One-Step Spectral Clustering via Dynamically Learning Affinity Matrix and Subspace

Xiaofeng Zhu,[1] **Wei He,**[1] **Yonggang Li,**[1*] **Yang Yang,**[2]
Shichao Zhang,[1†] **Rongyao Hu,**[1] **Yonghua Zhu**[3]
[1]Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, 541004, China
[2]University of Electronic Science and Technology of China, Chengdu, 611731, China
[3]Guangxi University, Nanning, 530004, China
seanzhuxf@gmail.com, zhangsc@gxnu.edu.cn

Abstract

This paper proposes a one-step spectral clustering method by learning an intrinsic affinity matrix (*i.e.,* the clustering result) from the low-dimensional space (*i.e.,* intrinsic subspace) of original data. Specifically, the intrinsic affinity matrix is learnt by: 1) the alignment of the initial affinity matrix learnt from original data; 2) the adjustment of the transformation matrix, which transfers the original feature space into its intrinsic subspace by simultaneously conducting feature selection and subspace learning; and 3) the clustering result constraint, *i.e.,* the graph constructed by the intrinsic affinity matrix has exact c connected components where c is the number of clusters. In this way, two affinity matrices and a transformation matrix are iteratively updated until achieving their individual optimum, so that these two affinity matrices are consistent and the intrinsic subspace is learnt via the transformation matrix. Experimental results on both synthetic and benchmark datasets verified that our proposed method outputted more effective clustering result than the previous clustering methods.

Introduction

Spectral clustering has drawn growing concern due to finding the cluster membership of the data by considering the inherent structure among data points to naturally reflect the relationships of the data (Wang et al. 2011; Lu et al. 2012). The previous spectral clustering is a two-step strategy, *i.e.,* first learning an affinity matrix to measure the similarity among data points (*i.e.,* the affinity matrix learning step) and then conducting a k-means clustering on the resulting affinity matrix to output final clustering result (*i.e.,* the k-means clustering step). Usually, the affinity matrix (*i.e.,* similarity graph) learning step, *i.e.,* transferring the finding of cluster membership to an optimal graph partition problem, is the most key step of spectral clustering (Nie and Huang 2016; Wang and Siskind 2003). Representation methods have been well-known as the most popular methods for the affinity matrix learning, by assuming that each data point may be represented by other data points (Peng, Zhang, and Yi 2013; Zhu et al. 2013). Specifically, representation methods use the resulting representation coefficient to measure the similarity among data points, *i.e.,* large coefficient indicates close relationship between two data points while small coefficient indicates distant relationship.

Representation methods for the construction of the affinity matrix include global representation methods (Lu et al. 2012; Liu et al. 2013) and local representation methods (Ng et al. 2002; Luo et al. 2011; Nie et al. 2016). Global representation methods represent each data point by all data points, such as the low-rank representation method (Liu et al. 2013) and the least square representation method (Lu et al. 2012; Peng, Zhang, and Yi 2013). Local representation methods represent each data point by its nearest neighbors, with the assumption that high-dimensional data usually lie on a low-dimensional space, *i.e.,* an intrinsic subspace. For example, a global representation method in (Elhamifar and Vidal 2013) and a local representation method in (Nie et al. 2016) used an ℓ_1-norm sparse model and a k Nearest Neighbor (kNN) graph, respectively, to conduct the affinity matrix. In a nutshell, the previous representation methods conduct the affinity matrix learning step by sharing a common strategy, *i.e.,* representing each data point by other data points with different criteria.

However, the previous spectral clustering methods still have drawbacks to be overcome. First, the affinity matrix learning is sensitive to the data quality. The previous spectral clustering methods (Ng et al. 2002; Lu et al. 2012; Liu et al. 2013; Elhamifar and Vidal 2013) learn the affinity matrix from original data, which are often corrupted by noise and outliers, thus unavailable to correctly disclose the similarity among data points. Second, the k-means method is well-known as sensitive to the initialization of clustering centers (Ng et al. 2002). Lastly, even though each step achieves their individual optimum, the two-step strategy easily leads to suboptimal result since individual optimum cannot ensure the global optimum of the two-step strategy.

In this paper, we propose a one-step spectral learning method to learn an intrinsic affinity matrix from the intrinsic subspace. Moreover, the intrinsic affinity matrix is actually the clustering result without conducting the k-means clustering step. Different from the previous spectral clustering methods learning a fixed affinity matrix from original data, our proposed method learns the intrinsic affinity matrix by: 1) the alignment of the initial affinity matrix learnt from the original feature space. The motivation is that these

[*] Wei He and Yonggang Li equally contributed to this work.

[†] Corresponding author: Shichao Zhang.

two affinity matrices have different illustrations in real applications due to the influence of noise and outliers, but they measure the similarity of the same data. Thus this paper proposes to first learn them individually and then align them to be consistent. 2) the adjustment of the transformation matrix, which transfers the original feature space into the intrinsic subspace by simultaneously conducting feature selection and subspace learning. The resulting intrinsic subspace is interpretable and robust, and thus enabling to learn the real similarity among data points, *i.e.,* yielding an optimal intrinsic affinity matrix. 3) the clustering result constraint, *i.e.,* the graph constructed by the intrinsic affinity matrix ideally has *c* connected components where *c* is the number of clusters. The clustering result constraint removes the *k*-means clustering step to result in one-step spectral clustering, and thus enabling to output optimal clustering result. Since these two affinity matrices and the transformation matrix are unknown, we propose to iteratively update one of them by fixing the others. As a result, all of them achieve their individual optimum. That is, the intrinsic subspace spanned by the transformation matrix is approximately found. Moreover, the intrinsic affinity matrix is consistent to the initial affinity matrix as well as is the final clustering result.

Compared to the previous two-step spectral clustering methods, we conclude the contributions of our proposed method as follows. First, unlike the previous methods learning either a fixed affinity matrix (Ng et al. 2002; Elhamifar and Vidal 2013) or a dynamic affinity matrix (Liu et al. 2013; Lu et al. 2012), *from original data*, this paper learns a dynamic intrinsic affinity matrix *from the intrinsic subspace* which removes the influence of noise and outliers. Moreover, the learnt intrinsic affinity matrix is consistent to the initial affinity matrix. Furthermore, we couple the learning of these two affinity matrices with the learning of the intrinsic subspace to achieve their individual optimum. Second, our proposed method conducts a *one-step spectral clustering* by only learning the affinity matrix (*i.e.,* the clustering result) without the *k*-means clustering step, which is sensitive to its initialization and used in *two-step spectral clustering*. Moreover, our one-step strategy can obviously avoid the suboptimal issue of the previous two-step spectral clustering.

Method

Notation

In this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters; We also denote the *i*-th row and *j*-th column of a matrix $\mathbf{X} = [x_{ij}]$ as $\mathbf{x}^i$ and $\mathbf{x}_j$, and its Frobenius norm and $\ell_{2,1}$-norm as $||\mathbf{X}||_F = \sqrt{\sum_i \sum_j x_{i,j}^2}$, and $||\mathbf{X}||_{2,1} = \sqrt{\sum_j x_{i,j}^2}$; We further denote the transpose, the trace, the rank, and the inverse, of a matrix $\mathbf{X}$, as $\mathbf{X}^T$, $tr(\mathbf{X})$, $rank(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

Initial affinity matrix learning

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ be the *d*-dimensional feature matrix, where *n* is the number of data points, we can use either global representation methods or local representation methods to construct the affinity matrix $\mathbf{G}$ of graph $\mathbb{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ and $\mathbf{E}$, respectively, represent the set of vertices (*i.e.,* data points) and the set of the edges. Since the local representation methods linearly represent each data point by its nearest neighbors to remove the influence of distant data points (especially outliers) (Roweis and Saul 2000; Yu, Zhang, and Gong 2009), the local representation methods have been demonstrated more robust than global representation methods for the construction of the affinity matrix. Different from that the previous global representation methods learn a fixed affinity matrix from the original feature space (Ng et al. 2002), this paper devises a new local representation method to dynamically learn an initial affinity matrix from the original feature space (this subsection) and an intrinsic affinity matrix from the intrinsic subspace in Section 2.3.

The construction of an affinity matrix actually finds a similarity measurement among data points. With the local representation assumption, *i.e.,* each data point is only connected with its nearest neighbors, we expect that close data points have large similarity while distant data points have small or even zero similarity. Thus, we propose to minimize the following objective function:

$$\min_{\mathbf{G}} \sum_{i,j}^{n} g_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2, \text{ s.t.}, \mathbf{G} \in \mathcal{C}, \quad (1)$$

where the initial affinity matrix $\mathbf{G} = [\mathbf{g}_1, ..., \mathbf{g}_n] \in \mathbb{R}^{n \times n}$, $\mathcal{C} = \{\forall i | \mathbf{c}_i^T \mathbf{1} = 1, c_{i,i} = 0, c_{i,j} \geq 0 \text{ if } j \in \mathbb{N}(i), \text{ otherwise } 0.\}$, $\mathbf{1}$ and $\mathbb{N}(i)$, respectively, represent an all-ones vector and the set of nearest neighbors of the *i*-th data point. The constraint $\mathbf{c}_i^T \mathbf{1} = 1$ in $\mathcal{C}$ enables to result in shift invariant similarity. Eq. (1) leads to small or even zero value of $g_{i,j}$ while $\mathbf{x}_i$ and $\mathbf{x}_j$ are far apart, and large value of $g_{i,j}$ while $\mathbf{x}_i$ and $\mathbf{x}_j$ are close.

It is noteworthy that similar objective function can be found in (Nie and Huang 2016), which used a global representation method to learn the affinity matrix by representing each data point by all data points, while Eq. (1) uses a local representation method to learn the representation of each data point by its nearest neighbors, where the number of nearest neighbors can be tuned by cross-validation methods. Besides, (Nie and Huang 2016) conducted a two-step clustering analysis, while the goal of Eq. (1) is to conduct one-step spectral clustering.

Intrinsic affinity matrix learning

The previous methods (*e.g.,* (Belkin and Niyogi 2001; He and Niyogi 2003)) assume that the affinity matrix constructed in the original feature space represents the real similarity among data points, and thus can be transferred to guide the predictions of the original feature matrix $\mathbf{X}$, *i.e.,*

$$\min_{\mathbf{Y}} \sum_{i,j}^{n} g_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2, \text{ s.t.}, \mathbf{G} \in \mathcal{C}, \quad (2)$$

where the *i*-th vector $\mathbf{y}_i$ of the prediction matrix $\mathbf{Y}$ is the prediction of $\mathbf{x}_i$. In Eq. (2), a fixed similarity $g_{i,j}$ between the *i*-th data point $\mathbf{x}_i$ and the *j*-th data point $\mathbf{x}_j$ learnt from the original feature space is used to guide the predictions of $\mathbf{y}_i$ and $\mathbf{y}_j$ (Belkin and Niyogi 2001; He and Niyogi 2003). However, this assumption usually does not hold since the real distribution of the data are often highly complex.

It is apparent that the affinity matrices generated in different feature spaces have different illustrations due to the influence of noise and outliers. Thus, there is no guarantee that the affinity matrix learnt from the original feature space can effectively guide the clustering process in the intrinsic subspace. Unlike the previous methods learning a fixed affinity matrix, this paper proposes to learn an initial affinity matrix from original data and an intrinsic affinity matrix from the intrinsic subspace of original data. The motivation is that different feature spaces result in different affinity matrices, which also meets the goal of spectral clustering methods, *i.e.,* finding an intrinsic subspace of the original feature space since original data actually lie on a low-dimensional space (Vidal 2011; Zhu et al. 2012). By denoting $\mathbf{W} \in \mathbb{R}^{d\times d'}$ (where $d' \leq d$) as the transformation matrix mapping original data $\mathbf{X}$ to its intrinsic subspace spanned by $\mathbf{W}^T\mathbf{X}$, we design to learn an intrinsic affinity matrix $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_n] \in \mathbb{R}^{n\times n}$ in the intrinsic subspace via:

$$\begin{aligned}\min_{\mathbf{S},\mathbf{W}} \sum_{i,j}^{n} s_{i,j}\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 + \gamma\|\mathbf{W}\|_{2,1},\\ \text{s.t.}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_{d'}, \mathbf{S} \in \mathcal{C},\end{aligned} \tag{3}$$

where $\mathbf{I}_{d'} \in \mathbb{R}^{d'\times d'}$ and γ, respectively, are an identity matrix and a tuning parameter. The penalty $\|\mathbf{W}\|_{2,1}$ conducts feature selection by outputting the row sparsity on $\mathbf{W}$ to remove the noisy/redundant features of $\mathbf{X}$, while the orthogonal constraint on the scatter matrix $\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$ actually conducts subspace learning to transfer original d-dimensional feature space into a statistically uncorrelated d'-dimensional space.

The literature (Gu, Li, and Han 2011; Zhu et al. 2016) has demonstrated that subspace learning enables to output robust models and feature selection outputs interpretable models. Therefore, Eq. (3) simultaneously conducts subspace learning (via the orthogonal constraint on the scatter matrix) and feature selection (via the the row sparsity on $\mathbf{W}$), and thus achieving robust and interpretable models for finding the ideally intrinsic subspace (via $\mathbf{W}$), where the intrinsic affinity matrix $\mathbf{S}$ is yielded. Unfortunately, we have no prior knowledge on either the dimensions of the intrinsic subspace or the intrinsic affinity matrix. As a consequence, Eq. (3) is unavailable to output the optimal result for either the intrinsic affinity matrix or the intrinsic subspace. In this paper, we propose two solutions to address this issue, *i.e.,* coupling the intrinsic affinity matrix with the initial affinity matrix (please see Section 2.4) and regarding the intrinsic affinity matrix as the clustering results by ideally expecting that the graph constructed by the intrinsic affinity matrix has exact c connected components where c is the number of clusters (please see Section 2.5).

The consistency of two affinity matrices

With the motivation of that: 1) the initial affinity matrix $\mathbf{G}$ is learnt from original data and thus may be influenced by noise and outliers. As a result, the quality of $\mathbf{G}$ cannot be guaranteed; 2) both $\mathbf{G}$ and $\mathbf{S}$ are used to measure the similarity of the same data points, so their difference, measured by the summation of element-wise similarity, should be as small as possible; and 3) we have no prior knowledge on the dimensions of the intrinsic subspace $\mathbf{W}$, in this paper, we allow $\mathbf{S}$ to be progressively refined by $\mathbf{G}$, aiming at suppressing possible noise and outliers to find the approximate intrinsic dimensions of $\mathbf{X}$, *i.e.,* $\mathbf{W}$. To do this, we couple the estimation of $\mathbf{G}$ with the estimation of $\mathbf{S}$ by designing a dynamic affinity matrix learning model as follows:

$$\begin{aligned}\min_{\mathbf{G},\mathbf{S},\mathbf{W}} \sum_{i,j}^{n} g_{i,j}(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + \alpha\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2)\\ + \beta\sum_{i=1}^{n}\|\mathbf{g}_i - \mathbf{s}_i\|_2^2 + \gamma\|\mathbf{W}\|_{2,1},\\ \text{s.t.}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_{d'}, \mathbf{G} \in \mathcal{C}, \mathbf{S} \in \mathcal{C},\end{aligned} \tag{4}$$

where α and β are tuning parameters, and the constraint $\sum_{i=1}^{n}\|\mathbf{g}_i - \mathbf{s}_i\|_2^2$ is used to preserve the consistency between $\mathbf{G}$ and $\mathbf{S}$.

In Eq. (4), by fixing two of three variables, *i.e.,* $\mathbf{G}$, $\mathbf{S}$, and $\mathbf{W}$, the remaining one can be optimized. After the iteration optimization, $\mathbf{S}$ is aligned to $\mathbf{G}$ by the adjustment of $\mathbf{W}$, while $\mathbf{W}$ is optimized by the adjustment of $\mathbf{G}$ and $\mathbf{S}$, and thus the intrinsic dimensions of $\mathbf{X}$ (via $\mathbf{W}$) is approximately approached. As a consequence, the construction of both $\mathbf{G}$ and $\mathbf{S}$ are with the high quality of the data (controlled by $\mathbf{W}$), so they become the ideal affinity matrices of $\mathbf{X}$. This is different from the previous spectral clustering methods, such as learning a fixed local representation affinity matrix in (Elhamifar and Vidal 2013; He and Niyogi 2003; Ng et al. 2002), learning a fixed global representation affinity matrix in (Liu et al. 2013; Lu et al. 2012), and learning a dynamic global representation affinity matrix in (Nie et al. 2016; Nie and Huang 2016), *from original data*.

It is noteworthy that Eq. (4) solves the first issue of learning affinity matrix on the quality of the data, but does not touch the last two issues of spectral clustering methods on explicitly yielding the clustering result, *i.e.,* the removal of the k-means clustering step and the suboptimal clustering result of the two-step strategy.

One-step spectral clustering

In graph theory, if an n-vertex graph $\mathbb{S}$ has exactly c connected components, where any two vertices are connected to each other by paths, then we can permutate its affinity matrix (constructed by these n data points) to a new matrix. In the resulting new matrix, the data points in the same connected components are put together to form a block. As a result, the resulting matrix becomes a block diagonal matrix with c blocks (where c is the number of clusters) (Mohar et al. 1991). That is, the data points in the same blocks can be regarded as having the same cluster membership, and n data points form c clusters. In this case, we can say that the matrix has an explicit clustering result.

In our case, if we want the intrinsic affinity matrix $\mathbf{S}$ to have explicit clustering result, then the graph constructed by $\mathbf{S}$ should have exactly c connected components, which is called '*clustering result constraint*' ($\mathbb{K}$ for short) in this paper. To do this, we first follow the literature (Chung 1997; Mohar et al. 1991) to have Theorem 1 as follows:

Theorem 1. *The number of connected components of the graph $\mathbb{S}$ is equal to the multiplicity of 0 as an eigenvalue of the Laplacian matrix $\mathbf{L}$*[1].

Theorem 1 implies that "the graph $\mathbb{S}$ has c connected components $\Leftrightarrow$ the Laplacian matrix $\mathbf{L}$ has c zero eigenvalues". By rearranging the eigenvalues of $\mathbf{L}$ as an ascending-order set of $\{\lambda_1, ..., \lambda_n\}$, we relax the above constraint to the following: "$\mathbf{L}$ has c zero eigenvalues $\Leftrightarrow \sum\limits_{i=1}^{c} \lambda_i \to 0$" according to (Moslehian 2012). By following the Ky Fan's theorem in (Fan 1949), we have:

$$\sum_{i=1}^{c} \lambda_i \to 0 \Leftrightarrow \left\{ \begin{array}{l} \min\limits_{\mathbf{W}} tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}), \\ \qquad s.t., \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_c, \end{array} \right. \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ (*i.e.,* $d' = c$) is the transformation matrix and $\mathbf{I}_c \in \mathbb{R}^{c \times c}$ is an identity matrix. Thus we obtain our final objective function for conducting one-step spectral clustering as follows:

$$\begin{aligned} \min_{\mathbf{G},\mathbf{S},\mathbf{W}} & \sum_{i,j}^{n} (\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 g_{i,j} + \alpha \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{i,j}) \\ & + \beta \sum_{i=1}^{n} \|\mathbf{g}_i - \mathbf{s}_i\|_2^2 + \gamma \|\mathbf{W}\|_{2,1}, \\ & \text{s.t.}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_c, \mathbf{G} \in \mathcal{C}, \mathbf{S} \in \mathcal{C}, \end{aligned} \tag{6}$$

Eq. (6) considers the optimization of $\mathbf{S}$ as a function of $\mathbf{G}$, $\mathbf{W}$, and $\mathbb{K}$, *i.e.,* $\mathbf{S} = \mathbb{F}(\mathbf{G}, \mathbf{W}, \mathbb{K})$ where $\mathbb{F}$ denotes a function operator, as well as considers the optimization of $\mathbf{W}$ directly influenced by $\mathbf{S}$ and indirectly influenced by $\mathbf{G}$. Specifically, $\mathbf{G}$ and $\mathbf{W}$ are used to help learn $\mathbf{S}$ with the high quality of the data, while the clustering result constraint $\mathbb{K}$ is used to directly make $\mathbf{S}$ as the clustering result. As a consequence, although we have no prior knowledge on either $\mathbf{S}$ or $\mathbf{W}$, Eq. (6) makes conduct a one-step spectral clustering to learn an intrinsic affinity matrix $\mathbf{S}$ from the intrinsic subspace $\mathbf{W}$. That is, besides finding the intrinsic subspace (via $\mathbf{W}$) of the original feature space, Eq. (6) also enables the resulting intrinsic affinity matrix $\mathbf{S}$ to 1) measure the real similarity among data points in the intrinsic subspace, and 2) be the final clustering result.

Optimization

The objective function in Eq. (6) is not jointly convex with respect to the three variables, *i.e.,* $\mathbf{S}$, $\mathbf{W}$, and $\mathbf{G}$. In this paper, we employ the framework of Iteratively Reweighted Least Square (IRLS) (Björck 1996) to solve Eq. (6), by iteratively optimizing each of the parameters (*i.e.,* $\mathbf{S}$, $\mathbf{W}$, and $\mathbf{G}$) while fixing the remaining parameters.

i) Update W by fixing S and G By fixing $\mathbf{S}$ and $\mathbf{G}$, we have the following objective function:

$$\begin{aligned} \min_{\mathbf{W}} \; & tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \gamma \|\mathbf{W}\|_{2,1}, \\ & \text{s.t.}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_c, \end{aligned} \tag{7}$$

By setting $\mathbf{X}^T\mathbf{W} - \mathbf{Z} = 0$ and $\mathbf{W} - \mathbf{M} = 0$, we then employ the framework of Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011) to optimize $\mathbf{W}$ with the corresponding augmented Lagrangian as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{M},\mathbf{Z}} & tr(\mathbf{Z}^T\mathbf{L}\mathbf{Z}) + \gamma \|\mathbf{M}\|_{2,1} + \rho_1 \|\mathbf{M} - \mathbf{W} + \mathbf{U}\|_F^2 \\ & + \rho_2 \|\mathbf{Z} - \mathbf{X}^T\mathbf{W} + \mathbf{V}\|_F^2, \; s.t. \; \mathbf{Z}^T\mathbf{Z} = \mathbf{I}_c, \end{aligned} \tag{8}$$

In each iteration of ADMM, the closed form solution of the variables $\mathbf{W}$, M, and $\mathbf{Z}$, can be obtained by:

$$\left\{ \begin{array}{l} \mathbf{W} = (\rho_1 \mathbf{I}_d + \rho_2 \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{B}, \\ \mathbf{m}_i = \max\{\|(\mathbf{W} - \mathbf{U})^i\|_2^2 - \frac{\gamma}{\rho_1}, 0\} \frac{(\mathbf{W}-\mathbf{U})^i}{\|(\mathbf{W}-\mathbf{U})^i\|_2^2}, \\ \mathbf{Z} = \mathbf{A}\mathbf{A}^T, \end{array} \right. \tag{9}$$

where $\mathbf{B} = \rho_1(\mathbf{M}+\mathbf{U}) + \rho_2\mathbf{X}(\mathbf{Z}+\mathbf{V})$, $\mathbf{I}_d$ is an $d \times d$ identity matrix, and $(\mathbf{W}-\mathbf{U})^i$ represents the i-th row of $\mathbf{W}-\mathbf{U}$, $i = 1, ..., d$. The result of Singular Value Decomposition (SVD) of $(\mathbf{X}^T\mathbf{W} - \mathbf{V})^T(\mathbf{R}^{-1})^T\mathbf{R}$ is denoted as $\mathbf{A}\mathbf{\Omega}\mathbf{A}^T$, where $\mathbf{R}$ is a lower triangular matrix (*i.e.,* $\mathbf{R}\mathbf{R}^T = \mathbf{L} + \rho_2\mathbf{I}_n$) and $\mathbf{I}_n$ denotes an $n \times n$ identity matrix.

ii) Update S by fixing W and G By fixing $\mathbf{W}$ and $\mathbf{G}$, we have the following objective function:

$$\begin{aligned} \min_{\mathbf{S}} \; & \sum_{i,j}^{n} \alpha \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{i,j} + \beta \sum_{i=1}^{n} \|\mathbf{g}_i - \mathbf{s}_i\|_2^2 \\ & \text{s.t.}, \mathbf{S} \in \mathcal{C}. \end{aligned} \tag{10}$$

We first calculate k nearest neighbors of each data point, and then set the value of $s_{i,j}$ as 0 if the j-th data point is not one of k nearest neighbors of the i-th data point, otherwise, the value of $s_{i,j}$ can be solved by Karush–Kuhn–Tucker (KKT) conditions, *i.e.,*

$$s_{i,j} = \left\{ \begin{array}{ll} \frac{e_{i,k+1} - e_{i,j}}{ke_{i,k+1} - \sum_{v=1}^{k} e_{i,v}}, & j \leq k, \\ 0, & j > k, \end{array} \right. \tag{11}$$

where $\mathbf{e}_i = \{e_{i,1}, ..., e_{i,n}\}$ is the descend order of $\mathbf{f}_i$ (where $f_{i,j} = \frac{\alpha}{2} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||_2^2, i, j = 1, ..., n$), and k is the number of nearest neighbors of i-th data point, which can be tuned by cross-validation methods.

iii) Update G by fixing W and S Similar to the optimization of $\mathbf{S}$, if the j-th data point is one of k nearest neighbors of the i-th data point, the close-form solution of $g_{i,j}$ is:

$$g_{i,j} = \left\{ \begin{array}{ll} \frac{e'_{i,k+1} - e'_{i,j}}{ke'_{i,k+1} - \sum_{v=1}^{k} e'_{i,v}}, & j \leq k, \\ 0, & j > k, \end{array} \right. \tag{12}$$

where $\mathbf{e}'_i = \{e'_{i,1}, ..., e'_{i,n}\}$ is the descend order of $\mathbf{f}'_i$ (where $f'_{i,j} = \frac{1}{2} ||\mathbf{x}_i - \mathbf{x}_j||_2^2, i = 1, ..., n$), and k is the number of nearest neighbors of i-th data point.

[1] where $\mathbf{L} = \mathbf{P} - \mathbf{S}$ and $\mathbf{P}$ is a diagonal matrix with the i-th element $p_{i,i}$ as $p_{i,i} = \sum\limits_{j=1}^{n} s_{i,j}, i = 1, ..., n.$

Table 1: The information of the benchmark datasets

Datasets	Data points	Features	Classes
Umist	165	3456	15
Wine	178	13	3
Ecoli	336	343	8
YaleB	640	2016	10
Jaffe	213	1024	10
Coil	1440	1024	20

Experimental analysis

We evaluate our one-step spectral clustering method and eight clustering methods on both synthetic datasets and real benchmark datasets, in terms of clustering ACCuracy (ACC) and Normalized Mutual Information (NMI).

Experimental result on synthetic datasets

The first synthetic dataset includes 200 3-D data points within 3 blocks, where the affinity matrix is 200×200 and the block size is 50×50, 100×100, and 50×50. It is noteworthy that the data points in the same block have the same cluster membership. In our experiments, we randomly generated noise with the noise level σ ($0 \leq \sigma \leq 1$) to the data points in the blocks and outside of the blocks (*i.e.,* outliers). The larger the value of σ, the larger the percentage of the noise is. Figs. 1(a) and 1(c) visualized the affinity matrices of two datasets with different noise levels, *i.e.,* $\sigma = 0.5$ and $\sigma = 0.9$, respectively. Figs. 1(b) and 1(d) illustrated the corresponding affinity matrices yielded by our method. It was obvious that our method could output clearly separated blocks. We then conducted clustering analysis using our method and the classic clustering method Normalize Cut (NCut) (Shi and Malik 2000) on these two datasets. The ACC results of these two methods are 100% for the dataset with a moderate noise level, *i.e.,* $\sigma = 0.5$, but our method (*i.e.,* 95% for ACC) outperformed NCut (*i.e.,* 85% for ACC) on the dataset with the high noise level (*i.e.,* $\sigma = 0.9$). This indicated that these two methods were robust to noise but our method was more robust than NCut.

Experimental result on benchmark datasets

Comparison methods The comparison methods include three classic clustering methods (*e.g.,* NCut, *k*-means (Hartigan and Wong 2013), and Ratio Cut (RCut) (Wang and Siskind 2003)), two global representation methods (*e.g.,* Low-Rank Representation (LRR) (Liu et al. 2013), and Constrained Laplacian Rank (CLR) (Nie et al. 2016)), and one local representation method (*e.g.,* Sparse subspace clustering (SSC) (Elhamifar and Vidal 2013)). The brief description of the comparison methods in this paper is described as follows:

- **NCut** identifies the data points into *k* disjoint vertex sets so that the weights of the edges between the vertex sets are minimum, while giving an graph measuring the similarity among samples.
- **k-means** aims to partition all the data points into *k* clusters/groups in which each data point belongs to the cluster with the nearest mean. Since *k*-means is sensitive to the initial values, we ran this algorithm 10 times and reported their averaging result.
- **RCut** is a graph-based clustering method and was designed to find partitions minimizing the ratio of the sums of two different weights.
- **LRR** seeks the lowest rank representation among all the other data points that can represent the data samples as linear combinations of the bases in a given dictionary.
- **CLR** is a graph-based method by learning a graph with exactly k connected components where k is the number of clusters.
- **SSC** is based on the fact that each point in a union of subspaces has a sparse representation with respect to a dictionary formed by all other data points. It could cluster data drawn from multiple low-dimensional linear or affine subspaces embedded in a high-dimensional space.

Datasets The used datasets (shown in Table 1 for more detail) include image datasets (such as Umist, Ecoli, YaleB, Coil and Jaffe) and the datasets Wine is downloaded from (Zhong and Fukushima 2007).

Umist (Graham and Allinson 1995) consists of 575 face images of 20 people. Each of images covers a range of poses from profile to frontal views and is disposed into 23×28 pixels.

Ecoli (Athitsos and Sclaroff 2005) contains 336 data samples drown from 8 groups and each of the samples has 343 features.

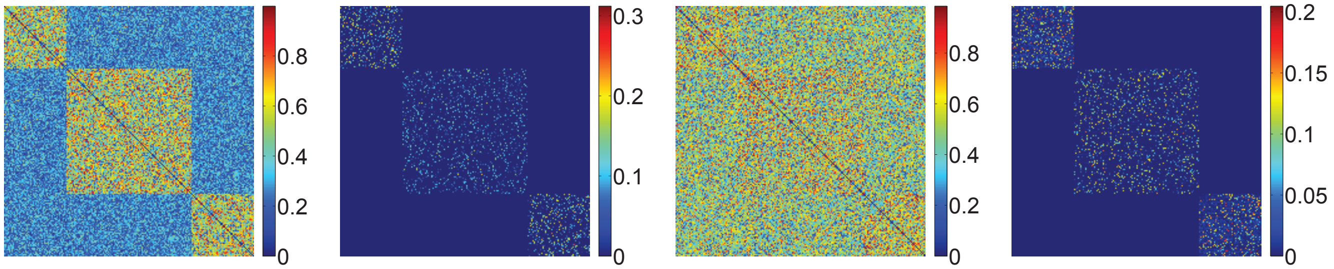
YaleB (Lee, Ho, and Kriegman 2005) has 16128 facial images of 28 persons under 9 postures (center-light, happy, w/no glasses, normal, sad, sleepy, surprised, and so on) and 64 illumination conditions. All the images are cut into 2016 dimensions. In our experiments, we used 64 images of the first 10 people to test the clustering performance of all the methods.

Coil (Rate and Retrieval 2011) consists of 1440 grid images of 20 objects and all the images are cut into 1024 features where the backgrounds of all the images have been discarded.

Jaffe (Nie, Wang, and Huang 2014) has totally 213 images and each of 10 distinct persons with 7 facial expressions (6 basic facial expressions plus 1 neutral). Each image is preprocessed into 256 pixels so that the number of the dimensions of each image is 256.

Wine is the results of chemical analysis of wines grown in the same region and is derived from 3 cultivars. It has 178 samples with 13 features.

Experimental analysis For fair comparison, we used the self-tune Gaussian method to construct the initial affinity matrix and set the number of *k* (in a range of $\{5, 10, 15\}$) by following the setting of CLR for the methods (*i.e.,* SSC, LRR, RCut, and NCut); We repeated the experiments 100 times for the methods (*i.e.,* k-means, RCut, and NCut) and reported the average performance of the k-means to eliminate the random error; We tuned all the parameters in a range of $\{0.01, 1, 10, 100\}$ to report the best performance for the spectral clustering methods (*i.e.,* SSC, LRR, CLR, and our method).



(a) Original graph ($\sigma = 0.5$) (b) Proposed graph ($\sigma = 0.5$) (c) Original graph ($\sigma = 0.9$) (d) Proposed graph ($\sigma = 0.9$)

Figure 1: Clustering result on a block diagonal synthetic data by our proposed method.

Table 2: Clustering accuracy (ACC) of all methods on benchmark datasets.

		k-means	RCut	NCut	CLR	SSC	LRR	Proposed
ACC	Umist	43.65	59.13	59.13	69.22	63.22	64.00	**79.83**
	Wine	53.65	71.03	71.03	**72.47**	65.17	71.01	**72.47**
	Ecoli	35.21	48.04	47.44	50.86	47.02	46.61	**53.27**
	YaleB	29.38	33.59	33.75	30.16	36.56	31.56	**45.94**
	Jaffe	74.21	96.24	96.24	81.69	82.63	81.60	**96.71**
	Coil	73.82	79.58	79.44	85.35	80.63	85.14	**94.72**

Table 3: NMI of all methods on benchmark datasets.

		k-means	RCut	NCut	CLR	SSC	LRR	Proposed
NMI	Umist	63.40	80.11	80.12	83.89	75.05	64.48	**88.48**
	Wine	33.40	37.14	37.14	**39.27**	36.76	35.56	**39.27**
	Ecoli	41.50	39.61	39.12	42.56	37.83	41.69	**42.77**
	YaleB	29.59	31.89	32.54	45.07	38.16	31.03	**49.70**
	Jaffe	89.38	96.23	96.23	90.44	88.25	91.35	**96.71**
	Coil	77.94	88.94	88.77	94.50	82.26	85.94	**97.69**

We reported all clustering result in Tables 2 and 3, which showed that our method achieved the best performance, compared to all the comparison methods. The reason is that our method could learn a robust affinity matrix from the intrinsic subspace of the original feature space and thus resulting a one-step clustering to yield robust clustering result.

Conclusion

This paper proposed a novel one-step spectral clustering method by learning the affinity matrix (also the clustering result) from the intrinsic subspace of the original feature space. Different from the previous two-step spectral clustering methods, our proposed method directly outputs the clustering result for avoiding the suboptimal issue of the two-step strategy. Experimental result on both synthetic datasets and benchmark datasets showed that our proposed one-step spectral clustering method outperformed the comparison clustering methods.

In our future work, this framework will be extended to conduct clustering on the datasets with incomplete data since missing data are often found in real applications (Zhu et al. 2007; Zhu, Suk, and Shen 2014).

Acknowledgement

This work was supported in part by the China "1000-Plan" National Distinguished Professorship; the Nation Natural Science Foundation of China (Grants No: 61263035, 61573270 and 61672177), the China 973 Program (Grant No: 2013CB329404); the China Key Research Program (Grant No: 2016YFB1000905); the Guangxi Natural Science Foundation (Grant No: 2015GXNSFCB139011); the China Postdoctoral Science Foundation (Grant No: 2015M570837); the Innovation Project of Guangxi Graduate Education under grants YCSZ2016045 and YCSZ2016046; the Guangxi High Institutions' Program of Introducing 100 High-Level Overseas Talents; the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing; the Guangxi "Bagui" Teams for Innovation and Research, and the project "Application and Research of Big Data Fusion in Inter-City Traffic Integration of The Xijiang River - Pearl River Economic Belt(da shu jv rong he zai xijiang zhujiang jing ji dai cheng ji jiao tong yi ti hua zhong de ying yong yu yan jiu)".

References

Athitsos, V., and Sclaroff, S. 2005. Boosting nearest neighbor classifiers for multiclass recognition. In *CVPR Workshops*, 45–45.

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, 585–591.

Björck, A. 1996. *Numerical methods for least squares problems.*

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1):1–122.

Chung, F. R. K. 1997. Spectral graph theory cbms series. *American Mathematical Society* 9(6):55.

Elhamifar, E., and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence* 35(11):2765–2781.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America* 35(11):652.

Graham, D. B., and Allinson, N. M. 1995. Characterising virtual eigensignatures for general purpose face recognition. *Journal of Nursing Management* 3(2):87–91.

Gu, Q.; Li, Z.; and Han, J. 2011. Joint feature selection and subspace learning. In *IJCAI*, volume 22, 1294.

Hartigan, J. A., and Wong, M. A. 2013. A k-means clustering algorithm. *Applied Statistics* 28(1):100–108.

He, X., and Niyogi, P. 2003. Locality preserving projections. In *NIPS*, 153–160.

Lee, K. C.; Ho, J.; and Kriegman, D. J. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 27(5):684–698.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1):171–184.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, 347–360.

Luo, D.; Nie, F.; Ding, C.; and Huang, H. 2011. Multi-subspace representation and discovery. In *ECML/PKDD*, 405–420.

Mohar, B.; Alavi, Y.; Chartrand, G.; and Oellermann, O. 1991. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* 2(871-898):12.

Moslehian, M. S. 2012. Ky fan inequalities. *Linear and Multilinear Algebra* 60(11-12):1313–1325.

Ng, A. Y.; Jordan, M. I.; Weiss, Y.; et al. 2002. On spectral clustering: Analysis and an algorithm. *NIPS* 2:849–856.

Nie, F., and Huang, H. 2016. Subspace clustering via new low-rank model with discrete group structure constraint. In *IJCAI*, 1874–1880.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, 1962–1968.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *KDD*, 977–986.

Peng, X.; Zhang, L.; and Yi, Z. 2013. Scalable sparse subspace clustering. In *ICCV*, 430–437.

Rate, C., and Retrieval, C. 2011. Columbia object image library (coil-20). *Computer*.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8):888–905.

Vidal, R. 2011. Subspace clustering. *IEEE Signal Process. Mag.* 28(2):52–68.

Wang, S., and Siskind, J. M. 2003. Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(6):675–690.

Wang, S.; Yuan, X.; Shen, J.; Yao, T.; and Yan, S. 2011. Efficient subspace segmentation via quadratic programming. In *AAAI*.

Yu, K.; Zhang, T.; and Gong, Y. 2009. Nonlinear learning using local coordinate coding. In *NIPS*, 2223–2231.

Zhong, P., and Fukushima, M. 2007. Regularized nonsmooth newton method for multi-class support vector machines. *Optimization Methods & Software* 22(1):225–236.

Zhu, X.; Zhang, S.; Zhang, J.; and Zhang, C. 2007. Cost-sensitive imputing missing values with ordering. In *AAAI*, 1922–1923.

Zhu, X.; Huang, Z.; Shen, H. T.; Cheng, J.; and Xu, C. 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition* 45(8):3003–3016.

Zhu, X.; Huang, Z.; Yang, Y.; Shen, H. T.; Xu, C.; and Luo, J. 2013. Self-taught dimensionality reduction on the high-dimensional small-sized data. *Pattern Recognition* 46(1):215–229.

Zhu, X.; Li, X.; Zhang, S.; Ju, C.; and Wu, X. 2016. Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhu, X.; Suk, H.-I.; and Shen, D. 2014. Multi-modality canonical feature selection for alzheimers disease diagnosis. In *MICCAI*, 162–169.