mathbf{1}^T + \alpha P) \odot Z\|_1 + \lambda \sum_{j=1}^K \|Z^{(j)}\|_* + \beta \Phi(E) \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0. \end{aligned} \quad (10)$$

With the guide of the cluster indicator matrix Q , the first term in problem (10) induces discriminative affinity between clusters through the weighted ℓ_1 -norm while the second term induces coherent affinity within clusters: the coefficients of data from the same cluster are highly correlated, through the nuclear norm. The problem (10) is equivalent to the following problem

$$\begin{aligned} \min_{Z, C, J, E} & \|(\mathbf{1}\mathbf{1}^T + \alpha P) \odot Z\|_1 + \lambda \sum_{j=1}^K \|J^{(j)}\|_* + \beta \Phi(E) \\ \text{s.t. } & X = XC + E, C = Z - \text{diag}(Z), C = J. \end{aligned} \quad (11)$$

We solve this problem by using the Alternating Direction Method of Multipliers (ADMM) [52,53]. The Augmented Lagrange function is given by

$$\begin{aligned} L(Z, C, J, E, Y_1, Y_2, Y_3) = & \|(\mathbf{1}\mathbf{1}^T + \alpha P) \odot Z\|_1 + \lambda \sum_{j=1}^K \|J^{(j)}\|_* + \beta \Phi(E) \\ & + \text{tr}[Y_1^T (X - XC - E)] + \text{tr}[Y_2^T (C - Z \\ & + \text{diag}(Z))] + \text{tr}[Y_3^T (C - J)] \\ & + \frac{\mu}{2} (\|X - XC - E\|_F^2 + \|C - Z \\ & + \text{diag}(Z)\|_F^2 + \|C - J\|_F^2), \end{aligned} \quad (12)$$

where Y_1 , Y_2 and Y_3 are Lagrange multipliers and $\mu > 0$ is a penalty parameter. Since $L(Z, C, J, E, Y_1, Y_2, Y_3)$ is separable, we can update

Z, C, J, E, Y_1, Y_2 and Y_3 alternately while fixing others. The solutions of the subproblems are as follows:

(a) Z -subproblem: We first update Z by fixing C, J, E, Y_1, Y_2 and Y_3 by solving the following problem:

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

where $U^t = C^t + \frac{1}{\mu^t} Y_2^t$. The solution for problem (13) is given by

$$Z^{t+1} = \bar{Z} - \text{diag}(\bar{Z}) \quad (14)$$

and $\bar{Z}$ can be solved by using the soft-thresholding operator:

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

(b) C -subproblem: fixing Z, J, E, Y_1, Y_2 and Y_3 , the problem for C becomes

$$C^{t+1} = \arg \min_C \left\| X - XC - E^t + \frac{Y_1^t}{\mu^t} \right\|_F^2 + \left\| C - Z^{t+1} + \text{diag}(Z^{t+1}) + \frac{Y_2^t}{\mu^t} \right\|_F^2 + \left\| C - J^t + \frac{Y_3^t}{\mu^t} \right\|_F^2. \quad (16)$$

The objective function in Eq. (16) is differentiable. Let its derivative with respect to C be zero, we have the following closed-form solution:

$$C^{t+1} = (X^T X + 2\mu^t I)^{-1} \left(X^T \left(X - E^t + \frac{Y_1^t}{\mu^t} \right) + Z^{t+1} - \text{diag}(Z^{t+1}) - \frac{Y_2^t}{\mu^t} + J^t - \frac{Y_3^t}{\mu^t} \right) \quad (17)$$

(c) J -subproblem: fixing other variables, we update J as follows:

$$J^{t+1} = \arg \min_J \frac{\lambda}{\mu} \sum_j \|J^{(j)}\|_* + \frac{1}{2} \left\| C^{t+1} + \frac{Y_3^t}{\mu^t} - J \right\|_F^2. \quad (18)$$

Note that the squared Frobenius norm is separable and let $H^{(j)} = \left(C^{(t+1)} + \frac{Y_3^t}{\mu^t} \right) Q^{(j)}$, then the problem (18) can be divided into the following subproblems:

$$J^{(j)t+1} = \arg \min_{J^{(j)}} \frac{\lambda}{\mu} \|J^{(j)}\|_* + \frac{1}{2} \|H^{(j)} - J^{(j)}\|_F^2. \quad (19)$$

Each subproblem is a standard low-rank approximation problem and the solution can be obtained by thresholding the singular values [54] of $H^{(j)}$. Specifically,

$$J^{(j)t+1} = U^{(j)} S_\eta(\Sigma^{(j)}) V^{(j)T}, \quad (20)$$

where $S_\eta(\cdot)$ is the shrinkage thresholding operator, and $U^{(j)} \Sigma^{(j)} V^{(j)T}$ is the Singular Value Decomposition (SVD) of $H^{(j)}$. Finally, according to Q , we assemble $J^{(j)t+1} (j = 1, 2, \dots, K)$ into the matrix J^{t+1} .

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

$$E^{t+1} = \arg \min_E \frac{\beta}{\mu^t} \Phi(E) + \frac{1}{2} \left\| E - \left(X - XC^{t+1} + \frac{Y_1^t}{\mu^t} \right) \right\|_F^2 \quad (21)$$

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

$$E_{ij}^{t+1} = \text{sgn} \left(\left(X - XC^{t+1} + \frac{Y_1^t}{\mu^t} \right)_{ij} \right) \max \left(\left| \left(X - XC^{t+1} + \frac{Y_1^t}{\mu^t} \right)_{ij} \right| - \frac{\beta}{\mu^t}, 0 \right) \quad (22)$$

If squared Frobenius norm is used, the solution for problem (21) is

$$E^{t+1} = (2\beta + \mu^t)^{-1} (\mu^t X - \mu^t XC^{t+1} + Y_1^t) \quad (23)$$

(e) Y_1, Y_2 and Y_3 subproblems: Update of the multipliers is the standard gradient ascent procedure:

$$\begin{aligned} Y_1^{t+1} &= Y_1^t + \mu^t (X - XC^{t+1} - E^{t+1}) \\ Y_2^{t+1} &= Y_2^t + \mu^t (C^{t+1} - Z^{t+1} + \text{diag}(Z^{t+1})) \\ Y_3^{t+1} &= Y_3^t + \mu^t (C^{t+1} - J^{t+1}). \end{aligned} \quad (24)$$

For clarity, we outline the ADMM algorithm for solving problem (11) in Algorithm 1.

Algorithm 1 Solving Problem (11) by ADMM.

Input: Data matrix $X, P^0, \lambda, \alpha, \beta$ and K

Initialize: $P = P^0, C = J = Z = Z^0 = \mathbf{0}, Y_1^0 = Y_2^0 = Y_3^0 = \mathbf{0}, \mu_0 = 0.1, \mu_{\max} = 10^{10}, t = 0, \rho = 1.1, \varepsilon = 10^{-5}$

While not converge do

1: update Z^{t+1} by formula (14) and (15)

2: update C^{t+1} by formula (17).

3: update J^{t+1} by formula (20).

4: update E^{t+1} by formula (22) or (23).

5: update the multipliers Y_1^{t+1}, Y_2^{t+1} and Y_3^{t+1} by formula (24).

6: update the parameter μ^{t+1} by $\mu^{t+1} = \min(\mu_{\max}, \rho\mu^t)$.

7: check the convergence conditions

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

8: $t = t + 1$

end while

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

3.3.2. Spectral clustering

Given the coefficient matrix Z and noise matrix E , our DCSC model (9) become the following problem:

$$\arg \min_Q \frac{\alpha}{2} \sum_{i,j=1}^N \|Q(i,:) - Q(j,:)\|_2^2 |z_{ij}| + \lambda \sum_{j=1}^K \|Z \text{Diag}(Q(:,j))\|_* \quad (25)$$

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

Similar to [54], we drop the second term and hence the optimization problem in (25) reduces to:

$$\arg \min_Q \frac{\alpha}{2} \sum_{i,j=1}^N \|Q(i,:) - Q(j,:)\|_2^2 |z_{ij}| \quad \text{s.t. } Q \in \mathbb{Q}. \quad (26)$$

Problem (26) is equivalent to

$$\min_Q \frac{1}{2} \sum_{i,j=1}^N A_{ij} \|Q(i,:) - Q(j,:)\|_2^2 = \min_Q \text{trace}(Q^T L Q) \quad \text{s.t. } Q \in \mathbb{Q}, \quad (27)$$

where $A_{ij} = \frac{1}{2}(\alpha(|z_{ij}| + |z_{ji}|))$ is the affinity similarity defined in Eq. (5), $L = D - A$ is a graph Laplacian, and D is a diagonal matrix whose diagonal entries are $D_{ii} = \sum_{j \neq i} A_{ij}$. We use the method in [54] to solve problem (27). Let $D^{\frac{1}{2}}$ be a diagonal matrix with diagonal entries $\sqrt{D_{ii}}$, $\tilde{Q} = D^{\frac{1}{2}} Q$, $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ and relax the constraint $Q \in \mathbb{Q}$ to $Q^T D Q = I$, the problem (27) can be written as follows:

$$\min_Q \text{trace}(\tilde{Q}^T \tilde{L} \tilde{Q}) \quad \text{s.t. } \tilde{Q}^T \tilde{Q} = I. \quad (28)$$

The solution to this problem can be found efficiently by the spectral clustering [45]. In particular, the columns of the solution of $\tilde{Q}$ is given by the eigenvectors (N-dimensional) of the normalized graph Laplacian matrix $\tilde{L}$ corresponding to the smallest K

eigenvalues. We use the hard/soft strategy [44]: in H-DCSC, we use the K-means algorithm to classify the rows of $\tilde{Q}$ into K clusters and the clustering results are used to produce a binary matrix $Q \in \{0, 1\}^{N \times K}$ such that $Q\mathbf{1} = \mathbf{1}$. The matrix P defined by Eq. (3) is a binary matrix. In S-DCSC, we use $\tilde{Q}$, instead of Q , to construct the matrix P by

$$p_{ij} = \frac{1}{2} \left\| \frac{\tilde{Q}(i, :)}{\|\tilde{Q}(i, :)\|_2} - \frac{\tilde{Q}(j, :)}{\|\tilde{Q}(j, :)\|_2} \right\|_2^2 \in [0, 2] \quad (29)$$

3.4. Connections and differences between DCSC and other related methods

Compared with SSSC, DCSC added one term $\sum_{j=1}^K \|Z \text{Diag}(Q(:, j))\|_*$. In SSSC, the second term combines the cluster labels and the affinity, and as we analyzed in the Section 3.1, it enforces the coherence of labels for data points from the same cluster and the discrimination of the affinity for data points from different clusters. However, SSSC does not consider the coherence of the affinity and the discrimination of the labels. As we analyzed in Section 3.2, the added term $\sum_{j=1}^K \|Z \text{Diag}(Q(:, j))\|_*$ further enforces the coherence of the affinity for data points from the same cluster as well as the discrimination of the labels for data points from different clusters. So, DCSC enforces both discrimination and coherence of the affinity as well as the cluster labels. Our experimental results show that using DCSC can improve the performance of SSSC.

The main difference of the added term $\sum_{j=1}^K \|Z \text{Diag}(Q(:, j))\|_*$ in the proposed DCSC and the $\|Z\|_*$ in LRR is: the former tends to enforce the coefficients of the data points from the same subspace highly correlated, while the later does not consider whether the data points are from the same cluster or not, it essentially enforces the coefficients of all data points highly correlated. In addition, the LRR does not consider the sparseness of the affinity. Moreover, it divides the problem into two separate stages, and the relationship between the affinity matrix and the labels of the data is not fully exploited.

The NSLLR [55] model enforces Z to be sparse by using $\|Z\|_1$ as well as $\|Z\|_*$. It is quite different from our model. The main difference is that, either $\|Z\|_1$ or $\|Z\|_*$ in the NSLLR model does not consider whether the data points are from the same cluster or not, thus may result in wrong discrimination or wrong coherence. By combining the affinity and the cluster labels, our model iteratively uses each to guide the other to have expected discrimination and coherence.

In the DTL-FSSC [56] model, the authors use $\sum_{k=1}^K \|F \text{diag}(Q(:, k))\|_* - \|F\|_*$ to learn a cluster-discriminative transformation matrix and then use it to transform the data points into another domain such that the data points from the same cluster have small angles while the data points from different clusters have large angles. In our DCSC model, we use the term $\sum_{k=1}^K \|Z \text{Diag}(Q(:, j))\|_*$ to learn the data points affinity and we use the data points directly.

3.5. Summary of the proposed algorithm

We summarize the solution of our problem (9) in Algorithm 2. The algorithm alternatively solve for the coefficient matrix Z and the error matrix E by fixing the segmentation Q using Algorithm 1, and solve for Q by fixing Z and E using spectral clustering.

The major computation burden of the DCSC lies in step 3 since it involves the singular value decomposition (SVD). Specifically, the SVD in step 3 is operated on matrix $H^{(j)} \in \mathbb{R}^{n \times N_j}$ ($j = 1, \dots, K$). Singular vector threshold (SVT) operator will be leveraged for low rank requirement with $O(K \min(np^2, n^2p))$ complexity, where $p = \min(N_1, N_2, \dots, N_K)$. The computation complexity of step

Algorithm 2 Solving Problem (9).

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

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

While not converge do

- 1: Fix Q , solve problem (10) via Algorithm 1 to obtain Z and E ;
- 2: Fix Z and E , solve problem (27) via spectral clustering to obtain Q ;

end while

Output: the cluster indicator matrix Q

2 is $O(N^3)$. Meanwhile, in steps 1 and 3, we employ soft thresholding to update the matrix whose complexity is $O(N^2) + O(nN)$. So the total computation complexity of our DCSC is $O(tT(N^3 + K \min(np^2, n^2p) + N^2 + nN))$, where t is the number of iteration in Algorithm 1 with ADMM and T is the number of outer iterations in Algorithm 2. Note that, in Algorithm 2, the previous solution is used to initialize the next execution of ADMM and thus starting from the second iteration in Algorithm 2, t could be remarkably reduced.

4. Experiments

To validate the clustering performance of our method, we test it on three commonly used benchmark datasets: Extended Yale B [57], USPS [58] and Hopkins 155 Database [59]. As used by most state-of-the-art methods, the clustering error rate is used to evaluate the methods in comparison:

$$\text{error} = \frac{N_{\text{error}}}{N_{\text{total}}}, \quad (30)$$

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.

4.1. Experiments on extended Yale B dataset

The Extended Yale B dataset consists of 192×168 pixel face images of 38 subjects and each subject has 64 frontal face images acquired under various pose and lighting conditions. Fig. 1 shows some sample images of the first, third and tenth subjects. We classify the faces by applying subspace clustering. In our experiments, we follow the protocol introduced in [28,29,44]: (a) to reduce the computational cost and the memory requirements, each image is down-sampled to 48×42 pixels and rearranged into a 2,016-dimensional vector; (b) the 38 subjects are divided into 4 groups: the first three groups correspond to subjects 1 to 10, 11 to 20, 21 to 30, and the fourth group corresponds to subjects 31 to 38. 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 $K \in \{2, 3, 5, 8\}$. We use the ℓ_1 -norm to measure the representation error matrix E .

We use the H-DCSC and the S-DCSC to solve our DCSC model (9) in this experiment. We compare the clustering error rates of our method with SSC [29], LRR [30], LatLRR [46], LRSC [32], LSR [33], CASS [34], BDSSC [47], BDLRR [47], TSC [39], OMP [48], NSN[41], DTL-FSSC [56] and SSSC [44].

Our model (9) involves three parameters. Similarly to SSSC, we do experiments for extended choices of 2 subjects. Actually, the first three groups have totally 45×3 choices of samples and the last group has 28 choices. Experiments on all 163 choices of 2 subjects of the Extended Yale B dataset show that the H-DCSC generally performs well with $\alpha \in [0.05, 0.15]$, $\beta \in [0.1, 0.2]$, $\lambda \in [0.01, 0.1]$



Fig. 1. Sample images from the first, third and tenth subjects of the Extended Yale B dataset.

Table 1

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

No.subjects	2		3		5		8		10	
ERR(%)	Ave.	Med.	Ave.	Med.	Ave.	Med.	Ave.	Med.	Ave.	Med.
LRR	6.74 ± 4.22	7.03	9.30 ± 3.63	9.90	13.94 ± 3.36	14.38	25.61 ± 5.08	24.80	29.54 ± 4.32	30.00
LSR1	6.72 ± 4.16	7.03	9.25 ± 3.64	9.90	13.87 ± 3.40	14.22	25.98 ± 5.48	25.10	28.33 ± 5.65	30.00
LSR2	6.74 ± 4.22	7.03	9.29 ± 3.64	9.90	13.91 ± 3.40	14.38	25.52 ± 5.47	24.80	30.73 ± 3.29	33.59
CASS	10.95 ± 12.22	6.25	13.94 ± 14.22	7.81	21.25 ± 13.70	18.91	29.58 ± 5.66	29.20	32.08 ± 11.59	35.31
LRSC	3.15	2.34	4.71	4.17	13.06	8.44	26.83	28.71	35.89	34.84
BDSSC	3.90	–	17.70	–	25.70	–	33.20	–	39.53	–
BDLRR	3.91	–	10.02	–	12.97	–	27.70	–	30.84	–
LatLRR	2.54	0.78	4.21	2.60	6.90	5.63	14.34	10.06	22.92	23.59
TSC	8.06	–	9.00	–	10.14	–	12.58	–	17.86	–
OMP	4.45	–	6.35	–	8.93	–	12.90	–	9.82	–
NSN	1.71	–	3.63	–	5.81	–	8.46	–	9.82	–
SSC	1.87 ± 6.39	0.00	3.35 ± 7.02	0.78	4.32 ± 4.60	2.81	5.99 ± 4.13	4.49	7.29 ± 4.28	5.47
H-SSSC	1.27 ± 5.54	0.00	2.71 ± 6.80	0.52	3.41 ± 4.88	1.25	4.15 ± 3.22	2.93	5.16 ± 4.30	4.22
DTL-FSSC	–	–	1.93	1.04	3.22	2.19	–	–	5.02	2.81
S-SSSC	0.76 ± 3.90	0.00	0.82 ± 1.14	0.52	1.32 ± 0.99	1.25	2.14 ± 1.05	1.95	2.40 ± 1.10	2.50
H-DCSC	0.20 ± 0.69	0.00	0.37 ± 0.92	0.00	0.55 ± 1.05	0.31	1.00 ± 0.18	0.18	1.87 ± 0.47	2.71
S-DCSC	0.07 ± 0.30	0.00	0.11 ± 0.33	0.00	0.11 ± 0.23	0.00	0.12 ± 0.18	0.00	0.10 ± 0.18	0.00

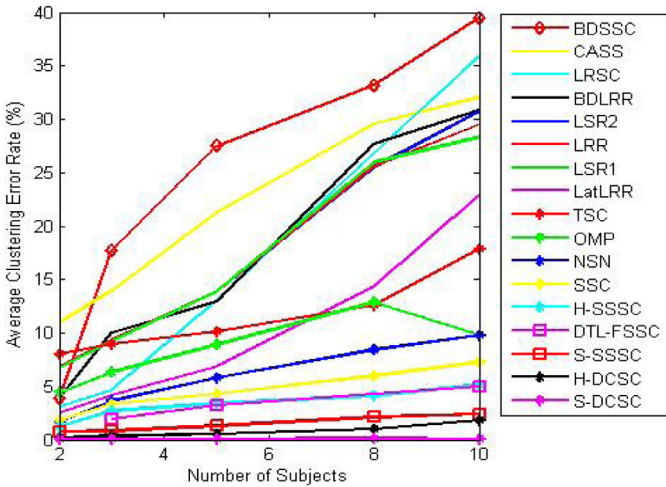


Fig. 2. Clustering performance on the Extended Yale B: average clustering error rate versus number of subjects.

and the S-DCSC generally performs well with $\alpha \in [0.2, 0.6]$, $\beta \in [0.1, 0.2]$, $\lambda \in [0.2, 0.5]$. We choose $\alpha = 0.1$, $\beta = 0.14$, $\lambda = 0.07$ for the H-DCSC and $\alpha = 0.4$, $\beta = 0.15$, $\lambda = 0.37$ for the S-DCSC because which result in the best average clustering accuracy on all 163 choices of 2 subjects. Then we use this setting of parameters for all experiments on this dataset. To show the performance of our method, we perform clustering on all choices of data points for each subject number and each group. Finally, we report the mean, the standard deviation and the median of the clustering error rates of all choices in Table 1, where “–” denotes the data is not reported. For better comparison, we plot the average clustering error rates of different methods for all numbers of subjects in Fig. 2.

From Table 1 and Fig. 2, one can see that, among all methods in comparison, our method performs best in terms of the average clustering error rates for all numbers of subjects. The small deviations indicate that our method is stable most to the pose and lighting conditions. Moreover, as the number of subjects increases, the average clustering error rates of our methods increase most slowly, showing that our method is stable most to the increasing numbers of subjects. In all, our method outperforms other existing methods in terms of the average clustering accuracy and the robustness. For 10 subjects, compared with SSC (the best two-stage method), our H-DCSC and S-DCSC improve the average clustering error rate of the S-SSSC by 5.42% and 7.19%. Compared with SSSC, for 10 subjects, our H-DCSC improves the clustering error rate of the H-SSSC by 3.29%, while the S-DCSC improves the clustering error rate of the S-SSSC by 2.30%. And compared with TDL-FSSC, our H-DCSC improves the clustering error rate by 3.15%, while the S-DCSC improves the clustering error rate by 4.92% for 10 subjects. Our DCSC method far outperforms the two-stage methods mainly because it combines the affinity learning and label inferring so that they can interactively guide each other to have the expected properties for clustering. Our DCSC outperforms the existing unified methods mainly because it utilizes more fully the affinity and the labels to guide each other so that both are discriminative between clusters and coherent within clusters.

To better understand why our DCSC outperforms SSSC, we compare them visually. Fig. 3 shows (from top to bottom) the self-representation coefficient matrix Z , the affinity matrix A , the row vectors of $\tilde{Q}$ and the data connection matrix P learnt from the data matrix consisting all samples of the first, third and tenth subjects. Recalling that the column of $\tilde{Q}$ is the eigenvectors of the normalized graph Laplacian matrix $\tilde{L}$ corresponding to the smallest K eigenvalues, $\tilde{Q}$ has 192 rows and each row is a 3-dimensional vector. The row vectors of $\tilde{Q}$ obtained by H-SSSC, S-SSSC, H-DCSC and

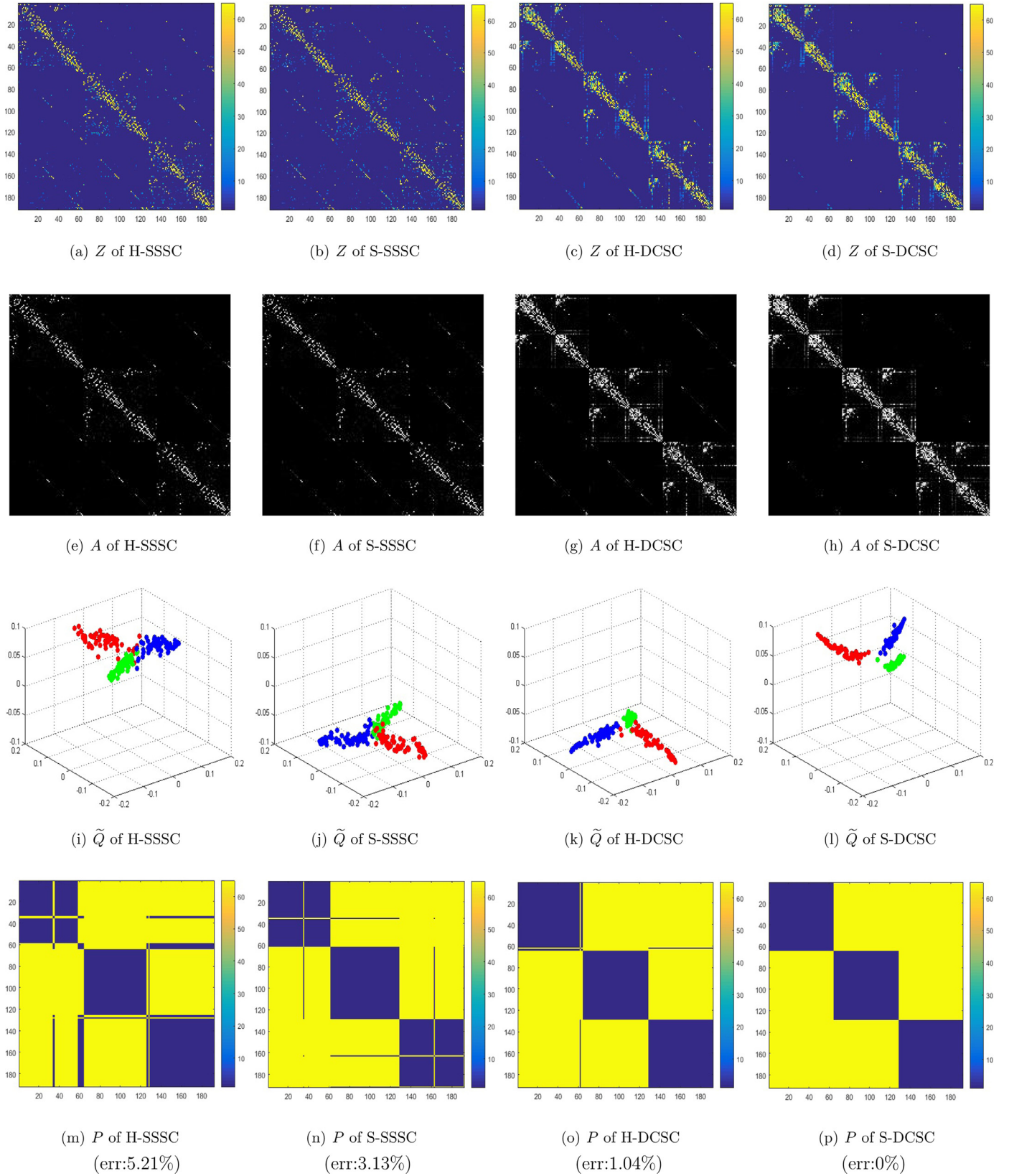


Fig. 3. Visualization of the self-representation matrices Z , the affinity matrices A , the eigenvectors corresponding to the three smallest eigenvalues of A , and the matrix P obtained by SSSC and our method. The percentage numbers are the corresponding clustering error rates. For ease of visualization, we show the absolute value of the matrixes and rescale each entry by a factor of 800. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

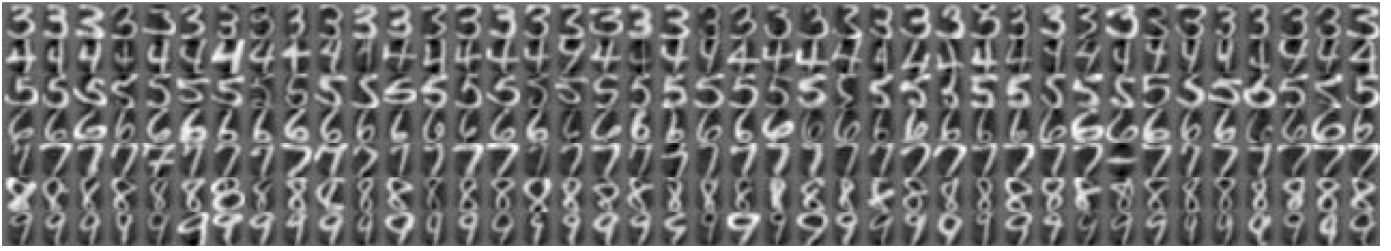


Fig. 4. Sample images from the USPS dataset.

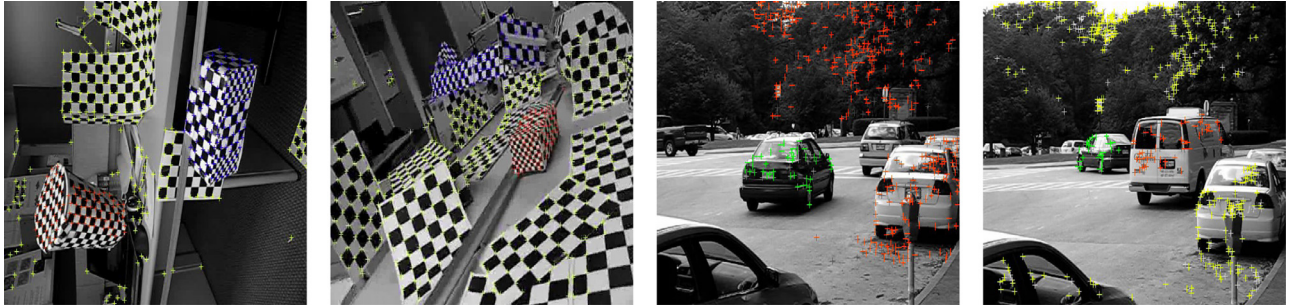


Fig. 5. Sample images from the Hopkins 155 dataset.

S-DCSC are shown in Fig. 3-(i-l), where the red, green and blue points correspond to the 1–64, 64–128, 129–192 rows of the matrix $\tilde{Q}$, respectively.

Ideally, the representation coefficients and the affinities of data points from different clusters (the entries off the diagonal blocks) should be zero, thus exhibits the discriminative property. Or in other words, the matrix Z and A should exhibit block diagonal structure, where the diagonal blocks correspond to the data within clusters and the entries off the diagonal blocks correspond to from different clusters. One can see that the matrix Z and A obtained by our method have very few nonzero entries off the diagonal blocks and the entries within the diagonal blocks dominate the matrix entries in amplitude, thus have better discriminative property. In comparison, the matrixes Z and A obtained by H-SSSC and S-SSSC have much more nonzero entries off the diagonal blocks, thus have weaker discriminative ability.

The row vectors of $\tilde{Q}$ are used by the K-means algorithm to obtain the clustering results and the data connection matrix P indicates which samples are finally segmented into the same cluster. It is important for the row vectors of $\tilde{Q}$ to exhibit good cluster property. One can see that the row vectors of $\tilde{Q}$ of our S-DCSC are well concentrated in three groups (corresponding to the three subjects) and there is little overlap between groups, thus result in exact clustering result. As shown in Fig. 3-(p), the matrix P of our S-DCSC exactly recovers the cluster membership of the data. Most rows of $\tilde{Q}$ of our H-DCSC are well concentrated in three groups, but there is a little overlap between the red group and the blue group. This causes that a few samples from the first subject are wrongly segmented into the third cluster, as shown by the matrix P in Fig. 3-(o). The row vectors of $\tilde{Q}$ obtained by SSSC have heavier overlaps so result in more wrong clustering results, as shown by the matrix P in Fig. 3-(m) and (n). The S-DCSC is better than the H-DCSC, so in the next experiments, we use the S-DCSC method.

4.2. Experiments on the USPS dataset

The USPS dataset consists of 9298 images of 10 subjects, corresponding to 10 handwritten digits 0–9, and each image has 16×16 pixels. Fig. 4 shows some sample images. We use the first 100 images of each digit in the experiments. To reduce computa-

Table 2

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

Methods	LRR	LSR1	LSR2	CASS	SMR	SSC	S-SSSC	S-DCSC
Subjects 10	26.90	42.90	25.30	18.00	11.10	10.10	8.20	6.90

tional cost and restoration requirement, we use the standard PCA to reduce the dimension 256-dim data into 40-dim. The squared Frobenius norm is used to measure the error matrix E . We perform the clustering experiment on all data points of all subjects. We compare our S-DCSC with LRR [30], LSR [33], CASS [34], SMR [43], SSC [29] and SSSC [44] on the USPS dataset. Experiments show that the S-DCSC method performs best with $\alpha = 0.12$, $\beta = 2.5$ and $\lambda = 0.2$. The clustering error rates are reported in Table 2. Our method outperforms other methods on the USPS dataset.

4.3. Experiments on the Hopkins 155 dataset

Motion segmentation refers to the problem of segmenting a video sequence with multiple rigidly moving objects into multiple spatiotemporal regions that correspond to the different motions in the scene (see some examples in Fig. 5). This problem is often solved by first extracting and tracking the spatial positions of a set of N feature points $\mathbf{x}_{f_i} \in \mathbb{R}^{2F}$ through each frame $f = 1, \dots, F$ of the video, and then clustering these feature point trajectories according to each one of the motions. Under the affine projection model, a feature point trajectory is formed by stacking the feature points $\mathbf{x}_{f_i}$ in the video as $\mathbf{y}_i = [\mathbf{x}_{1i}^T, \mathbf{x}_{2i}^T, \dots, \mathbf{x}_{Fi}^T]^T \in \mathbb{R}^{2F}$. Since the trajectories associated with a single rigid motion lie in an affine subspace (of $\mathbb{R}^{2F}$) of dimension at most 3, the trajectories of n rigid motions lie in a union of n low-dimensional subspaces of $\mathbb{R}^{2F}$. Therefore, the multi-view affine motion segmentation problem reduces to the subspace clustering problem.

We consider the Hopkins 155 dataset [59], which consists of 155 video sequences with 2 or 3 motions in each video corresponding to 2 or 3 low-dimensional subspaces. The squared Frobenius norm is used to measure the error matrix E . We compare our S-DCSC method with SSC [29], LSA [23], LRR [30], LSR [33], BDSSC [47], BDLRR [47], DTL-FSSC [56] and SSSC [44] on the Hopkins 155 motion segmentation dataset for the multi-view affine motion seg-

Table 3

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

Methods		LSA	LRR	BDLRR	LSR1	LSR2	BDSSC	SSC	DTL-FSSC	S-SSSC	S-DCSC
2 motions	Ave.	3.27	3.76	3.70	2.20	2.22	2.29	1.95	1.80	1.64	1.43
	ERR(%)	Med.	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std.	8.41	7.73	10.31	5.73	5.73	7.75	7.19	–	6.15	4.08
3 motions	Ave.	9.15	9.92	6.49	7.18	7.18	4.95	4.94	4.20	4.27	4.17
	ERR(%)	Med.	1.66	1.42	1.20	2.40	0.91	0.89	0.21	0.73	1.11
	Std.	14.58	11.33	12.32	8.96	8.86	9.72	9.91	–	8.97	6.68
Total	Ave.	4.60	5.15	4.33	3.31	3.34	2.89	2.63	–	2.20	2.04
	ERR(%)	Med.	0.69	0.00	0.00	0.22	0.23	0.00	–	0.00	0.00
	Std.	10.37	9.07	10.82	6.72	6.86	8.28	7.95	–	6.80	4.90

mentation without any other post processing (e.g., coefficients selection, thresholding, or ℓ_∞ normalization).

Experiments show our S-DCSC performs approximately best with $\alpha = 0.22$, $\beta = 870$ and $\lambda = 10$. Experimental results presented in Table 3 indicate that our S-DCSC performs better than other existing methods.

The experimental results motivate some insights into the subspace clustering: first, the unified methods: TDL-FSSC, SSSC and our DCSC generally outperform the two-stage methods. This confirms that combining the affinity learning and label inferring can actively affect each other and thus obtain higher clustering accuracy. This also shows that exploiting the discrimination of the affinity and the coherence of the labels helps clustering. More importantly, our DCSC method performs much better than SSSC and TDL-FSSC. This suggests that, combining the affinity and the labels more fully to guide each other so that both are discriminative between clusters and coherent within clusters is advantageous to subspace clustering.

5. Conclusion

In this work, we present a new regularity which combines the labels and the affinity so that they interactively enforce each other to have expected properties. Incorporating the new regularity into the SSSC model, we give a new unified optimization framework for subspace clustering. It enforces the coherence and discrimination of the affinity as well as the labels, thus called Discriminative and Coherent Subspace Clustering (DCSC). Extended experiments 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 grant nos. [61472303](#), [61271294](#), and the Fundamental Research Funds for the Central Universities under grant no. NSIY21 for supporting our research works.

References

- [1] L. Parsons, E. Hague, H. Liu, Subspace clustering for high dimensional data: a review, *ACM SIGKDD Explor. Newsl.* 6 (1) (2004) 90–105.
- [2] R. Basri, D. Jacobs, Lambertian reflectance and linear subspaces, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2) (2003) 218–233.
- [3] A.Y. Yang, J. Wright, Y. Ma, S. Sastry, Unsupervised segmentation of natural images via lossy data compression, *Comput. Vis. Image Underst.* 110 (2) (2008) 212–225.
- [4] R. Vidal, R. Tron, R. Hartley, Multiframe motion segmentation with missing data using powerfactorization and GPCA, *Int. J. Comput. Vis.* 79 (1) (2008) 85–105.
- [5] J. Ho, M.H. Yang, J. Lim, K.C. Lee, D. Kriegman, Clustering appearances of objects under varying illumination conditions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- [6] W. Hong, J. Wright, K. Huang, Y. Ma, Multi-scale hybrid linear models for lossy image representation, *IEEE Trans. Image Process.* 15 (12) (2006) 3655–3671.
- [7] R. Vidal, S. Soatto, Y. Ma, S. Sastry, An algebraic geometric approach to the identification of a class of linear hybrid systems, in: *Proceedings of the Conference on Decision and Control*, 2003, pp. 167–172.
- [8] P.S. Bradley, O.L. Mangasarian, K-plane clustering, *J. Glob. Optim.* 16 (1) (2000) 23–32.
- [9] P. Tseng, Nearest q-flat to m points, *J. Optim. Theory Appl.* 105 (1) (2000) 249–252.
- [10] T. Zhang, A. Szlam, G. Lerman, Median K-flats for hybrid linear modeling with many outliers, in: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2009, pp. 234–241.
- [11] P. Agarwal, N. Mustafa, K-means projective clustering, in: *Proceedings of the ACM Symposium on Principles of Database Systems*, 2004, pp. 155–165.
- [12] E.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.
- [13] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (12) (2005) 1–15.
- [14] Y. Ma, A.Y. Yang, H. Derksen, R. Fossum, Estimation of subspace arrangements with applications in modeling and segmenting mixed data, *SIAM Rev.* 50 (3) (2008) 413–458.
- [15] K. Huang, Y. Ma, R. Vidal, Minimum effective dimension for mixtures of subspaces: a robust GPCA, algorithm and its applications, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 631–638.
- [16] M.C. Tsakiris, R. Vidal, Algebraic clustering of affine subspaces, *Adv. Pure Math.* 05 (2) (2015) 62–70.
- [17] T. Boulton, L. Brown, Factorization-based segmentation of motions, in: *Proceedings of the IEEE Workshop on Motion Understanding*, 1991, pp. 179–186.
- [18] A. Leonardis, H. Bischof, J. Mayer, Multiple eigenspaces, *Pattern Recognit.* 35 (11) (2002) 2613–2627.
- [19] C. Archambeau, N. Delannay, M. Verleysen, Mixtures of robust probabilistic principal component analyzers, *Neurocomputing* 71 (7) (2008) 1274–1282.
- [20] A. Gruber, Y. Weiss, Multibody factorization with uncertainty and missing data using the EM algorithm, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2004, pp. 707–714.
- [21] 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.
- [22] A.Y. Yang, S. Rao, Y. Ma, Robust statistical estimation and segmentation of multiple subspaces, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2006, p. 99.
- [23] J. Yan, M. Pollefeys, A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate, in: *Proceedings of the European Conference on Computer Vision*, 2006, pp. 94–106.
- [24] A. Goh, R. Vidal, Segmenting motions of different types by unsupervised manifold clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.
- [25] 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.
- [26] G. Chen, G. Lerman, Spectral curvature clustering (SCC), *Int. J. Comput. Vis.* 81 (3) (2009) 317–330.
- [27] T. Zhang, A. Szlam, Y. Wang, G. Lerman, Hybrid linear modeling via local best-fit flats, *Int. J. Comput. Vis.* 100 (3) (2012) 217–240.
- [28] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.
- [29] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (11) (2013) 2765–2781.
- [30] 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.
- [31] P. Favaro, R. Vidal, A. Ravichandran, A closed form solution to robust subspace estimation and clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1801–1807.
- [32] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognit. Lett.* 43 (1) (2014) 47–61.

- [33] C.Y. Lu, H. Min, Z.Q. Zhao, L. Zhu, D.S. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, *Proceedings of the European Conference on Computer Vision*, 2012.
- [34] C. Lu, Z. Lin, S. Yan, Correlation adaptive subspace segmentation by trace lasso, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1345–1352.
- [35] C. Lu, S. Yan, Z. Lin, Correntropy induced l2 graph for robust subspace clustering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1801–1808.
- [36] Y.X. Wang, H. Xu, C. Leng, Provable subspace clustering: when LRR meets SSC, *Proceedings of the Neural Information Processing Systems*, 2013.
- [37] J. Chen, H. Zhang, H. Mao, Y. Sang, Z. Yi, Symmetric low-rank representation for subspace clustering, *Neurocomputing* 173 (P3) (2016) 1192–1202.
- [38] J. Wang, D. Shi, D. Cheng, Y. zhang, J. Gao, LRSR: low-rank-sparse representation for subspace clustering, *Neurocomputing* 214 (19) (2016) 1026–1037.
- [39] R. Heckel, H. Bolcskei, Robust subspace clustering via thresholding, *IEEE Trans. Inf. Theory* 61 (11) (2015) 6320–6342.
- [40] Y. Zhang, Z. Sun, R. He, T. Tan, Robust subspace clustering via half-quadratic minimization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3096–3103.
- [41] D. Park, C. Caramanis, S. Sanghavi, Greedy subspace clustering, in: *Proceedings of the Neural Information Processing Systems*, 2014, pp. 2753–2761.
- [42] B. Li, Y. Zhang, Z. Lin, H. Lu, Subspace clustering by mixture of Gaussian regression, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2094–2102.
- [43] H. Hu, Z. Lin, J. Feng, J. Zhou, Smooth representation clustering, in: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2014, pp. 3834–3841.
- [44] C.G. Li, C. You, R. Vidal, Structured sparse subspace clustering: a joint affinity learning and subspace clustering framework, *IEEE Trans. Image Process.* 26 (2017) 2988–3001.
- [45] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (8) (2000) 888–905.
- [46] G. Liu, S. Yan, Latent low-rank representation for subspace segmentation and feature extraction, in: *Proceedings of the International Conference on Computer Vision*, IEEE Computer Society, 2011, pp. 1615–1622.
- [47] J. Feng, Z. Lin, H. Xu, S. Yan, Robust subspace segmentation with block-diagonal prior, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3818–3825.
- [48] E.L. Dyer, A.C. Sankaranarayanan, R.G. Baraniuk, Greedy feature selection for subspace clustering, *J. Mach. Learn. Res.* 14 (1) (2013) 2487–2517.
- [49] D. Pham, S. Budhaditya, D. Phung, S. Venkatesh, Improved subspace clustering via exploitation of spatial constraints, in: *Proceedings of the Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 550–557.
- [50] V.M. Patel, H.V. Nguyen, R. Vidal, Latent space sparse subspace clustering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 225–232.
- [51] V.M. Patel, R. Vidal, Kernel sparse subspace clustering, in: *Proceedings of the International Conference on Image Processing*, 2014, pp. 2790–2797.
- [52] 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.
- [53] Z. Lin, M. Chen, L. Wu, Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, 9, 2010, Eprint ArXiv:1009.5055v2.
- [54] C.G. Li, R. Vidal, A structured sparse plus structured low-rank framework for subspace clustering and completion, *IEEE Trans. Signal Process.* 64 (24) (2016) 6557–6570.
- [55] M. Yin, J. Gao, Z. Lin, Laplacian regularized low-rank representation and its applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (3) (2016) 504.
- [56] Z. Wen, B. Hou, W. Qian, L. Jiao, Discriminative transformation learning for fuzzy sparse subspace clustering, *IEEE Trans. Cybern. PP* (99) (2017) 1–14.
- [57] 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.
- [58] J.J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Anal. Mach. Intell.* 16 (5) (1994) 550–554.
- [59] R. Tron, R. Vidal, A benchmark for the comparison of 3-d motion segmentation algorithms, in: *Proceedings of the 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.



Ruiqiang He received the M.S. degree from Xi'an University of Architecture and Technology, Xi'an, China, in 2009. Currently he is pursuing his Ph.D. degree at the School of Mathematics and Statistics, Xidian University. His current research interests include inverse problem in image processing, computer vision and machine learning.