RS^{3}CIS: Robust Single-Step Spectral Clustering with Intrinsic Subspace

Yun Xiao,[1] Pengzhen Ren,[1] Zhihui Li,[2*] Xiaojiang Chen,[1] Xin Wang,[1,3] Dingyi Fang [1]
[1]School of Information Science and Technology, Northwest University, Xi'an 710127, P.R. China
[2]School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[3]Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N1N4, Canada
{yxiao, xjchen, xinwang, dyf}@nwu.edu.cn,{pzhren, zhihuilics}@gmail.com

Abstract

Spectral clustering has been widely adopted because it can mine structures between data clusters. The clustering performance of spectral clustering depends largely on the quality of the constructed affinity graph, especially when the data has noise. Subspace learning can transform the original input features to a low-dimensional subspace and help to produce a robust method. Therefore, how to learn an intrinsic subspace and construct a pure affinity graph on a dataset with noise is a challenge in spectral clustering. In order to deal with this challenge, a new Robust Single-Step Spectral Clustering with Intrinsic Subspace (RS^{3}CIS) method is proposed in this paper. RS^{3}CIS uses a local representation method that projects the original data into a low-dimensional subspace through a row-sparse transformation matrix and uses the $\ell_{2,1}$-norm of the transformation matrix as a penalty term to achieve noise suppression. In addition, RS^{3}CIS introduces Laplacian matrix rank constraint so that it can output an affinity graph with an explicit clustering structure, which makes the final clustering result to be obtained in a single-step of constructing an affinity matrix. One synthetic dataset and six real benchmark datasets are used to verify the performance of the proposed method by performing clustering and projection experiments. Experimental results show that RS^{3}CIS outperforms the related methods with respect to clustering quality, robustness and dimension reduction.

Introduction

Spectral clustering has long been favored by researchers because of its ability to mine structures between data clusters. The quality of the constructed affinity graph is crucial for the clustering performance of spectral clustering, especially when the data has noise. The existing optimization methods for constructing affinity graphs in spectral clustering can be roughly divided into two categories: global representation (Nie et al. 2016; 2017; Nie, Li, and Li 2016; 2017; Ren et al. 2018) and local representation (Nie, Wang, and Huang 2014; Zhu et al. 2017a).

Local representation methods tend to be more robust than global representation methods due to using its neighbor nodes to represent each data point, which can effectively remove the influence of noise points (especially outliers) (Roweis and Saul 2000; Yu 2009). Moreover, subspace learning helps to produce a robust method, while feature selection helps to produce an interpretable method (Gu, Li, and Han 2011; Zhu et al. 2017b). This is also an important reason that local representation methods are widely concerned. In addition, the high-dimensional feature space of the original dataset is actually located in a low-dimensional subspace (Zhu et al. 2012; Vidal 2011).

The above advantages motivate researchers to focus on local representations of data. The projected clustering with adaptive neighbors (PCAN) method proposed by Nie et al. (Nie, Wang, and Huang 2014) learns the data similarity matrix and clustering structure simultaneously by projecting the original dataset into a low-dimensional subspace. Although PCAN has achieved certain noise suppression effect by reducing the data dimension, the clustering performance of PCAN is still greatly affected when there is more noise in the data.

In addition, the $\ell_{2,1}$-norm has a good effect in feature selection, noise suppression and redundant information removal, etc., which has attracted the attention of many researchers (Zhu et al. 2017a; Liao et al. 2018; Nie et al. 2010). Inspired by these papers, we apply the $\ell_{2,1}$-norm to spectral clustering to obtain a robust clustering performance.

Therefore, in this paper, we construct a low-dimensional intrinsic subspace by projecting the original dataset into a low-dimensional subspace using a row-sparse transformation matrix. The suppression of noise and outliers is achieved by a row-sparse transformation matrix. At the same time, the intrinsic affinity matrix is further studied according to the low-dimensional intrinsic subspace. Different from the previous method (Nie, Cai, and Li 2017; Li et al. 2018; Nie et al. 2016), which learns an affinity matrix directly from the original dataset, this paper proposes a new learning strategy for the intrinsic affinity matrix. We introduce a row-sparse penalty term in our method and avoid those trivial solutions by adjusting its coefficients. In terms of parameters, our method mainly focuses on the optimization of the dimension reduction and the row-sparse penalty coefficient, which is relatively simple. Moreover, a new iterative update method is used to optimize the newly added row-sparse items, and this optimization problem is well solved with our method.

In addition, affinity matrices constructed based on tradi-

* Corresponding author

tional spectral clustering algorithms are often criticized by researchers (Nie, Li, and Li 2017; Li et al. 2018; Nie et al. 2017; Zhu et al. 2017a) because they do not have an explicit clustering structure. And these methods (Huang, Nie, and Huang 2013; Nie, Li, and Li 2016) often require a method like K-means for post-processing to obtain the final label of the dataset. However, the results obtained by this two-step approach may be sub-optimal. In order to solve the above problem, in this paper, we introduce Laplacian matrix rank constraints in the learning of affinity matrices. The Laplacian matrix rank constraint can make the learned affinity graph have an explicit cluster structure with our method. Therefore the final clustering label can be obtained in a single-step of constructing an affinity matrix. For the above reasons, we call our method Robust Single-Step Spectral Clustering with Intrinsic Subspace (RS^{3}CIS).

The proposed RS^{3}CIS is compared with four classical clustering methods and two related clustering methods on the synthetic dataset and the real world benchmark dataset. The final experimental results show that our method RS^{3}CIS has better clustering quality, robustness, and dimension reduction than that of the related methods.

In general, the main contributions of this paper are as follows:

- In our method RS^{3}CIS, subspace learning is performed by applying orthogonal constraints to the hash matrix, and feature selection is achieved by adding row-sparse penalty terms. Therefore, our method is robust and interpretable.
- The proposed method can simultaneously learn an intrinsic subspace and an affinity graph with an explicit clustering structure. Thus the final clustering result can be obtained in a single-step of constructing an affinity matrix.
- Since the optimization problem is non-smooth and difficult to solve, a new optimization method of $\ell_{2,1}$-norm is proposed and the final optimization problem is solved effectively by an iterative update algorithm.
- We have designed extensive experiments to verify the effectiveness of our method RS^{3}CIS. And in the projection experiments, we prove that our method still maintains superiority under different reduced dimensions.

The Proposed RS^{3}CIS Methodology

Notation

In the entire paper, we use uppercase letters to represent matrices. For example, in the matrix $M \in \mathbb{R}^{n_1 \times n_2}$, we define m_i to represent the i-th column vector of the matrix M, and m_{ij} represents the j-th element of the column vector m_i. The Frobenius norm and $\ell_{2,1}$-norm of matrix M are represented as $\|M\|_F = \sqrt{\sum_i \sum_j m_{ij}^2}$, and $\|M\|_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2}$. In addition, we further define the transpose, the trace, the rank, and the inverse of matrix M, as M^T, $Tr(M)$, $rank(M)$, M^{-1}, respectively. $\mathbf{1}$ in the paper represents a unit column vector, and I represents an identity matrix.

Initial affinity graph revisit

Constructing an affinity graph with high quality is critical to the performance of clustering tasks. Given a dataset $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{d \times n}$ with k clusters, d is the feature dimension in the dataset X, n is the number of data points, and $x_i \in \mathbb{R}^{d \times 1}$ is the i-th data point in dataset X. A natural idea of constructing an affinity matrix is to measure similarity based on the distance between different data points. That is, there is a high similarity between data points that are close to each other in distance, and a low similarity between data points that are far away from each other. Based on this idea, Nie et al. (Nie, Wang, and Huang 2014) proposed the following optimization goals:

$$\begin{aligned} &\min_S \sum_{i,j=1}^{n} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2, \\ &s.t., S \geq 0, s_i^T \mathbf{1} = 1, rank(L_S) = n - c, \end{aligned} \tag{1}$$

where $S \in \mathbb{R}^{n \times n}$ is the affinity matrix with the j-th element as $s_j \in \mathbb{R}^{n \times 1}$ and the i-th element of s_j as s_{ij}, $L_S = D_S - (S^T + S)/2$ is the Laplacian matrix of the affinity matrix S, the degree matrix D_S is a diagonal matrix with the i-th element as $d_{ii} = \sum_j (s_{ij} + s_{ji})/2$, c is the number of 0 eigenvalues of the Laplacian matrix L_S, and s_{ij} represents the similarity between the data point x_i and the data point x_j.

The model in the problem (1) has an adaptive neighborhood and can flexibly explore the affinity relationship between data points. However, it cannot learn the local structure of the intrinsic subspace with the initial affinity graph construction method, thus the performance of clustering is easily affected by noise and outliers.

Moreover, Zhu et al. (Zhu et al. 2017a) also proposed a similar optimization objective function. They remove the second item in the optimization problem (1), and modify the constraints to $S \in \mathcal{C}$, where $\mathcal{C} = \{\forall i | s_i^T \mathbf{1} = 1,\ s_{ii} = 0,\ s_{ij} \geq 0, if\ j \in \mathbb{N}(i), otherwise\ 0.\}$, $\mathbb{N}(i)$ represents the set of neighbor nodes of the i-th data point. The number of neighbor nodes is determined by cross-validation. This method implements a local representation of each data point. However, the number of its neighbor points is fixed, which is inconvenient to the local flexible representation of the data points. In addition, the above optimization problem does not solve the local representation problem of the intrinsic subspace on the dataset.

Local representation of intrinsic subspace

Based on the superiority of local representation in the construction of affinity graphs, we introduce a row-sparse transformation matrix into our method. That is, we define a transformation matrix $W \in \mathbb{R}^{d \times d'}$ (where $d' \leq d$) to project the original dataset X to a low-dimensional intrinsic subspace, where d' is the feature dimension of the dataset X projected to the low-dimensional subspace. This low-dimensional intrinsic subspace is represented as $W^T X$, which allows us to further explore the intrinsic structure existing in the original dataset and effectively suppress the noise points. Therefore,

we design the following optimization object function:

$$\min_{S,W} \sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1} + \alpha \|S\|_F^2,$$
$$s.t., S \geq 0, s_i^T \mathbf{1} = 1, W \in \mathbb{R}^{d \times d'}, W^T X X^T W = I, \tag{2}$$

where γ and α are the tuning parameters, s_{ij} represents the similarity between the data points in the corresponding low-dimensional subspace (the s_{ij} in the following equations has the same meaning if the data point x_i and the data point x_j are projection through the transformation matrix W.)

In problem (2), feature selection is performed by applying a penalty term $\|W\|_{2,1}$ to row sparsity, which can suppress noise and remove redundant features. Another penalty term $\|S\|_F^2$ is used to avoid those insignificant solutions, in which only the similarity between the points closest to the data point x_i is assigned to a value of 1, and the similarity of the other points is 0. The addition of the sparse penalty term $\|S\|_F^2$ allows our method to construct a more flexible affinity graph with adaptive neighbors. Applying orthogonal constraints to the scattering matrix $W^T X X^T W$ is actually for intrinsic subspace learning which transfers the d-dimensional feature space into the statistically uncorrelated d'-dimensional intrinsic subspace on the original dataset X.

Robust Single-Step Spectral Clustering with Intrinsic Subspace

In order to obtain an affinity graph with an explicit clustering structure, we introduce Laplacian matrix rank constraints in affinity matrices learning. An important property about the Laplacian matrix (Fan 1997; Mohar 1991) is presented as follows:

Theorem 1. *The multiplicity c of the eigenvalue 0 of the Laplacian matrix L_S (nonnegative) is equal to the number of connected components in the graph with the similarity matrix S.*

Based on Theorem 1, we add a Laplacian matrix rank constraint to the affinity matrix S in problem (2). The learned affinity graph are permutated by using the added Laplacian matrix rank constraint to obtain an affinity graph with exactly c connected components. Therefore, problem (2) becomes

$$\min_{S,W} \sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1} + \alpha \|S\|_F^2,$$
$$s.t., S \geq 0, s_i^T \mathbf{1} = 1, rank(L_S) = n - c,$$
$$W \in \mathbb{R}^{d \times d'}, W^T X X^T W = I. \tag{3}$$

For optimal solution to problem (3), we can project the original dataset to a low-dimensional intrinsic subspace, and learn a row-sparse transformation matrix W and an intrinsic affinity graph S of a low-dimensional subspace with an explicit clustering structure. Therefore, the method is interpretable and robust. In addition, based on the learned intrinsic affinity graph, the final clustering label can be obtained directly without any post-processing steps. Therefore, we name the method expressed in problem (3) as Robust Single-Step Spectral Clustering with Intrinsic Subspace (RS^{3}CIS).

In fact, the optimization of problem (3) is difficult. This is because the degree matrices D_S and the Laplacian matrix $L_S = D_S - (S^T + S)/2$ in the problem (3) are both dependent on the affinity matrix S, but the affinity matrix S is the unknown amount that we need to solve. At the same time, the Laplacian matrix rank constraint $rank(L_S) = n - c$ is also difficult to handle. In addition, the orthogonal constraints imposed by the scattering matrix and the optimization of the $\ell_{2,1}$-norm applied to the transformation matrix are two challenges that cannot be ignored. In the next section, we present an effective iterative optimization method to address these challenges.

Optimization of RS^{3}CIS

For the problem (3), let $\sigma_i(L_S)$ denote the i-th smallest eigenvalue of L_S. Because Laplacian matrix L_S is positively semidefinite, so $\sigma_i(L_S)$ is non-negative. Given a sufficiently large η, problem (3) can be rewritten as follows:

$$\min_{S,W} \sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1} + \alpha \|S\|_F^2 + 2\eta \sum_{i=1}^{c} \sigma_i(L_S),$$
$$s.t., S \geq 0, s_i^T \mathbf{1} = 1, W \in \mathbb{R}^{d \times d'}, W^T X X^T W = I. \tag{4}$$

Because η is large enough, and $\sigma_i(L_S) \geq 0$ for each i, the optimal solution S to the problem (4) ensures the second term $\sum_{i=1}^{c} \sigma_i(L_S)$ equal to 0 and the constraint $rank(L_S) = n - c$ can be satisfied. Furthermore, according to Ky Fan's Theory (Fan 1949), the following equation is true:

$$\sum_{i=1}^{c} \sigma_i(L_S) = \min_{F} Tr(F^T L_S F),$$
$$s.t., F \in \mathbb{R}^{n \times c}, F^T F = I. \tag{5}$$

According to Eq.(5), the problem (4) can be further equivalent to

$$\min_{S,W} \sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1} + \alpha \|S\|_F^2 + 2\eta Tr(F^T L_S F),$$
$$s.t., S \geq 0, s_i^T \mathbf{1} = 1, F \in \mathbb{R}^{n \times c}, W \in \mathbb{R}^{d \times d'},$$
$$W^T X X^T W = I, F^T F = I. \tag{6}$$

Then we optimize the problem (6) with an effective iterative algorithm.

i. Update F by fixing S and W By fixing S and W, problem (6) can be rewritten as

$$\min_{F \in R^{n \times c}, F^T F = I} Tr(F^T L_S F). \tag{7}$$

Algorithm 1: Optimization of transformation matrix W in problem (10)

Input:
Dataset $X \in \mathbb{R}^{d\times n}$, Laplace matrix L_S, parameters γ, a large enough η and $\varepsilon = 1e-8$.

Output: Row-sparse transformation matrix W.

Initialize $W^{(0)}$ by the optimal solution to the following problem:

$$\min_W Tr(W^T X L_S X^T W), s.t., W^T X X^T W = I.$$

1. Calculate L_S by Eq.(1).
2. Calculate A and B by Eq.(11).
3. Calculate D by Eq.(12) and calculate G by Eq.(13).

repeat

i. Update W: The transformation matrix W is updated by Eq.(13).

ii. Update δ: Calculate $\delta = |J(W^{(t+1)}) - J(W^{(t)})|$, where $J(W^{(t)}) = Tr(W^{(t)T} X L_S X^T W^{(t)}) + \frac{\gamma}{2}\|W\|_{2,1}$.

iii. Update t: $t \leftarrow t+1$.

until $\delta < \varepsilon$

By the c eigenvectors of L_S corresponding to the c smallest eigenvalues, the optimal solution of F is well composed with this method.

ii. Update W by fixing S and F By fixing S and F, problem (6) can be rewritten as

$$\begin{aligned} &\min_W \sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma\|W\|_{2,1}, \\ &s.t., W \in \mathbb{R}^{d\times d'}, W^T X X^T W = I. \end{aligned} \tag{8}$$

Suppose each node i is assigned to a function value as $W^T x_i \in \mathbb{R}^{c\times 1}$, the following equation is established:

$$\sum_{i,j}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} = 2Tr(W^T X L_S X^T W). \tag{9}$$

According to Eq.(9), problem (8) can be rewritten as

$$\begin{aligned} &\min_W Tr(W^T X L_S X^T W) + \frac{\gamma}{2}\|W\|_{2,1}, \\ &s.t., W \in R^{d\times d'}, W^T X X^T W = I. \end{aligned} \tag{10}$$

Unfortunately, due to the introduction of the $\ell_{2,1}$-norm, it is difficult to optimize the problem (10) for W. In this paper, we use a simple and effective iterative approach to solve this optimization problem.

Problem (10) is a constrained optimization problem, and we use the Lagrangian multiplier method to solve this problem. The Lagrangian function of problem (10) is

$$\mathcal{L}(W) = Tr(W^T A W) + \frac{\gamma}{2}\|W\|_{2,1} - Tr(\Lambda(W^T B W - I)), \tag{11}$$

where $A = X L_S X^T, B = X X^T$ and Λ is a diagonal matrix that is used to enforce the constraint on problem (10). Let the derivative of Eq.(11) $w.r.t.$ be zero, that is

$$\frac{\partial \mathcal{L}(W)}{\partial W} = (A + A^T)W + 2\gamma D W - 2BW\Lambda = 0, \tag{12}$$

where $D = diag(\frac{1}{4\|\widetilde{W}(1,:)\|_2}, ..., \frac{1}{4\|\widetilde{W}(d,:)\|_2})$, and $\widetilde{W}$ is the current solution. Since $A = A^T$, Eq.(12) can be rewritten as follows by simple algebraic operations:

$$GW = W\Lambda, \tag{13}$$

where $G = B^{-1}(A + \gamma D)$. Note that Λ is a diagonal matrix, therefore, the optimal transformation matrix W is actually a matrix composed of eigenvectors corresponding to the first c smallest eigenvalues except the zero eigenvalues in the Eigen-equation $Gw_i = \lambda_i w_i (i = 1, 2, ..., d')$, where w_i is the i-th column vector of the transformation matrix W, and λ_i is the i-th smallest eigenvalue except the zero eigenvalues in the Eigen-equation.

We summarize the detailed algorithm for solving problem (10) into the Algorithm 1. Liao et al (Liao et al. 2018), demonstrate the convergence of optimization targets with the $\ell_{2,1}$-norm. It is similar to our problem and the convergence of our optimization targets can be obtained easily according to the this paper (Liao et al. 2018).

iii. Update S by fixing W and F By fixing W and F, problem (6) can be rewritten as

$$\begin{aligned} &\min_S \sum_{i,j=1}^{n} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + 2\eta Tr(F^T L_S F) + \alpha\|S\|_F^2, \\ &s.t., S \geq 0, s_i^T \mathbf{1} = 1. \end{aligned} \tag{14}$$

According to Eq.(9) and setting $Z = X^T W$, the problem (14) can be further rewritten as

$$\begin{aligned} &\min_S \sum_{i,j=1}^{n} (\|z_i - z_j\|_2^2 s_{ij} + \alpha s_{ij}^2 + \eta\|f_i - f_j\|_2^2 s_{ij}), \\ &s.t., s_i \geq 0, s_i^T \mathbf{1} = 1, \end{aligned} \tag{15}$$

where z_i is the i-th column vector of Z, f_i is the i-th column vector of F. Since problem (15) is independent of different i, so we can solve the following problem individually for each i:

$$\begin{aligned} &\min_{s_i} \sum_{j=1}^{n} (\|z_i - z_j\|_2^2 s_{ij} + \alpha s_{ij}^2 + \eta\|f_i - f_j\|_2^2 s_{ij}), \\ &s.t., s_i \geq 0, s_i^T \mathbf{1} = 1. \end{aligned} \tag{16}$$

Denote

$$d_{ij}^z = \|z_i - z_j\|_2^2, d_{ij}^f = \|f_i - f_j\|_2^2, \tag{17}$$

and denote $d_i \in \mathbb{R}^{n\times 1}$ as a vector with the j-th element as $d_{ij} = d_{ij}^z + \eta d_{ij}^f$, then the problem (16) can be written in a vector form as

$$\min_{s_i^T \mathbf{1} = 1, s_i \geq 0} \|s_i + \frac{1}{2\alpha} d_i\|_2^2 \tag{18}$$

Algorithm 2: RS^{3}CIS optimization in problem (3)

Input:
dataset $X \in \mathbb{R}^{d\times n}$, cluster number k, parameters γ and α, a large enough η.

Output: Intrinsic affinity matrix $S \in R^{n\times n}$ with exactly $c = k$ connected components and row-sparse transformation matrix W.

Initialize S by the optimal solution to the following problem:

$$\min_{S} \sum_{j=1}^{n} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2, s.t., S \geq 0, s_i^T \mathbf{1} = 1.$$

repeat

i. Update F. F is formed by the c eigenvectors of $L_S = D_S - \frac{S^T+S}{2}$ corresponding to the c smallest eigenvalues.

ii. Update W. The transformation matrix W is updated by Algorithm 1.

iii. Update S. For each i, update the i-th row of S by solving the problem (18), where j-th element of vector d_i is defined as $d_{ij} = d_{ij}^z + \eta d_{ij}^f$.

until converge

This problem can be solved by a closed form solution which is put forward by (Nie, Wang, and Huang 2014). By updating s_i, we can get the matrix S with exactly k strong connected subgraphs. The matrix S corresponds to a graph with an explicit structure that can be used to obtain the final clustering directly.

We summarize the detailed algorithm for solving problem (3) into the Algorithm 2.

Update of parameters α and η In addition, the value of the regularization parameter α is arbitrary from zero to infinite, and its tuning is difficult. We also note that when we remove the Laplacian matrix rank constraint and the row-sparse matrix constraint in problem (3), the global representation in problem (3) will degenerate to the following form:

$$\begin{aligned} &\min_{S} \sum_{i,j=1}^{n} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2, \\ &s.t., S \geq 0, s_i^T \mathbf{1} = 1. \end{aligned} \tag{19}$$

Therefore we determine the value of the regularization parameter α by (Nie, Wang, and Huang 2014):

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} (\frac{K}{2} d_{i,K+1} - \frac{1}{2} \sum_{j=1}^{K} d_{ij}), \tag{20}$$

where K is the number of neighbor nodes. The initial value of η for acceleration of the method can be set to α. And η is constantly adjusted during the iteration of the algorithm 2 until the number of connected components of the Laplacian matrix L_S is equal to the number of clusters k.

Experiments

We use the clustering ACCuracy (ACC) to evaluate the performance of our method RS^{3}CIS through projection experiments and clustering experiments on both synthetic dataset and real benchmark datasets.

Experiment setting

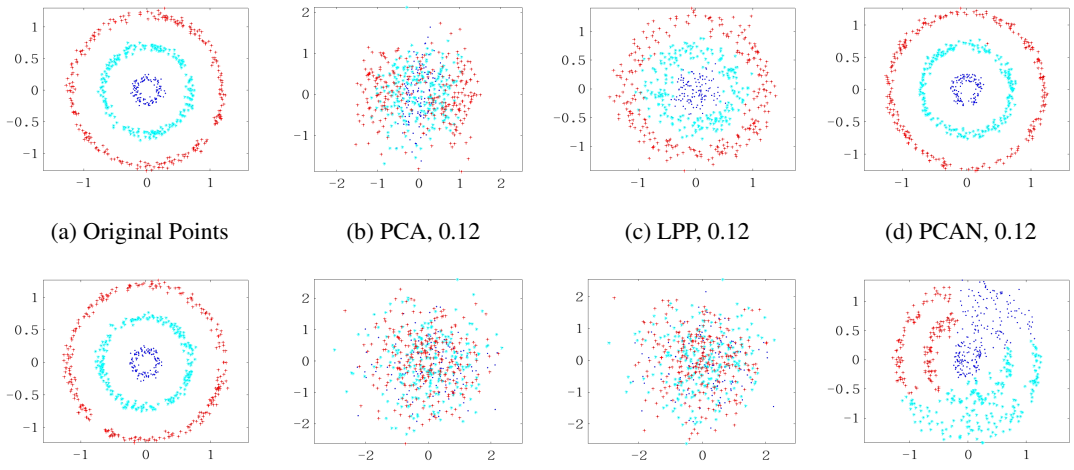
During the experiments, for the methods that need to perform K-means (e.g., Principal Component Analysis (PCA) (Jolliffe 2002), Locality Preserving Projections (LPP) (He and Niyogi 2004), K-means (Hartigan and Wong 1979), Ratio Cut (RCut) (Wang and Siskind 2003), Normalized Cut (NCut) (Shi and Malik 2000)), we run them 100 times with random initializations, and report their average performance (*Ave*), standard deviation (*std*) and optimal clustering result (*Best (min_obj)*). For the methods (e.g., LPP, RCut, NCut, Nonnegative Matrix Factorization (NMF) (Xu, Liu, and Gong 2003), Constrained Laplacian Rank (CLR) (Nie et al. 2016)) that take the affinity graph as input, we use the Self-tuning Gaussian method (Zelnik-Manor and Perona 2004) to construct the affinity graph. And the corresponding number of neighbor nodes is set to ten. In our method, K in parameter α is set to fifteen.

Experiments on synthetic dataset

In order to verify the performance of our method RS^{3}CIS in reducing the dimension on datasets containing noise and achieving clustering result by learning the intrinsic affinity matrix, some experiments are performed on synthetic data sets.

Construction of three-ring dataset We design a three-ring dataset which consists of sine functions, and their number of data points is 120, 220, 260. This dataset is 600×5. It includes two-dimension useful information and three-dimension noise data. We set two noise values: $noise1$ controls the dispersion within the cluster of two-dimensional useful information, and $noise2$ controls the size of the noise value in the other three dimensions. In this way, the two-dimension, three-cluster dataset in a circular arrangement is sequentially produced as shown in Fig.1(a). We use different colors and dot shapes to represent different data clusters. Our goal is to extract useful data dimensions in the five dimensions noise-containing dataset by dimension reduction and restore the useful information in the dataset as much as possible.

Methods for comparison Our method RS^{3}CIS is compared with the two most commonly used dimension reduction methods: PCA and LPP. In addition, our method RS^{3}CIS is also compared with another method PCAN. It should be noted, in order to facilitate the comparison of dimension reduction capabilities on the synthetic data, we set the reduced dimension to 2 for all compared method. For PCA and LPP, we only reduce the dimension of the synthetic dataset and mark the clusters according to their real labels.



(a) Original Points (b) PCA, 0.12 (c) LPP, 0.12 (d) PCAN, 0.12

(e) RS^{3}CIS, 0.12, 0.18 (f) PCA, 0.18 (g) LPP, 0.18 (h) PCAN, 0.18

Figure 1: Comparison of learnt subspaces on three-ring synthetic datasets with different $noise2$.

Table 1: Projection clustering results on a three-ring dataset with $noise2 = 0.18$

Methods	Best	Ave
PCA	0.3883	0.3733 ± 0.0093
LPP	0.3650	0.3582 ± 0.0076
PCAN		0.5267
RS^{3}CIS		1.0

Table 2: Statistics of benchmark datasets

Datasets	# of Samples	Features	Classes
USPS	1854	256	10
Palm	2000	256	100
Ecoli	336	7	8
Coil	1440	1024	20
Yeast	1484	8	10
Wine	178	13	3

Result analysis As shown in Figs.1(d) and 1(e), when $noise2$ is small, both PCAN and RS^{3}CIS can learn a good subspace and complete the clustering task correctly. Then, as we continue to increase the value of $noise2$, as shown in Figs.1(f), 1(g) and 1(h), both PCA and LPP fail completely and the PCAN method is gradually becoming powerless. Fortunately, as shown in Fig.1(e), our method RS^{3}CIS still maintains good dimension reduction and robust clustering capabilities when other methods are no longer valid. It is worth pointing out that the learn subspaces with our method (shown in Fig.1(e)) are consistent with the first 2-dimensional useful data (shown in Fig.1(a)), which also prove the interpretability of our method. This is mainly because the introduction of the row-sparse transformation matrix W allows our method to simultaneously perform feature selection and subspace learning, which theoretically ensures the robustness and interpretability of our method. The resulting During the experiment, we also notice that when $noise2$ is small, LPP is better than PCA in learning subspace (shown in Figs.1(b) and 1(c)). This is mainly because LPP pays more attention to the local structure of data, while PCA pays more attention to the global structure of data. Therefore, when $noise2$ is small, LPP has an advantage over PCA in learning subspace.

In addition, we also perform projection experiments and report the clustering ACC when $noise2 = 0.18$. For RS^{3}CIS and PCAN, all parameters are the same setting, and the two methods are run only once. The results of the above experiments are reported in Table 1. The above experimental results show that the proposed method RS^{3}CIS can still complete the corresponding subspace learning and clustering tasks correctly in a noisy environment, while other compared methods are severely ineffective.

Experiments on real benchmark dataset

In this section, we further verify the validity of our proposed method RS^{3}CIS through a clustering experiment and a projection experiment.

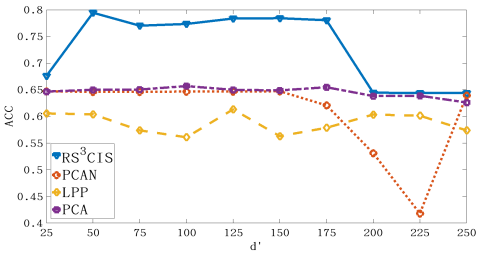
Datasets Among them, four image datasets: USPS[1], Palm[2], Ecoli (Athitsos and Sclaroff 2005) and Coil (Nene et al. 1996). Two biological datasets: Yeast (Asuncion and Newman 2007) comes from the UCI Machine Learning Repository, Wine is downloaded from (Zhong and Fukushima 2007). The details of these datasets are summarized in Table 2.

Clustering experiments We compare our method RS^{3}CIS with four classical clustering methods (K-means, NCut, RCut and NMF), a global representation methods CLR and a local projection representation method PCAN.

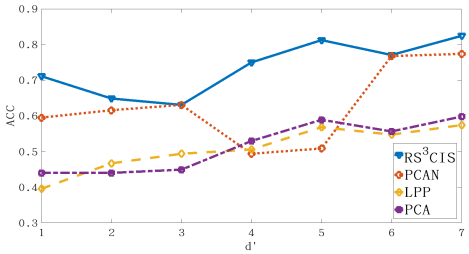
[1]http://www-i6.informatik.rwth-aachen.de/keysers/usps.html
[2]http://www.escience.cn/people/fpnie/index.html

Table 3: Experimental results on benchmark datasets

Methods		USPS	Palm	Ecoli	Coil	Yeast	Wine
K-means	Ave	0.6292±0.0357	0.6899±0.0245	0.5367±0.0594	0.5957±0.0558	0.3508±0.0207	0.6725±0.0547
	Best	0.6499	0.7390	0.5982	0.6625	0.3659	0.7022
RCut	Ave	0.6578±0.0818	0.5432±0.0241	0.5149±0.0469	0.6265±0.0710	0.3619±0.0185	0.7156±0.0313
	Best	0.6812	0.6115	0.5774	0.7258	0.4036	0.7170
NCut	Ave	0.6588±0.0844	0.4293±0.0287	0.5092±0.0431	0.5775±0.0787	0.3543±0.0206	0.6815±0.0811
	Best	0.6818	0.5095	0.5744	0.7007	0.3956	0.7037
NMF		0.6618	0.6265	0.5446	0.7674	0.3558	0.7126
CLR		0.7044	0.8040	0.5208	0.7688	0.4340	**0.7247**
PCAN		0.6494	0.8935	0.7738	0.7944	0.3888	0.7135
RS^{3}CIS		**0.7843**	**0.9045**	**0.8244**	**0.8058**	**0.4420**	**0.7247**



(a) USPS



(b) Ecoli

Figure 2: Projection clustering ACC comparison with different methods.

For a fair comparison, we set the parameters of our method RS^{3}CIS to the same values of PCAN. In addition, the reduced dimension d' is searched equally within the dimensions of the original dataset and the optimal clustering results are reported. The γ in our method RS^{3}CIS is taken from {1e-6,1e-3,1,1e3,1e6}. For the CLR method, we adjust all parameters to the optimal.

We report the experimental results on the six benchmark datasets in Table 3. From the experimental results, we can see that our method RS^{3}CIS is optimal (at least equal) on each dataset compared with other methods. This is mainly because our method RS^{3}CIS can learn an intrinsic subspace from the original dataset through the row-sparse transformation matrix W. The addition of a row-sparse penalty term $\gamma\|W\|_{2,1}$ makes our method more robust (proven in projection experiments). Therefore, the noise points and outliers contained in the dataset can be suppressed effectively on the real benchmark dataset, and better clustering results can be achieved. In addition, we also found that when the γ takes a small value, the clustering performance of our method is similar to that of PCAN. This is because when γ is infinitely small, our method RS^{3}CIS degenerates to PCAN.

Projection experiments In fact, we don't have any prior knowledge of the true subspace dimensions of the original dataset. Therefore, in order to further test the projection effect of our method RS^{3}CIS, we design the following projection experiments: we reduce the dataset to a different value of reduced dimension d' and then compare the clustering accuracy with different methods in the same reduced dimension.

For PCAN, we adjust all the parameters to the optimum. For PCA and LPP, we first reduce the dimension of the data set, then run K-means 100 times to achieve the best clustering ACC directly. Accordingly, we adjust the γ of the parameters in our method to the optimal. Due to space constraints, we only show the projected clustering ACC results on the datasets USPS and Ecoli in Fig.2.

From the experimental results shown in Fig.2, we can see that our method RS^{3}CIS always gets the best clustering ACC (at least equal) compared with other methods. This because that our method RS^{3}CIS can effectively learn a pure intrinsic subspace that is more conducive to clustering than other dimension reduction methods.

Conclusion

In this paper, a new Robust Single-Step Spectral Clustering with Intrinsic Subspace method is put forward. It can learn an intrinsic subspace and conduct a pure affinity matrix from the original dataset through a row-sparse transformation matrix. And the proposed problem in our method is effectively solved by a new iterative update algorithm. In addition, the introduction of the Laplacian matrix rank constraint allows our method to obtain an affinity graph with an explicit cluster structure, so the final clustering results can be obtained in a single step without any post-processing steps. In the future, we will extend our method to multi-view clustering problems and aim to achieve robust clustering results in multi-view clustering tasks with noise.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61602379 and 61772420),

International cooperation and exchange program of Shaanxi Province (No. 2016KW-034) and the ShaanXi Science and Technology Innovation Team Support Project under grant agreement (No. 2018TD-O26). The authors are grateful for the constructive advice on the revision of the manuscript from the anonymous reviewers.

References

Asuncion, A., and Newman, D. 2007. Uci machine learning repository.

Athitsos, V., and Sclaroff, S. 2005. Boosting nearest neighbor classifiers for multiclass recognition. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 45–45.

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):652–655.

Fan, R. K. C. 1997. *Spectral graph theory*. Conference Board of the mathematical sciences.

Gu, Q.; Li, Z.; and Han, J. 2011. Joint feature selection and subspace learning. In *International Joint Conference on Artificial Intelligence*, 1294–1299.

Hartigan, J. A., and Wong, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1):100–108.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in neural information processing systems*, 153–160.

Huang, J.; Nie, F.; and Huang, H. 2013. Spectral rotation versus k-means in spectral clustering. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 431–437.

Jolliffe, I. T. 2002. Principal component analysis. *Journal of Marketing Research* 87(100):513.

Li, Z.; Nie, F.; Chang, X.; Nie, L.; Zhang, H.; and Yang, Y. 2018. Rank-constrained spectral clustering with flexible embedding. *IEEE Transactions on Neural Networks & Learning Systems* PP(99):1–10.

Liao, S.; Gao, Q.; Yang, Z.; Cheng, F.; Nie, F.; and Han, J. 2018. Discriminant analysis via joint euler transform and $\ell_{2,1}$-norm. *IEEE Transactions on Image Processing*.

Mohar, B. 1991. The laplacian spectrum of graphs. *graph theory, combinations and applications* 18(7):871–898.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20). *Columbia University*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *International Conference on Neural Information Processing Systems*, 1813–1821.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *Thirtieth AAAI Conference on Artificial Intelligence*, 1969–1976.

Nie, F.; Wang, X.; Deng, C.; and Huang, H. 2017. Learning a structured optimal bipartite graph for co-clustering. In *Advances in Neural Information Processing Systems*, 4129–4138.

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, 2408–2414.

Nie, F.; Li, J.; and Li, X. 2016. Parameter-free auto-weighted multiple graph learning: a framework for multi-view clustering and semi-supervised classification. In *International Joint Conference on Artificial Intelligence*, 1881–1887.

Nie, F.; Li, J.; and Li, X. 2017. Self-weighted multi-view clustering with multiple graphs. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2564–2570.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 977–986.

Ren, P.; Xiao, Y.; Xu, P.; Guo, J.; Chen, X.; Wang, X.; and Fang, D. 2018. Robust auto-weighted multi-view clustering. In *IJCAI*, 2644–2650.

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. *Signal Processing Magazine IEEE* 28(2):52–68.

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

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. *Proc Acm Sigir* 267–273.

Yu, K. 2009. Nonlinear learning using local coordinate coding. *Proc Nips* 22:2223–2231.

Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, 1601–1608.

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.; 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.; He, W.; Li, Y.; Yang, Y.; Zhang, S.; Hu, R.; and Zhu, Y. 2017a. One-step spectral clustering via dynamically learning affinity matrix and subspace. In *AAAI*, 2963–2969.

Zhu, X.; Li, X.; Zhang, S.; Ju, C.; and Wu, X. 2017b. Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE Transactions on Neural Networks & Learning Systems* 28(6):1263–1275.