Spectral Embedded Clustering: A Framework for In-Sample and Out-of-Sample Spectral Clustering

Feiping Nie, Zinan Zeng, Ivor W. Tsang, Dong Xu, *Member, IEEE*, and Changshui Zhang, *Member, IEEE*

***Abstract*— Spectral clustering (SC) methods have been successfully applied to many real-world applications. The success of these SC methods is largely based on the *manifold assumption*, namely, that two nearby data points in the high-density region of a low-dimensional data manifold have the same cluster label. However, such an assumption might not always hold on high-dimensional data. When the data do not exhibit a clear low-dimensional manifold structure (e.g., high-dimensional and sparse data), the clustering performance of SC will be degraded and become even worse than K-means clustering. In this paper, motivated by the observation that the true cluster assignment matrix for high-dimensional data can be always embedded in a linear space spanned by the data, we propose the spectral embedded clustering (SEC) framework, in which a linearity regularization is explicitly added into the objective function of SC methods. More importantly, the proposed SEC framework can naturally deal with out-of-sample data. We also present a new Laplacian matrix constructed from a local regression of each pattern and incorporate it into our SEC framework to capture both local and global discriminative information for clustering. Comprehensive experiments on eight real-world high-dimensional datasets demonstrate the effectiveness and advantages of our SEC framework over existing SC methods and K-means-based clustering methods. Our SEC framework significantly outperforms SC using the Nyström algorithm on unseen data.**

***Index Terms*— Linearity regularization, out-of-sample clustering, spectral clustering, spectral embedded clustering.**

I. Introduction

CLUSTERING is one of the fundamental topics in machine learning and data mining. It has been widely used in various domains ranging from engineering and science to economics. The primary goal of clustering is to group similar patterns into the same cluster, and discover the meaningful structure of the data [1]. Over the past decades, a large family of clustering algorithms such as K-means clustering and mixture models [2]–[4] have been widely studied. Self-organizing map methods [5], [6] have also been used to reveal and visualize the cluster structures of data. Moreover, clustering methods such as kernel-based clustering [7], [8], spectral clustering (SC) [9]–[13] and support vector clustering [14] have been developed to capture nonlinear cluster structures. Recently, maximum margin clustering methods [15]–[17] have been proposed to find a decision boundary in some low-density region of the data that separates data points into two opposite clusters. Structured outputs and loss functions have also been incorporated into the design of clustering algorithms [18].

Although many clustering methods have been proposed, partitioning high-dimensional data points into their relevant clusters remains one of major challenges for clustering. For instance, K-means clustering iteratively assigns each data point to the cluster with the closest center based on some distance/similarity measurements and updates the center of each cluster. However, the estimated distance/similarity measures on high-dimensional data may not be accurate, resulting in degraded clustering performances of K-means clustering. In practice, many high-dimensional data may exhibit dense grouping in a low-dimensional subspace. Hence, researchers usually resort to projecting the high-dimensional data onto the low-dimensional subspace via some dimension reduction techniques such as principal component analysis before performing cluster analysis. Several works have been proposed to perform K-means clustering and dimension reduction iteratively for high-dimensional data [19]–[21]. Recently, Ye *et al.* [22] proposed discriminative K-means (DisKmeans) clustering which unifies the iterative procedure of dimension reduction and K-means clustering into a trace maximization problem. However, DisKmeans clustering does not consider the local geometry structure (a.k.a. low-dimensional manifold) of the data [23], [24].

The use of such manifold information in SC has brought the state-of-the-art clustering performance in many high-dimensional applications, such as image segmentation [9], [10]. The basic idea of SC is to find a cluster assignment of the data points by using the spectrum of the similarity matrix that captures the nonlinear and low dimensional manifold structure of the data. Moreover, the variants of the SC methods have demonstrated many interesting properties for clustering. For instance, normalized cuts can balance the volume of clusters using data density information [9]. Self-tuning SC can learn

Manuscript received June 22, 2010; accepted June 27, 2011. Date of publication September 29, 2011; date of current version November 2, 2011. This work was supported in part by the Singapore National Research Foundation Interactive Digital Media R&D Program under Grant NRF2008IDM-IDM004-018, the China State Key Science and Technology Project on Marine Carbonate Reservoir Characterization under Project 2011ZX05004-003, and the National Natural Science Foundation of China under Grant 61021063 and Grant 60835002.

F. Nie was with the Department of Automation, Tsinghua University, Beijing 100084, China. He is now with the University of Texas, Arlington, TX 76019 USA (e-mail: feipingnie@gmail.com).

Z. Zeng, I. W. Tsang, and D. Xu are with the School of Computer Engineering, Nanyang Technological University, 639798, Singapore (e-mail: znzeng@ntu.edu.sg; ivortsang@ntu.edu.sg; dongxu@ntu.edu.sg).

C. Zhang is with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: zcs@mail.tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2011.2162000

the parameters automatically in an unsupervised setting [25]. Note that SC heavily relies on the manifold assumption [24], namely, that two nearby data points in the high-density region of a low-dimensional data manifold have the same cluster label. However, for high-dimensional and sparse data, nearest neighbors may actually still be far away from each other due to the bias introduced by the curse of dimensionality. Hence, the similarity matrix of the data could not effectively reflect an evident low-dimensional manifold structure (i.e., the manifold assumption does not hold here), and the clustering performance of SC would be degraded dramatically.

In addition, traditional SC methods usually do not provide a natural extension to cope with out-of-sample data points. Several methods [11], [26], [27] have been proposed to address this issue. For example, the methods described in [11] and [26] respectively used some heuristics and the Nyström method to approximate the implicit eigenfunction for the new data points. However, its clustering performance heavily depends on the approximation quality of the affinity matrix A defined between the in-sample and new data points. Apart from approximation methods, Alzate and Suykens [27] recently proposed to use an error correcting output code (ECOC) approach [28], [29] for SC with out-of-sample extension. First, the row of the eigenvectors corresponding to each training data point is binarized into an encoding vector. Then, one can count the occurrences of different encoding vectors and find the k encoding vectors with the most occurrences, which would be used to form the code set. Based on this code set, the data points are partitioned into different clusters by Hamming distance. Similarly, for a new data point, its projection is binarized. Its cluster label is then assigned to that of the closest code in the code set according to the Hamming distance. However, as discussed in [30], the prediction performance highly depends on the design of error correcting output codes, which is a nontrivial task. To achieve the optimal prediction performance, a specific coding should be learned for different datasets [30].

To enhance the clustering performance of SC on high-dimensional datasets, in this paper, we propose a novel spectral embedded clustering (SEC) framework for high-dimensional data, which considers the underlying dense grouping structure of data in a low-dimensional subspace, and explicitly incorporates this prior knowledge into different variants of SC methods. Our main contributions include the following.

1) First, we prove that the eigenvector corresponding to the trivial eigenvalue of the Laplacian matrix should not be discarded when the spectral rotation method is adopted to obtain the final cluster assignment matrix.
2) More importantly, we prove that the cluster assignment matrix of the data can be embedded in a linear space spanned by the data, when the dimensionality of data is high enough [31].
3) Based on this observation on high-dimensional data, we propose the SEC, in which a linearity regularization is explicitly imposed on the objective function of the clustering methods in order to control the mismatch between the cluster assignment matrix and the low-dimensional embedding of the data.
4) In order to handle the case in which the data do not exhibit a clear manifold structure, we further propose a Laplacian matrix in local regression in place of the widely used Laplacian matrix to reflect the local geometry structure of the data. By using the newly proposed Laplacian matrix, we can capture more locally discriminative information.
5) We also theoretically discuss the connection between our proposed SEC methods and other clustering methods from a new perspective. Hence, we can unify variants of SC algorithms, K-means, and the recently proposed DisKmeans clustering methods into our proposed SEC framework. In addition, we can naturally deal with the out-of-sample data for the clustering methods under our proposed SEC framework.
6) Comprehensive experiments on eight real-world high-dimensional datasets demonstrate that the proposed framework outperforms the existing SC methods and K-means related clustering methods for in-sample clustering. The experiments also show the superior performance of the proposed SEC framework over the Nyström method and better generalization capability than the K-means-based clustering methods for out-of-sample clustering.

The rest of this paper is organized as follows. Section II first reviews SC and the cluster assignment methods. Our proposed SEC framework is then presented in Section III. Connections to other clustering methods are discussed in Section IV. A natural mechanism to cope with the out-of-sample data under the proposed SEC framework is presented in Section V. Experimental results on benchmark datasets are reported in Section VI and the concluding remarks are given in Section VII.

II. SC Revisited

Given a dataset $\mathcal{X} = \{x_i\}_{i=1}^n$, the main task of clustering is to partition $\mathcal{X}$ into c clusters. Denote the cluster assignment matrix by $Y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{B}^{n\times c}$, where $y_i \in \mathbb{B}^{c\times 1}(1 \le i \le n)$ is the cluster assignment vector for the pattern x_i. The jth element of y_i is 1 if the pattern x_i is assigned to the jth cluster, it is 0 otherwise. In addition, there is one and only one element being 1 in y_i. Clustering is a nontrivial problem because Y is constrained to be an integer solution. In this section, we first revisit the SC method and the corresponding techniques (i.e., K-means clustering and spectral rotation) to obtain the discrete cluster assignment matrix. We then prove that the eigenvector corresponding to the smallest eigenvalue should not be discarded when spectral rotation is used to obtain the discrete clustering assignment.

A. SC

From the last decade, SC, in which a weighted graph is used to partition the data, has attracted much attention. Several algorithms have been proposed in the literature [9], [10], [32], [33]. Here, we focus on the SC algorithm with k-way normalized cuts [10].

Let us denote $\mathcal{G} = \{\mathcal{X}, A\}$ as an undirected weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $A \in \mathbb{R}^{n\times n}$, in which each entry A_{ij} of the symmetric matrix A represents the affinity of a pair of vertices of the weighted graph. A common choice of A_{ij} is defined by

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) & x_i \text{ and } x_j \text{ are neighbors,} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where σ is the parameter to control the spread of neighbors. The Laplacian graph L is then defined by $L = D - A$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j A_{ij}, \forall i$. Let us denote $tr(A)$ as the trace operator of a matrix A. The minimization of the normalized cut criterion can be transformed into the following maximization problem [10]:

$$\max_{Z^T DZ = I_c} tr\left(Z^T AZ\right) \quad (2)$$

where $Z = Y(Y^T DY)^{-1/2}$ and I_c denotes the identity matrix of size c by c. Let us define a scaled cluster assignment matrix $\bar{F}$ by

$$\bar{F} = D^{1/2}Z = D^{1/2}Y\left(Y^T DY\right)^{-\frac{1}{2}} = f(Y). \quad (3)$$

Then the objective function (2) can be rewritten as

$$\max_{\bar{F}^T\bar{F} = I_c} tr\left(\bar{F}^T D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \bar{F}\right). \quad (4)$$

Note that the elements of $\bar{F}$ are constrained to be discrete values, which makes (4) hard to solve. A well-known solution to this problem is to relax the matrix $\bar{F}$ from the discrete values to continuous ones. Then the problem becomes

$$\max_{F^T F = I_c} tr\left(F^T KF\right) \quad (5)$$

where $K = D^{-1/2}AD^{-1/2}$ and $F \in \mathbb{R}^{n\times c}$. The optimal solution F of (5) can be obtained by eigenvalue decomposition of the matrix K.

B. Cluster Assignment Methods

With the relaxed continuous solution $F \in \mathbb{R}^{n\times c}$ from eigenvalue decomposition of (5), K-means clustering or spectral rotation can be used to calculate the discrete solution $Y \in \mathbb{B}^{n\times c}$.

1) K-Means Clustering: The input to K-means clustering is n data points, in which the ith data point is the ith row of F. The standard K-means clustering algorithm is performed to obtain the discrete-valued cluster assignment for each pattern. This technique is also used in [32] for assigning cluster labels.

2) Spectral Rotation: Note that the global optimal F of the optimization problem (5) is not unique. Let $F^* \in \mathbb{R}^{n\times c}$ be the matrix whose columns consist of the top c eigenvectors of K and $R \in \mathbb{R}^{c\times c}$ be an orthogonal matrix. Then F can be replaced by F^*R for any orthogonal matrix R. To obtain the final clustering result, we need to find a discrete-valued cluster assignment matrix which is closed to F^*R. The work in [10] also defined a mapping to obtain the corresponding Y^*

$$Y^* = f^{-1}\left(F^*\right) = \text{Diag}\left(F^*F^{*T}\right)^{-\frac{1}{2}} F^* \quad (6)$$

where f^{-1} denotes a reverse mapping function from a continuous solution space to a discrete one, and Diag(M) denotes a diagonal matrix with the same size and the same diagonal elements as the square matrix M. The mapping of (6) is to normalize each row of F^* such that its L2-norm is equal to one. It can be easily verified that $f^{-1}(F^*R) = Y^*R$.

As F^*R is the optimal solution to the relaxed (5) for an arbitrary orthogonal matrix R, a suitable R should be selected such that Y^*R is closest to a discrete cluster assignment matrix Y. The optimal R and Y are then obtained by solving the following optimization problem [10]:

$$\min_{Y\in\mathbb{B}^{n\times c}, R\in\mathbb{R}^{c\times c}} \|Y - Y^*R\|^2 \quad \text{s.t.} \quad Y\mathbf{1}_c = \mathbf{1}_n, \; R^T R = I_c \quad (7)$$

where $\mathbf{1}_c$ and $\mathbf{1}_n$ denote the $c \times 1$ and $n \times 1$ vectors of all 1s, respectively. Yu and Shi [10] used this technique to obtain the cluster assignment matrix by iteratively solving Y and R.

C. Eigenvectors of K for Cluster Assignment

In SC, the eigenvector corresponding to the largest eigenvalue of K is $D^{1/2}\mathbf{1}_n$, which is commonly considered as a trivial solution, and thus is discarded in the cluster assignment in two-cluster problems. However, we focus on k-way normalized cut in this paper, and two cluster assignment methods including K-means clustering and spectral rotation are used. When we use the K-means clustering method to obtain the discrete solution, whether this eigenvector is discarded or not will have little change to the final clustering result. However, based on the following proposition, we know that this eigenvector should not be discarded when the spectral rotation method is used to obtain the discrete solution.

Proposition 1: When the spectral rotation method is used to obtain the discrete solution of k-way normalized cuts with the normalized Laplacian matrix $\tilde{L} = D^{-1/2}LD^{-1/2}$, discarding the eigenvector $D^{1/2}\mathbf{1}_n$ will change the eigenspace and hence will lead to a different solution to the optimization problem in (4).

Proof: Suppose the eigenvector $D^{1/2}\mathbf{1}_n$ is discarded in the columns of the continuous solution F^*, then we have

$$\mathbf{1}_n^T D^{1/2} F^* = \mathbf{0}.$$

While for the discrete solution $\bar{F}$, from (3), we have

$$\mathbf{1}_n^T D^{1/2}\bar{F} = \mathbf{1}_n^T DY\left(Y^T DY\right)^{-\frac{1}{2}}.$$

Note that D is a diagonal matrix and Y is a cluster assignment matrix whose row has one and only one element being 1 and others being 0. Therefore, for any orthogonal matrix R, we have $\mathbf{1}_n^T D^{1/2}F^*R = \mathbf{0}$, but $\mathbf{1}_n^T D^{1/2}\bar{F} \neq \mathbf{0}$, which indicates that the continuous solution F^* cannot precisely approximate the discrete solution $\bar{F}$ with any orthogonal matrix R. ■

In contrast, if we preserve the eigenvector $D^{1/2}\mathbf{1}_n$ in the columns of the continuous solution F^*, then $\mathbf{1}_n^T D^{1/2}F^*R$ might be equal to $\mathbf{1}_n^T D^{1/2}\bar{F}$ with an appropriate orthogonal matrix R. Based on the above analysis, the eigenvector corresponding to the largest eigenvalue of K should not be discarded when we use spectral rotation to obtain the discrete solution.

III. General Framework for SEC

Many applications such as natural language processing, text mining, and bioinformatics need to deal with high-dimensional data, while cluster analysis for such high-dimensional data is still a challenging task, and the classical K-means clustering method might not perform well. Recently, SC methods have been successfully applied to some applications with high-dimensional data. As discussed in Section I, the success of SC methods greatly depends on the manifold assumption and the choice of affinity matrix A, which represents the low-dimensional manifold structure. For some high-dimensional data that exhibit a clear manifold structure in a low-dimensional space, SC methods could outperform K-means clustering methods. However, such a manifold assumption might not always hold for all problems with high-dimensional data. Moreover, the common choice of the affinity matrix A in (1) may not clearly reflect the local geometry structure of the data. In those cases, the clustering performance of SC may be even worse than the K-means clustering.

In the following subsections, we will first introduce an observation on the high-dimensional data that the ground truth cluster assignment matrix Y can always be embedded in a linear space spanned by the data to be clustered. Based on this prior, we then explicitly add this linearity regularization into clustering framework, which is referred to as SEC.

A. Low-Dimensional Embedding for Cluster Assignment Matrix

Denote the data matrix by $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$. For simplicity, we assume the data are centered, i.e., $X\mathbf{1}_n = \mathbf{0}$. Let us define the total scatter matrix S_t, the between-cluster scatter matrix S_b, and the within-cluster scatter matrix S_w as

$$S_t = XX^T \tag{8}$$

$$S_b = XGG^TX^T \tag{9}$$

$$S_w = XX^T - XGG^TX^T \tag{10}$$

where G is a weighted cluster assignment matrix defined by

$$G = Y\left(Y^TY\right)^{-\frac{1}{2}} \tag{11}$$

and Y is defined as in Section II. It is easy to verify that $G^TG = I_c$.

We have the following theorem on the cluster assignment matrix Y, which is the foundation of the proposed SEC framework.

Theorem 1: If $rank(S_b) = c - 1$ and $rank(S_t) = rank(S_w) + rank(S_b)$, then the true cluster assignment matrix can be represented by a low-dimensional linear mapping of the data, that is, there exist $W \in \mathbb{R}^{d\times c}$ and $b \in \mathbb{R}^{c\times 1}$ such that $Y = X^TW + \mathbf{1}_nb^T$.

The proof can be found in Appendix. As noted in [34], the conditions in Theorem 1 are usually satisfied for the high-dimensional and small-sample-size problem, which is usually the case in many real-world applications.

B. SEC

Many clustering methods can be reduced to minimize the following objective function:

$$\min_{F^TF=I_c} \mathcal{J}(F). \tag{12}$$

For example, in SC, the optimization problem (5) is equivalent to minimizing the objective function as follows:

$$\mathcal{J}(F) = tr\left(F^T\tilde{L}F\right)$$

where $\tilde{L} = D^{-(1/2)}LD^{-(1/2)} = I_n - D^{-(1/2)}AD^{-(1/2)}$ is the normalized Laplacian matrix.

According to Theorem 1, the true cluster assignment matrix Y can be always embedded into a linear mapping of data in many applications with high-dimensional data. In this paper, with the prior on the linearity property of Y, we propose the SEC framework by incorporating the linearity regularization into the SC methods. Specifically, we minimize the following objective function:

$$\min_{\substack{F,W,b\\F^TF=I_c}} \mathcal{J}(F) + \mu\left(\left\|X^TW + \mathbf{1}_nb^T - F\right\|^2 + \gamma_g tr\left(W^TW\right)\right) \tag{13}$$

where μ and γ_g are two regularization parameters and the second term characterizes the mismatch between the relaxed cluster assignment matrix F and the low-dimensional representation of the data.

Note that the data are centered, i.e., $X\mathbf{1}_n = \mathbf{0}$. By setting the derivatives of the objective function with respect to b and W to zeros, we have

$$\begin{cases} b = \frac{1}{n}F^T\mathbf{1}_n, \\ W = \left(XX^T + \gamma_g I_d\right)^{-1}XF. \end{cases} \tag{14}$$

By substituting W and b in (13) by (14), the optimization problem (13) becomes

$$\min_{F^TF=I_c} \mathcal{J}(F) + \mu\mathcal{R}(F) \tag{15}$$

where

$$\mathcal{R}(F) = tr\left(F^TL_gF\right) \tag{16}$$

$$L_g = H_n - X^T\left(XX^T + \gamma_g I_d\right)^{-1}X \tag{17}$$

and $H_n = I_n - 1/n\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix.

Analogous to SC, the cluster assignment matrix of the proposed SEC framework can be relaxed as the eigenvectors of $\tilde{L} + \mu L_g$ corresponding to the c smallest eigenvalues. Based on Proposition 1, all these c eigenvectors should be kept if the spectral rotation is used to find the final cluster assignment matrix.

C. Design of Laplacian Matrix Using Local Regression

As discussed in Section III-B, the objective function of SC methods can be expressed in the form of $tr(F^TLF)$ or $tr(F^T\tilde{L}F)$, where L and $\tilde{L}$, respectively, are the unnormalized and normalized Laplacian matrix defined based upon the affinity matrix A in (1). However, these Laplacian matrices

cannot capture the discriminative information of the clusters for data that do not exhibit an evident manifold structure. Recall that the regularization term $\mathcal{R}(F) = tr(F^T L_g F)$ in (16) is derived from the global regression

$$\min_{\substack{F,W,b \\ F^T F = I_c}} \left\| X^T W + \mathbf{1}_n b^T - F \right\|^2 + \gamma_g tr\left(W^T W\right) \tag{18}$$

where W and b capture the globally discriminative directions of each cluster. In order to capture the locally discriminative information without the manifold assumption, we can use a local regression function for each pattern

$$\min_{\substack{F_i,W_i,b_i \\ F^T F = I_c}} \sum_{i=1}^{n} \left(\left\| X_i^T W_i + \mathbf{1}_k b_i^T - F_i \right\|^2 + \gamma_l tr\left(W_i^T W_i\right) \right) \tag{19}$$

where $X_i = [x_{i_1}, x_{i_2}, \ldots, x_{i_k}]$, $x_{i_j}|_{j=1}^k$ are the k nearest neighbors of x_i, and $F_i = [f_{i_1}, f_{i_2}, \ldots, f_{i_k}]^T$ with $f_{i_j}|_{j=1}^k$ being the transpose of the i_jth row of F.

To obtain the optimal solution to (19), we set the derivatives of the objective function with respect to b_i and W_i to zeros. Then we have

$$\begin{cases} b_i = \frac{1}{k} F_i^T \mathbf{1}_k \\ W_i = \left(X_i H_k X_i^T + \gamma_l I_d\right)^{-1} X_i H_k F_i \end{cases} \tag{20}$$

where $H_k = I_k - 1/k \mathbf{1}_k \mathbf{1}_k^T$ is the local centering matrix.

By substituting W_i and b_i in (19) by (20), (19) is then reduced to $tr(F^T L_l F)$, where

$$L_l = [S_1, \ldots, S_n] \begin{bmatrix} L_{l_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_{l_n} \end{bmatrix} [S_1, \ldots, S_n]^T. \tag{21}$$

In (21), $L_{l_i}|_{i=1}^n = H_k - X_i^T (X_i X_i^T + \gamma_l I_d)^{-1} X_i$ and $S_i|_{i=1}^n$ is a selection matrix such that $F_i = S_i^T F$. Notice that L_l is also in the form of Laplacian matrix, which captures locally discriminative information of the data. Moreover, this Laplacian matrix L_l is similar to that in [35] which is based on local discriminative analysis, however, L_l is derived from local regression. Thereafter, we can replace L or $\tilde{L}$ in (15) by L_l, and the overall optimization problem (15) becomes

$$\min_{F^T F = I_c} tr\left(F^T (L_l + \mu L_g) F\right). \tag{22}$$

Again, the global optimal solution F^* to (22) can be obtained by eigenvalue decomposition. The columns of F^* are from the top c eigenvectors of the matrix $L_l + \mu L_g$. Based on F^*, the discrete-valued cluster assignment matrix can be obtained by K-means clustering or spectral rotation. The details of the proposed SEC are outlined in Algorithm 1. Recall that SEC using the Laplacian matrix L_l, which is constructed from local regression, can capture both locally and globally discriminative information of clusters from L_l and L_g, respectively. For better presentation, in the rest of this paper, we refer to the SEC algorithm using the standard Laplacian matrix in SC and the Laplacian matrix in local regression as SEC/SC and SEC/LR, respectively. Similarly, we can also apply any Laplacian matrix discussed in [36] and [37] to our proposed SEC framework.

Algorithm 1 SEC Algorithm

Given a sample set $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and the number of clusters c.

1: Compute the matrix $\tilde{L} + \mu L_g$ or $L_l + \mu L_g$.
2: Solve (15) or (22) with eigenvalue decomposition and obtain the optimal F^*.
3: Based on F^*, compute the discrete cluster assignment matrix Y by using K-means clustering or spectral rotation.

IV. Connections to Prior Work

Though the connection between classical K-means clustering and SC has been discussed in [38] and [39] from the viewpoint of weighted kernel K-means clustering, in this section we will explore the connection between our proposed SEC framework and SC, K-means clustering, the recently proposed discriminative K-means clustering [22], and clustering with local and global regularization (CLGR) [37] from a new perspective. Particularly, we can unify the aforementioned clustering methods into our proposed SEC framework. Hereafter, we will discuss how to naturally cater for the out-of-sample extension to the clustering methods under our proposed SEC framework in Section V.

A. Connection Between SEC and SC

In the local regression model (19), when $k \to n$, the local regression model is reduced to the global regression model. Thus, the constructed Laplacian matrix L_l, which is based on local regression, is a local version of the Laplacian matrix L_g, which is based on global regression. In other words, L_l is constructed based on local information of the data, and the Laplacian matrix L_g is constructed based on global information of the data.

The normalized Laplacian matrix $\tilde{L}$ used in SC also captures the local information of the data, so when $\mu = 0$ in SEC, SEC reduces to a variant of SC. On the other hand, when μ is set to an appropriate positive value, SEC can simultaneously capture the local and global structure information of the data and thus obtain a better clustering result.

B. Connection Between SEC and CLGR

Recently, Wang *et al*. proposed CLGR [37], which solves the following problem:

$$\min_{F^T F = I_c} tr\left(F^T (L + \eta L_o) F\right) \tag{23}$$

where L_o is another Laplacian matrix constructed using local learning method [40] and η is a tradeoff parameter.

Let us denote the cluster assignment matrix $F = [f_1, \ldots, f_n]^T \in \mathbb{R}^{n \times c}$. We also define the k nearest neighbors of x_i as $\mathcal{N}(x_i) = \{x_{i_1}, \ldots, x_{i_k}\}$, and $F_i = [f_{i_1}, \ldots, f_{i_k}]^T \in \mathbb{R}^{k \times c}$. In local learning regularization, for each x_i, a locally linear projection matrix $W_i \in \mathbb{R}^{d \times c}$ is learned by minimizing the following structural risk functional [37]:

$$\min_{W_i} \sum_{x_j \in \mathcal{N}(x_i)} \left\| W_i^T x_j - f_j \right\|^2 + \gamma\, tr\left(W_i^T W_i\right).$$

One can obtain the closed-form solution for W_i

$$W_i = \left(X_i X_i^T + \gamma I_d\right)^{-1} X_i F_i. \tag{24}$$

After all the locally linear projection matrices are learned, the cluster assignment matrix F can be found by minimizing the following criterion:

$$\mathcal{J}(F) = \sum\nolimits_{i=1}^{n} \left\| x_i^T W_i - f_i^T \right\|^2. \tag{25}$$

Substituting (24) back into (25), we have

$$\mathcal{J}(F) = tr\left(F^T (N - I_n)^T (N - I_n) F\right) = tr\left(F^T L_o F\right)$$

where $L_o = (N - I_n)^T (N - I_n)$ and $N \in \mathbb{R}^{n \times n}$ with its (i, j)th entry as

$$N_{ij} = \begin{cases} a_h^i, & \text{if } x_j \text{ is the } h\text{th nearest neighbors of } x_i, \\ 0, & \text{otherwise} \end{cases}$$

in which a_h^i denotes the hth entry of $a^i = x_i^T \left(X_i X_i^T + \gamma I_d\right)^{-1} X_i$.

One can observe that both L and L_o capture local information only, and $L + \eta L_o$ in (23) is also a Laplacian matrix, hence CLGR is just a variant of SC. Therefore, SEC reduces to CLGR when $\mu = 0$ and L_l is replaced by $L + \eta L_o$ in (22).

C. Connection Between SEC and K-Means Clustering

K-means clustering is a simple and frequently used clustering algorithm. As shown in [41], the objective function of K-means clustering is to minimize the following criterion:

$$\min_{G^T G = I_c} tr(S_w) = \min_{G^T G = I_c} tr\left(XX^T - XGG^T X^T\right) \tag{26}$$

where G is defined as in (11). Problem (26) is simplified as the following problem:

$$\max_{G^T G = I_c} tr\left(G^T X^T X G\right). \tag{27}$$

Traditional K-means clustering uses an EM-like iterative method to solve the above problem. Spectral relaxation can also be used to solve the K-means clustering problem [41].

We will prove that the objective function of the proposed SEC in (22) reduces to that of K-means clustering when $\mu \to \infty$ and $\gamma_g \to \infty$. When $\mu \to \infty$, the optimization problem of SEC in (22) becomes

$$\min_{F^T F = I_c} tr\left(F^T \left(H_n - X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) F\right) \tag{28}$$

which is equivalent to the following problem:

$$\max_{F^T F = I_c} tr\left(F^T \left(\frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T + X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) F\right). \tag{29}$$

The solution to this problem can be reduced to calculate the eigenvectors corresponding to the top eigenvalues of the matrix $(1/n)\mathbf{1}_n \mathbf{1}_n^T + X^T (XX^T + \gamma_g I_d)^{-1} X$. Note that $\mathbf{1}_n$ is the eigenvector corresponding to the largest eigenvalue, and the other eigenvectors are orthogonal to $\mathbf{1}_n$. Therefore, calculating the other top eigenvectors is equivalent to solving the following problem:

$$\begin{aligned} &\max_{\substack{F^T F = I_c \\ F^T \mathbf{1}_n = \mathbf{0}}} tr\left(F^T \left(\frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T + X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) F\right) \\ \Leftrightarrow &\max_{\substack{F^T F = I_c \\ F^T \mathbf{1}_n = \mathbf{0}}} tr\left(F^T \left(X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) F\right) \\ \Leftrightarrow &\max_{F^T F = I_c} tr\left(F^T \left(X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) F\right). \end{aligned} \tag{30}$$

When $\gamma_g \to \infty$, the optimization problem in (30) reduces to the optimization problem in (27). Therefore, the objective function of SEC reduces to that of K-means clustering algorithm if $\mu \to \infty$ and $\gamma_g \to \infty$.

D. Connection Between SEC and Discriminative K-Means Clustering

Subspace clustering methods were proposed to learn the low-dimensional subspace and data cluster simultaneously [42], [43], possibly because high-dimensional data may exhibit dense grouping in a low-dimensional space. For instance, discriminative clustering methods solve the following optimization problem:

$$\max_{W,G} tr\left(\left(W^T \left(S_t + \gamma_g I_d\right) W\right)^{-1} W^T S_b W\right) \tag{31}$$

where S_t and S_b are defined in (8) and (9), respectively.

There are two sets of variables, namely, the projection matrix W and the scaled cluster assignment matrix G, in (31). Most of the existing works optimize W and G iteratively [19]–[21]. However, a recent work, discriminative K-means [22], simplified (31) by optimizing G only, which is based on the following observation [44]:

$$\begin{aligned} &tr\left(\left(W^T \left(S_t + \gamma_g I_d\right) W\right)^{-1} W^T S_b W\right) \\ &\quad \le tr\left(\left(S_t + \gamma_g I_d\right)^{-1} S_b\right) \end{aligned} \tag{32}$$

where the equality holds when $W = VM$, and V is composed of the eigenvectors of $(S_t + \gamma_g I_d)^{-1} S_b$ corresponding to all the nonzero eigenvalues, with M as an arbitrary nonsingular matrix.

Based on (32), the optimization problem (31) can be simplified as

$$\max_{G} tr\left(\left(S_t + \gamma_g I_d\right)^{-1} S_b\right). \tag{33}$$

By substituting (8) and (9) into (33) and adding the constraint $G^T G = I_c$ in (33), we arrive at

$$\max_{G^T G = I_c} tr\left(G^T \left(X^T \left(XX^T + \gamma_g I_d\right)^{-1} X\right) G\right) \tag{34}$$

which is exactly the same as in (30). In other words, when $\mu \to \infty$ in SEC and the spectral relaxation is used to solve the cluster assignment matrix in the discriminative K-means clustering algorithm (referred to as DisKmeans), SEC reduces to DisKmeans.

V. Clustering for Out-of-Sample Data

Though Nyström-based methods [11], [26] and the ECOC method [27] have been proposed to provide an out-of-sample extension for SC, as discussed in Section I, the clustering performance of these approaches critically depends on either the approximation quality of the affinity matrix A or the nontrivial design of error correcting output codes [30], both of which are beyond the scope of this paper. Instead of using approximation methods or ECOC approach, in this section, we will show that our proposed SEC framework can provide a natural mechanism to cope with the out-of-sample data for many clustering methods including SC and K-means clustering.

As shown in Section IV, many clustering methods such as SC, CLGR, K-means clustering, and DisKmeans clustering can be deemed as a special variant of our SEC methods. Therefore, we can solve the unified objective function (15) to obtain a cluster assignment matrix F for the aforementioned clustering methods subject to different definition of Laplacian matrices and the parameters. Hereafter, we use the formulas in (14) to compute W and b, which is inherent in our SEC framework. Then, for any new data point $x \in R^d$, we can calculate the prediction

$$y = W^T x + b.$$

Based on this y, the spectral rotation method mentioned in Section II-B can be used to obtain the discrete cluster assignment for x. We first find a unit vector for y. Then, an orthogonal matrix R is obtained by the spectral rotation in (7). Finally, the data point x is assigned to the cluster

$$j = \arg\max_i \tilde{y}(i)$$

where $\tilde{y} = R^T y/\sqrt{\|y\|^2}$, and $\tilde{y}(i)$ is the ith element in the vector $\tilde{y}$.

Specifically, under the framework of SEC, with $\tilde{L}$ and $L + \eta L_o$, respectively, we can readily obtain an out-of-sample extension to SC and CLGR, and also their variants by setting $\mu \to 0$ in SEC. They are referred to as SEC/SC ($\mu \to 0$) and SEC/CLGR ($\mu \to 0$), respectively. Moreover, under the framework of SEC, we can also easily obtain an out-of-sample extension for K-means clustering and DisKmeans clustering by setting $\mu \to \infty$ in SEC. The final cluster assignment of these out-of-sample data can be obtained by using W and b in (14).

In addition, we observe that, if the spectral relaxation is used to solve the cluster assignment matrices, the optimization problem in K-means in (27) and DisKmeans in (34) will lead to the same results because of the fact that $X^T(XX^T + \gamma_g I_d)^{-1}X = I_n - \gamma_g(X^T X + \gamma_g I_n)^{-1}$, and thus $X^T(XX^T + \gamma_g I_d)^{-1}X$ in the optimization problem (34) and $X^T X$ in the optimization problem (27) have the same top c eigenvectors. Hereafter, we refer to K-means/DisKmeans with spectral relaxation as KM-r for in-sample clustering and SEC/KM-r under the framework of SEC for out-of-sample data. The results from K-means and DisKmeans are reported to be different because an EM-like method is used to solve the cluster assignment matrices of the optimization problem in (27) and (34) for K-means clustering and DisKmeans clustering, respectively.

TABLE I
Dataset Description

Dataset	Size	Dimensions	Classes
AR	840	768	120
YALE-B	2414	1024	38
CMU PIE	3329	1024	68
MPEG7	1400	6000	70
COIL20	1440	16384	20
Optdigits	5620	64	10
USPS	9298	256	10
MNIST	70000	784	10

VI. Experiments

In this section, we conduct extensive experiments to evaluate the clustering performance of different clustering methods in two settings, namely, in-sample clustering and out-of-sample clustering.

The first setting is to assign a cluster label to each unlabeled in-sample data point. We compare the proposed SEC method with K-means (KM) clustering, DisKmeans (DKM) clustering [22], SC [10], local learning for (LL) clustering [45], and CLGR [37]. We employ the spectral relaxation + spectral rotation method to compute the assignment matrix for SEC, SC, LL, and CLGR. For KM and DKM, we use an EM-like method to assign cluster labels as in [22]. We further use spectral relaxation + spectral rotation for KM and DKM. Since KM and DKM achieve the same results using the spectral relaxation + spectral rotation as discussed in Section V, we denote the results as KM-r in this paper.

In out-of-sample setting, we assign the cluster label of each unseen data point to the closest cluster center learned from the training set (or in-sample data points) for KM and DKM. Note that the variants of SC methods such as SC, LL, and CLGR do not have a natural out-of-sample extension. However, these methods can be unified under our proposed SEC framework, and thus the newly proposed out-of-sample approach discussed in Section V can be readily used to cope with unseen data for SC, LL, and CLGR. Therefore, we denote them as SEC/SC ($\mu \to 0$), SEC/LL ($\mu \to 0$), and SEC/CLGR ($\mu \to 0$) on unseen data for better presentation. Moreover, we also use the proposed out-of-sample approach for KM-r (referred to as SEC/KM-r). Since the Nyström method [11] can be used to estimate the distribution of unseen (i.e., out-of-sample) data points, we additionally report the results from the Nyström method for the out-of-sample SC.

A. Experimental Setup

Eight high-dimensional datasets are used in the experiments, including three face datasets (AR [46], YALE-B [47], and CMU PIE [48]), two shape image datasets (MPEG7 [49] and COIL20 [50]) and three handwritten digit datasets (Optdigits [51], USPS [52], and MNIST [53]). Some datasets are resized, and Table I summarizes the details of the datasets used in the experiments.

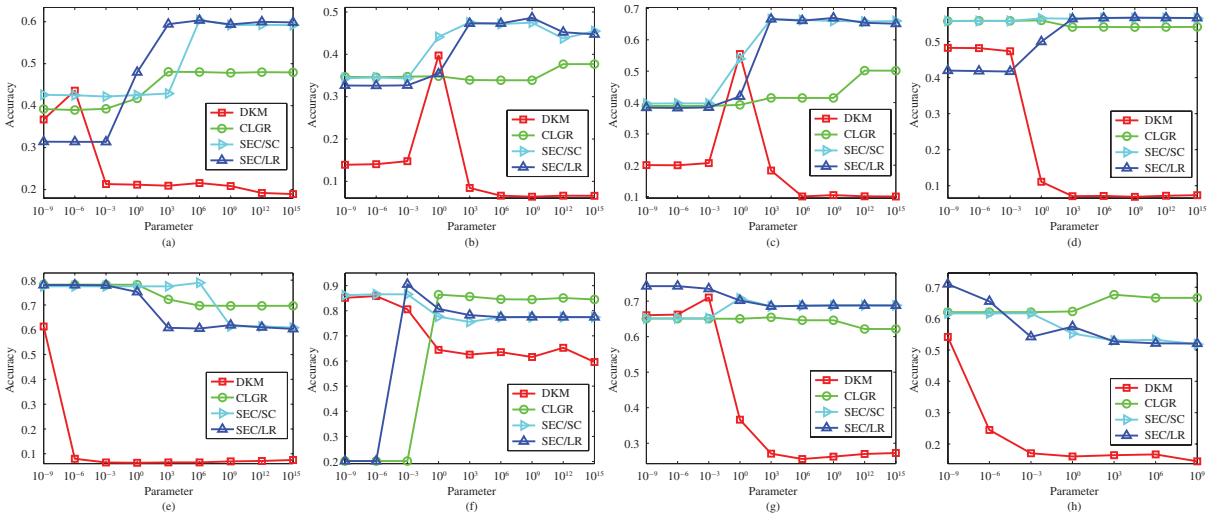

Fig. 1. Clustering accuracy of DisKmeans with different γ_g, SEC/SC, SEC/LR, and CLGR with different μ for the in-sample clustering on the eight datasets. (a) AR. (b) YALE-B. (c) CMU PIE. (d) MPEG7. (e) COIL20. (f) Optdigits. (g) USPS. (h) MNIST.

We randomly partition the data into seen and unseen datasets, where the seen data is used to obtain the clusters and perform cross validation for determining the optimal parameters of different clustering algorithms as suggested in [54], while the unseen data are used to test the performance of clustering algorithms with the optimal parameters in the in-sample setting. In the experiments, 60% of the data are randomly selected as seen data for the AR, YALE-B, CMU PIE, MPEG7, COIL20, and Optdigits datasets, while 20% and 5% of the data are randomly selected as seen data for the larger datasets, USPS and MNIST, respectively. All the remaining data are used as unseen data.

For SC and CLGR, the parameter σ in (1) needs to be determined. In this paper, we use the self-tuning SC method [25] to determine the parameter σ, and the k in k-nearest-neighbor graph for constructing Laplacian matrices is set to 5 for all the algorithms in the experiments. We also need to set the regularization parameters for SEC, CLGR, and DKM beforehand. For fair comparisons and to study the effect of the linearity regularization term on various datasets for SEC methods, we set the parameters γ_l, γ_g in SEC, γ in CLGR, and λ in LL as 1, and set the parameter μ in SEC, the parameter η in CLGR, and the parameter γ_g in DKM as $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}, 10^{12}, 10^{15}\}$. Since there are numerical problems for DKM with a large value of γ_g on the MNIST dataset, we limit the parameter in a range of $\{10^{-9}, 10^{-6}, 0^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$ for all the methods for fair comparisons on this dataset.

For the Nyström method, to obtain better performance, we define $\sigma = 2^r\sigma_0$, where $\sigma_0 = 1/A$, with A being the mean value of the square distance between the in-sample data as suggested in [55] and [56], and $r = \{-3, -2, -1, 0, 1, 2, 3\}$ except for the Optdigits and USPS datasets. To avoid the singular value problem on the Optdigits (resp. USPS) dataset, we set $r = -4$ (resp. $r = 0$) only. We report the best clustering results from the best parameters for SEC, CLGR, LL, and DKM in Table II for the in-sample clustering and their corresponding optimal parameters in Table IV. We also report the results for the out-of-sample clustering in Table III using the optimal parameters in Table IV.

Moreover, the results of all clustering algorithms depend on the initialization (either EM-like or spectral rotation). To reduce statistical variation for each parameter and each random partition, we independently repeat all clustering algorithms 50 times with random initializations, and then we report the results corresponding to the best objective values. Finally, we report the mean clustering accuracy and standard deviation corresponding to the best parameters over 20 random partitions on the seen and unseen data.

B. Evaluation Metrics

We use the clustering accuracy to evaluate the performance for all the clustering algorithms. The clustering accuracy (ACC) is defined as

$$ACC = \frac{\sum_{i=1}^{n} \delta(l_i, map(c_i))}{n} \tag{35}$$

where l_i is the true class label and c_i is the cluster label of x_i obtained from the clustering algorithm, $\delta(x, y)$ is the delta function, and $map(\cdot)$ is the best mapping function. Note $\delta(x, y) = 1$, if $x = y$, $\delta(x, y) = 0$, otherwise. A larger ACC indicates a better performance. The mapping function $map(\cdot)$ matches the true class labels and the obtained cluster labels, and the best mapping is solved by using the Kuhn–Munkres algorithm [57].

C. Experimental Results

In Fig. 1, we first study the sensitivity of the in-sample clustering performances of CLGR, SEC/SC, and SEC/LR (resp. DKM) with respect to the parameter μ (resp. γ_g). Generally, the clustering accuracy for CLGR is less sensitive to the parameter when compared with the other three methods. SEC/SC and SEC/LR generally favor a large value for μ

TABLE II

PERFORMANCE COMPARISON OF CLUSTERING ACCURACY USING KM, DKM, KM-r, SC, LL, CLGR, SEC/SC, AND SEC/LR FOR THE IN-SAMPLE CLUSTERING ON EIGHT DATASETS. THE FIRST ROW FOR EACH DATASET DENOTES THE ACCURACY ± STANDARD DEVIATION, AND THE CORRESPONDING PARAMETER IS SHOWN IN TABLE IV. THE SECOND (RESP. THIRD) ROW DENOTES THE STUDENT t-TEST RESULT OF SEC/SC (RESP. SEC/LR) AGAINST THE REST OF THE METHODS, WHERE 1 DENOTES SEC/SC (RESP. SEC/LR) IS BETTER THAN THE CORRESPONDING METHOD, 0 DENOTES A COMPARABLE RESULT, AND −1 DENOTES A WORSE RESULT

Dataset	KM	DKM	KM-r	SC	LL	CLGR	SEC/SC	SEC/LR
AR	36.3±1.8	43.6±2.2	59.7±2.0	42.4±1.8	47.9±1.4	48.1±1.6	60.4±2.1	60.3±1.9
	1	1	0	1	1	1	N.A.	0
	1	1	0	1	1	1	0	N.A.
YALE-B	14.1±0.7	39.7±3.2	47.3±1.2	34.5±1.0	32.6±1.4	37.7±1.9	47.5±1.4	48.6±1.9
	1	1	0	1	1	1	N.A.	0
	1	1	1	1	1	1	0	N.A.
CMU-PIE	20.1± 0.8	55.5±2.8	68.0±1.4	39.6±1.8	41.1±1.7	50.2±2.0	66.6±1.6	67.0±1.4
	1	1	0	1	1	1	N.A.	0
	1	1	0	1	1	1	0	N.A.
MPEG7	47.9±1.8	48.2±1.6	56.8±1.1	55.7±1.1	54.8±1.1	55.9±1.0	56.5±1.0	56.6±1.5
	1	1	0	1	1	1	N.A.	0
	1	1	0	1	1	1	0	N.A.
COIL20	68.0±3.0	61.3±4.4	62.3±3.3	77.5±3.9	68.9±1.1	78.2±2.9	79.0±2.2	78.0±3.1
	1	1	1	1	1	0	N.A.	0
	1	1	1	0	1	0	0	N.A.
Optdititis	85.3±0.8	85.8±0.9	78.5±0.9	86.4±2.8	88.7±1.6	86.4± 2.6	86.6±2.9	90.5±1.8
	1	0	1	0	−1	0	N.A.	−1
	1	1	1	1	1	1	1	N.A.
USPS	66.3±1.4	71.0±1.0	67.4±1.4	65.2±1.9	63.9±4.2	65.4±0.8	70.8±1.3	74.2±4.5
	1	0	1	1	1	1	N.A.	−1
	1	1	1	1	1	1	1	N.A.
MNIST	53.2±2.3	54.2±1.9	51.6±1.3	61.7±2.5	62.8±4.2	67.6±2.8	61.7±2.5	71.0±2.5
	1	1	1	0	0	−1	N.A.	−1
	1	1	1	1	1	1	1	N.A.

on the three face datasets and MPEG7 dataset, and their performances on these datasets are relatively stable when a large value is used for μ, while these methods prefer a small value of μ for the USPS and MNIST datasets. On the other hand, DKM is very sensitive to the parameter γ_g on all the datasets. Note that setting a large value for μ indicates that the linearity regularization, which captures the global discriminative information for the cluster assignment matrix, is more important for SEC/SC and SEC/LR on the datasets like AR, YALE-B, CMU PIE, and MPEG7, which probably have a strong linearity relationship between the data matrix X and the cluster assignment matrix Y.

For the comprehensive study of performances of various clustering methods, we also report the in-sample clustering results on all the datasets in Table II. Moreover, the Student t-test is also reported in order to evaluate the performance significance for different clustering algorithms, in which the significance level is set as 0.05. From the experimental results, we have the following observations for the in-sample clustering results.

1) Among the K-means-based algorithms (i.e., EM-like techniques such as KM and DKM as well as the spectral relaxation + spectral rotation method such as KM-r), there is no consistent winner. KM achieves the best result on the COIL20 dataset, DKM achieves the best result on the three digit datasets, and KM-r achieves the best result on the three face datasets and the MPEG7 dataset.
2) Variants of SC (e.g., SC, LL, and CLGR) outperform KM-r for the COIL20, Optdigits, and MNIST datasets, while they perform worse on the face datasets and achieve comparable result on the MPEG7 and USPS datasets. One possible explanation is that SC is designed to cluster the data that have a clear manifold structure in a low-dimensional space. If the data do not fulfill this assumption, it performs worse than the K-means-like algorithms.
3) LL outperforms SC on the AR, CMU PIE, Optdigits, and MNIST datasets, while SC outperforms LL on the rest of the datasets. CLGR outperforms SC and LL for all datasets except Optdigits.
4) Our methods SEC/SC and SEC/LR outperform KM, DKM, SC, LL, and CLGR in most cases, or at least achieve comparable results. For the image datasets with strong lighting variations, such as AR, Yale-B, and CMU PIE, we observe significant improvement of SEC/SC and SEC/LR over SC, LL, and CLGR. It clearly demonstrates the effectiveness of the proposed SEC methods on the datasets that do not have a clear manifold structure in a low dimensional space.
5) SEC methods perform better than KM-r on the COIL20 and the three digit datasets, while they achieve comparable results on the three face and MPEG7 datasets. This is probably because the face datasets and the MPEG7 dataset prefer a large value of μ for SEC methods, which can be observed in Fig. 1. On the other hand, as discussed in Section IV-C, when $\mu \to \infty$ and $\gamma_g \to \infty$ in SEC, it reduces to K-means clustering. When the datasets prefer a small value of μ, SEC methods outperform the KM-r method.

TABLE III

PERFORMANCE COMPARISON OF CLUSTERING ACCURACY USING NYSTRÖM SC, KM, DKM, SEC/KM-r ($\mu \to \infty$), SEC/SC ($\mu \to 0$), SEC/LL ($\mu \to 0$), SEC/CLGR ($\mu \to 0$), SEC/SC, AND SEC/LR FOR THE OUT-OF-SAMPLE CLUSTERING ON EIGHT DATASETS. THE FIRST ROW FOR EACH DATASET DENOTES THE ACCURACY ± STANDARD DEVIATION, AND THE CORRESPONDING PARAMETER IS SHOWN IN TABLE IV. THE SECOND (RESP. THIRD) ROW DENOTES THE STUDENT t-TEST RESULTS OF THE SEC/SC (RESP. SEC/LR) AGAINST EACH OF THE REST METHODS, WHERE 1 DENOTES THE SEC/SC (RESP. SEC/LR) IS BETTER THAN THE CORRESPONDING METHOD, 0 DENOTES A COMPARABLE RESULT, AND −1 DENOTES A WORSE RESULT

Dataset	Nyström SC	KM	DKM	SEC/KM-r ($\mu \to \infty$)	SEC/SC ($\mu \to 0$)	SEC/LL ($\mu \to 0$)	SEC/CLGR ($\mu \to 0$)	SEC/SC	SEC/LR
AR	35.2±1.4	33.6±1.7	26.4±1.3	66.6±1.9	43.0±1.7	47.3±1.7	47.4±2.0	67.1±2.5	67.1±2.0
	1	1	1	0	1	1	1	N.A.	0
	1	1	1	0	1	1	1	0	N.A.
YALE-B	16.7±1.8	13.0±0.5	20.8±1.0	43.5±1.0	35.9±1.8	35.0±1.1	35.6±1.4	42.5±1.7	42.8±1.3
	1	1	1	0	1	1	1	N.A.	0
	1	1	1	0	1	1	1	0	N.A.
CMU-PIE	14.8±1.3	19.2±0.8	29.8±1.8	65.4±1.5	45.3±2.4	48.6±2.1	49.4±1.8	64.0±1.6	64.4±1.3
	1	1	1	0	1	1	1	N.A.	0
	1	1	1	0	1	1	1	0	N.A.
MPEG7	43.5±2.9	47.6±1.4	47.6±1.5	55.7±1.1	55.2±1.1	54.7±0.8	55.2±1.4	55.7±1.0	55.7±1.1
	1	1	1	0	0	1	1	N.A.	0
	1	1	1	0	1	1	1	0	N.A.
COIL20	67.9±4.8	61.6±4.2	61.8±3.0	76.6±3.4	68.9±1.0	77.4±2.4	77.4±2.4	78.4±2.4	77.2±2.3
	1	1	1	1	1	1	1	N.A.	1
	1	1	1	1	0	1	0	−1	N.A.
Optditits	60.6±11.9	85.4±0.5	85.8±0.7	78.9±0.6	85.8±2.8	88.2±1.7	85.9±2.6	86.0± 2.7	90.0±1.8
	1	0	0	1	1	−1	0	N.A.	−1
	1	1	1	1	1	1	1	1	N.A.
USPS	28.3±5.2	66.0±1.2	68.4±0.6	67.1±1.3	61.5±2.0	60.2±3.3	61.8±0.6	68.1±0.7	70.3±3.3
	1	1	0	1	1	1	1	N.A.	−1
	1	1	1	1	1	1	1	1	N.A.
MNIST	N.A.	52.8±1.8	53.6±1.7	51.8±0.8	55.1±1.8	54.7±3.5	59.7±2.4	55.1±1.8	62.9±1.9
	N.A.	1	1	1	0	0	−1	N.A.	−1
	N.A.	1	1	1	1	1	1	1	N.A.

TABLE IV

PARAMETERS USED FOR THE OPTIMAL RESULT FOR DKM, LL, CLGR, SEC/SC, AND SEC/LR ON EIGHT DATABASES

Dataset	DKM	CLGR	SEC/SC	SC/LR
AR	10^{-6}	10^{15}	10^{6}	10^{6}
YALE-B	10^{0}	10^{15}	10^{9}	10^{9}
CMU-PIE	10^{0}	10^{12}	10^{6}	10^{9}
MPEG7	10^{-9}	10^{0}	10^{12}	10^{9}
COIL20	10^{-9}	10^{-9}	10^{6}	10^{-9}
Optdigits	10^{-6}	10^{0}	10^{-6}	10^{-3}
USPS	10^{-3}	10^{3}	10^{0}	10^{-9}
MNIST	10^{-9}	10^{3}	10^{-3}	10^{-9}

6) SEC/LR performs better than the SEC/SC on the three digit datasets, while they achieve comparable results on the rest of the datasets. One possible explanation is that discriminative information of the cluster for digit data can be better captured by both global and local regression as in (22) compared to using only the global regression as in (18).

We also study the performances of various clustering methods for the unseen data (i.e., out-of-sample), and the corresponding accuracies of various clustering methods on all the datasets are also reported in Table III, in which the significance level for the Student t-test is again set as 0.05. The optimal parameters are determined by using the cross-validation method from the in-sample clustering except for the Nyström method, in which we select the best parameter based on the out-of-sample results. The result for the Nyström method on the MNIST dataset is not available because it requires more than 32 GB memory to store the full similarity matrix, which is larger than the available memory on our servers. From the experimental results, we have the following observations.

1) The clustering accuracy of the Nyström method is significantly worse than the clustering methods under the proposed SEC framework. One possible explanation is that the Nyström method uses Nyström extension to approximate the similarity matrix C between the unseen data using the two similarity matrices A and B, where A is the similarity matrix between the seen data and B is the similarity matrix between the seen data and the unseen data. Depending on the nature of A and B, C may not be well approximated, and hence the whole similarity matrix constructed using matrices A, B, and C may not fully represent the relationship among the data. On the other hand, our proposed SEC framework relies on the linear property of the cluster assignment matrix, which is usually satisfied for real-world data [34].
2) The EM-like clustering methods (such as DKM) are significantly degraded on the three face datasets when compared to the corresponding results in the in-sample clustering setting (Table II), which is possibly due to the large variation of face data between the seen data

and the unseen data. However, the clustering accuracies for different clustering algorithms under the SEC framework are comparable to those of the seen data, which demonstrates the effectiveness of our SEC framework in terms of generalization performance. In particular, the clustering performance of SEC/KM-r is significantly better than that of KM and DKM on the three face datasets as well as MPEG7 dataset, which indicates that the proposed SEC framework can effectively cope with the out-of-sample data in clustering tasks.

3) Similar to the in-sample clustering, SEC/SC and SEC/LR outperform or achieve comparable results in most cases when compared to the SEC/KM-r, SEC/SC, SEC/LL, and SEC/CLGR methods.

VII. Conclusion

In this paper, we proposed the SEC framework based on the observation that for the high-dimensional data the true cluster assignment matrix can always be embedded in a linear space spanned by the data. The linear property of the cluster assignment matrix was incorporated in the clustering framework by imposing a linearity regularization. The proposed clustering framework also provided a natural mechanism to deal with the out-of-sample data in clustering. Moreover, we introduced a new algorithm for SEC using both local and global discriminative information. The experiments on eight real-world high-dimensional datasets showed the effectiveness of our SEC framework for both in-sample and out-of-sample clusterings.

Appendix

Proof of Theorem 1

Suppose the eigenvalue decomposition of S_t is $S_t = [U_1, U_0]\begin{bmatrix}\Lambda_t^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}[U_1, U_0]^T$. Let $B = \Lambda_t^{-1}U_1^T S_b U_1 \Lambda_t^{-1}$. Suppose the eigenvalue decomposition of B is $B = V_b \Lambda_b V_b^T$, and let $P = [U_1\Lambda_t^{-1}V_b, U_0]$. We have the following two lemmas.

Lemma 1 [34, Th. 5.1]: If $rank(S_t) = rank(S_w) + rank(S_b)$, then $P^T S_t P = \begin{bmatrix}I_{r_t} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} = D_t$ and $P^T S_b P = \begin{bmatrix}I_{r_b} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} = D_b$, where $I_t \in \mathbb{R}^{r_t \times r_t}$ and $I_{r_b} \in \mathbb{R}^{r_b \times r_b}$ are identity matrices, r_t is the rank of S_t, r_b is the rank of S_b, and $r_b \le r_t$.

Lemma 2 [44]: If $rank(S_t) = rank(S_w) + rank(S_b)$, then $S_t^+ = P D_t P^T$.

Then we proceed to the proof of Theorem 1.

Proof: According to Lemmas 1 and 2, we have

$$\begin{aligned} &S_w S_t^+ S_b\\ &= S_t S_t^+ S_b - S_b S_t^+ S_b\\ &= P^{-T}P^T S_t S_t^+ S_b P P^{-1} - P^{-T}P^T S_b S_t^+ S_b P P^{-1}\\ &= P^{-T} D_t D_t D_b P^{-1} - P^{-T} D_b D_t D_b P^{-1}\\ &= \mathbf{0}. \end{aligned}$$

Note that $S_b = XGG^T X^T$. Therefore, we have $S_w S_t^+ XGG^T X^T = \mathbf{0}$. Multiplying $(S_w S_t^+)^T$ on both sides, we arrive at

$$S_w S_t^+ XGG^T X^T \left(S_w S_t^+\right)^T = \mathbf{0}.$$

Alternatively, $S_w S_t^+ XG(S_w S_t^+ XG)^T = \mathbf{0}$, which indicates $S_w S_t^+ XG = \mathbf{0}$ and $S_w S_t^+ XY = \mathbf{0}$. Let us define W_0 to be

$$W_0 = S_t^+ XY.$$

Then we have

$$S_w W_0 = \mathbf{0}.$$

Therefore, W_0 is in the null space of S_w. From [58, Th. 1], for multiclass problems, all the data that belong to the same class will be projected onto the same point under the projection W_0, thus we have

$$\forall i, y_i = [\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{c-j}]^T \Rightarrow x_i^T W_0 = \bar{x}_j^T W_0 \quad (36)$$

where y_i^T is the ith row of the true cluster assignment matrix Y and $\bar{x}_j$ is the mean of the data that belong to class j.

Denote $\bar{X}_c = [\bar{x}_1, \ldots, \bar{x}_c]$. Note that $\bar{X}_c = XY\Sigma$, where $\Sigma \in \mathbb{R}^{c\times c}$ is a diagonal matrix with the ith diagonal element as $1/n_i$, and n_i is the number of the data that belong to Class i. Then

$$\begin{aligned} rank(\bar{X}_c^T W_0) &= rank\left(\Sigma Y^T X^T (XX^T)^+ XY\right)\\ &= rank((XX^T)^+ XY)\\ &= rank(S_b) = c - 1. \end{aligned}$$

Denote $Q = \bar{X}_c^T W_0 + \mathbf{1}_c\mathbf{1}_c^T$. Note that $Y\mathbf{1}_c = \mathbf{1}_n$ and $X\mathbf{1}_n = \mathbf{0}$, so

$$\bar{X}_c^T W_0 \mathbf{1}_c = \mathbf{0} \text{ and } \mathbf{1}_c^T \Sigma^{-1} \bar{X}_c^T W_0 = \mathbf{0}.$$

Thus, vector $\mathbf{1}_c$ is linearly independent of the rows or the columns of $\bar{X}_c^T W_0$. Suppose the full rank decomposition of $\bar{X}_c^T W_0$ is $\bar{X}_c^T W_0 = Q_1 Q_2^T$, where $Q_1, Q_2 \in \mathbb{R}^{c\times(c-1)}$ are column full-rank matrices. Then $Q = Q_1 Q_2^T + \mathbf{1}_c\mathbf{1}_c^T = [Q_1, \mathbf{1}_c][Q_2, \mathbf{1}_c]^T$. As $[Q_1, \mathbf{1}_c]$ and $[Q_2, \mathbf{1}_c]$ both are full-rank matrices, then Q is invertible. Hence we have

$$\begin{aligned} I_c &= QQ^{-1}\\ &= (\bar{X}_c^T W_0 + \mathbf{1}_c\mathbf{1}_c^T)Q^{-1}\\ &= \bar{X}_c^T W_0 Q^{-1} + \mathbf{1}_c(Q^{-T}\mathbf{1}_c)^T. \end{aligned}$$

Let $W = W_0 Q^{-1}$ and $b = Q^{-T}\mathbf{1}_c$. Then $(\bar{X}_c^T W + \mathbf{1}_c b^T) = I_c$. According to (36), we have $X^T W + \mathbf{1}_n b^T = Y$. Therefore, If $rank(S_b) = c-1$ and $rank(S_t) = rank(S_w) + rank(S_b)$, there exist $W \in \mathbb{R}^{d\times c}$ and $b \in \mathbb{R}^{c\times 1}$ such that $Y = X^T W + \mathbf{1}_n b^T$. ■

References

[1] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[3] A. P. Benavent, F. E. Ruiz, and J. Saez, "Learning Gaussian mixture models with entropy-based criteria," *IEEE Trans. Neural Netw.*, vol. 20, no. 11, pp. 1756–1771, Nov. 2009.

[4] K. Zhang and J. T. Kwok, "Simplifying mixture models through function approximation," *IEEE Trans. Neural Netw.*, vol. 21, no. 4, pp. 644–658, Apr. 2010.

[5] K. Tasdemir and E. Merényi, "Exploiting data topology in visualization and clustering of self-organizing maps," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 549–562, Apr. 2009.

[6] K. Tasdemir, "Graph based representations of density distribution and distances for self-organizing maps," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 520–526, Mar. 2010.

[7] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 669–688, Apr. 2002.

[8] A. Szymkowiak-Have, M. Girolami, and J. Larsen, "Clustering via kernel decomposition," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 256–264, Jan. 2006.

[9] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[10] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. Int. Conf. Comput. Vis.*, Beijing, China, 2003, pp. 313–319.

[11] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[12] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 176–190, Jan. 2008.

[13] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[14] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *J . Mach. Learn. Res.*, vol. 2, pp. 125–137, Nov. 2001.

[15] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2005, pp. 1537–1544.

[16] K. Zhang, I. W. Tsang, and J. T. Kwok, "Maximum margin clustering made practical," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 583–596, Apr. 2009.

[17] Y. Li, I. W. Tsang, J. T. Kwok, and Z. Zhou, "Tighter and convex maximum margin clustering," in *Proc. 12th Int. Conf. Artif. Intell. Stat.*, Clearwater Beach, FL, 2009, pp. 344–351.

[18] W. Zhong, W. Pan, J. T. Kwok, and I. W. Tsang, "Incorporating the loss function into discriminative clustering of structured outputs," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1564–1575, Oct. 2010.

[19] F. D. La Torre and T. Kanade, "Discriminative cluster analysis," in *Proc. Int. Conf. Mach. Learn.*, Pittsburgh, PA, 2006, pp. 241–248.

[20] C. H. Q. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and K-means clustering," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, 2007, pp. 521–528.

[21] J. Ye, Z. Zhao, and H. Liu, "Adaptive distance metric learning for clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, Jun. 2007, pp. 1–7.

[22] J. Ye, Z. Zhao, and M. Wu, "Discriminative K-means for clustering," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2007, pp. 1649–1656.

[23] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[24] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J . Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[25] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004, pp. 1601–1608.

[26] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for LLE, ISOMAP, MDS, eigenmaps, and spectral clustering," in *Proc. Neural Inf. Process. Syst.*, 2003, pp. 126–133.

[27] C. Alzate and A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 335–347, Feb. 2010.

[28] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Mach. Learn. Res.*, vol. 2, pp. 263–286, Jan. 1995.

[29] A. Passerini, M. Pontil, and P. Frasconi, "New results on error correcting output codes of kernel machines," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 45–54, Jan. 2004.

[30] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 201–233, Feb. 2002.

[31] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Spectral embedded clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, Pasadena, CA, 2009, pp. 1181–1186.

[32] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, pp. 849–856.

[33] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, Apr. 2009.

[34] J. Ye, "Least squares linear discriminant analysis," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, 2007, pp. 1087–1093.

[35] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.

[36] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[37] F. Wang, C. Zhang, and T. Li, "Clustering with local and global regularization," in *Proc. Assoc. Adv. Artif. Intell.*, Vancouver, BC, Canada, 2007, pp. 657–662.

[38] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel K-means: Spectral clustering and normalized cuts," in *Proc. KDD*, Seattle, WA, 2004, pp. 551–556.

[39] I. S. Dhillon, Y. Guan, and B. Kulis, "A unified view of kernel K-means, spectral clustering and graph partitioning," Dept. Comput. Sci., Univ. Texas at Austin, Austin, Tech. Rep. TR-04-25, Feb. 2005.

[40] M. Wu and B. Schölkopf, "Transductive classification via local learning regularization," in *Proc. Int. Conf. Artif. Intell. Stat.*, Mar. 2007, pp. 628–635.

[41] H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for K-means clustering," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, pp. 1057–1064.

[42] C. H. Q. Ding, X. He, H. Zha, and H. D. Simon, "Adaptive dimension reduction for clustering high dimensional data," in *Proc. Int. Conf. Data Min.*, Maebashi, Japan, 2002, pp. 147–154.

[43] T. Li, S. Ma, and M. Ogihara, "Document clustering via adaptive subspace iteration," in *Proc. 27th Annu. Int. Conf. SIGIR*, Sheffield, U.K., 2004, pp. 218–225.

[44] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, Apr. 2005.

[45] M. Wu and B. Schölkopf, "A local learning approach for clustering," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 1529–1536.

[46] A. Martinez and R. Benavente, "The AR face database," Centre Vis. Comput., Univ. Autonoma Barcelona, Barcelona, Spain, Tech. Rep. 24, Jun. 1998.

[47] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[48] T. Sim and S. Baker, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1617, Dec. 2003.

[49] L. Latecki, R. Lakämper, and U. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Hilton Head Island, SC, 2000, pp. 424–429.

[50] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Dept. Comput. Sci., Columbia Univ., New York, Tech. Rep. CUCS-005-96, 1996.

[51] A. Asuncion and D. Newman. (2007). *UCI Machine Learning Repository*. School Inf. Comput. Sci., Univ. California, Irvine [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[52] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2003.

[53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[54] J. Hjorth, *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap*. London, U.K.: Chapman & Hall, 2010.

[55] L. Duan, D. Xu, T. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, Jun. 2010, pp. 1959–1966.

[56] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Comput. Vis. Pattern Recognit.*, Anchorage, AK, Jun. 2008, pp. 1–8.

[57] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, Mar. 1957.

[58] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 4–13, Jan. 2005.

Feiping Nie received the B.S. degree from the North China University of Water Conservancy and Electric Power, Zhengzhou, China, in 2000, the M.S. degree from Lanzhou University, Lanzhou, China, in 2003, and the Ph.D. degree from Tsinghua University, Beijing, China in 2009, all in computer science.

He is currently a Research Assistant Professor with the University of Texas, Arlington. His current research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Zinan Zeng received the B.E. degree (First Hon.) from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2008. He is currently pursuing the Masters degree in the same university.

He is currently a Research Assistant with Nanyang Technological University. He was with R&D Engineering at Global Digital Creations Technology, Hong Kong, from 2008 to 2009. His current research interests include statistical learning, sparse coding, clustering, and their applications in computer vision.

Ivor W. Tsang received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2007.

He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. He is also the Deputy Director of the Center for Computational Intelligence in NTU.

Dr. Tsang received the IEEE Transactions on Neural Networks Outstanding 2004 Paper Award in 2006, and the second prize of the National Natural Science Award 2008, China, in 2009. He was awarded the Microsoft Fellowship in 2005 and the Best Paper Award from the IEEE Hong Kong Chapter of Signal Processing Postgraduate Forum in 2006. He also won the Best Student Paper Prize at CVPR'10.

Dong Xu (M'07) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively.

He was with Microsoft Research Asia, Beijing, China, and the Chinese University of Hong Kong, Shatin, Hong Kong, for more than two years, while pursuing the Ph.D. degree. He was a Post-Doctoral Research Scientist in Columbia University, New York, NY, for one year. He is currently an Assistant Professor with the Nanyang Technological University, Singapore. His current research interests include computer vision, statistical learning, and multimedia content analysis.

Dr. Xu was the co-author of a paper that won the Best Student Paper Award in the prestigious IEEE International Conference on Computer Vision and Pattern Recognition in 2010.

Changshui Zhang (M'02) received the B.S. degree in mathematics from Peking University, Beijing, China, in 1986, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, in 1992.

He is currently a Professor with the Department of Automation, Tsinghua University. His current research interests include artificial intelligence, image processing, pattern recognition, machine learning, evolutionary computation and complex system analysis, etc.

Prof. Zhang is an Associate Editor of the *Journal of Pattern Recognition*.