

Nonnegative self-representation with a fixed rank constraint for subspace clustering

Guo Zhong¹, Chi-Man Pun^{1,*}

Department of Computer and Information Science, University of Macau, Macau SAR, China



ARTICLE INFO

Article history:

Received 29 May 2019

Revised 10 January 2020

Accepted 10 January 2020

Available online 11 January 2020

Keywords:

Graph clustering

Subspace clustering

Least squares regression

Nonnegative representation

ABSTRACT

A number of approaches to graph-based subspace clustering, which assumes that the clustered data points were drawn from an unknown union of multiple subspaces, have been proposed in recent years. Despite their successes in computer vision and data mining, most neglect to simultaneously consider global and local information, which may improve clustering performance. On the other hand, the number of connected components reflected by the learned affinity matrix is commonly inconsistent with the true number of clusters. To this end, we propose an adaptive affinity matrix learning method, nonnegative self-representation with a fixed rank constraint (NSFRC), in which the nonnegative self-representation and an adaptive distance regularization jointly uncover the intrinsic structure of data. In particular, a fixed rank constraint as a prior is imposed on the Laplacian matrix associated with the data representation coefficients to urge the true number of clusters to exactly equal the number of connected components in the learned affinity matrix. Also, we derive an efficient iterative algorithm based on an augmented Lagrangian multiplier to optimize NSFRC. Extensive experiments conducted on real-world benchmark datasets demonstrate the superior performance of the proposed method over some state-of-the-art approaches.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Clustering is the organization of unlabeled data into similarity groups. A large amount of data in the fields of computer vision, machine learning, and data mining is high-dimensional. This results in the “curse of dimensionality,” by which the complete enumeration of all subspaces becomes difficult with an increasing number of dimensions [1]. How to effectively group these data has become an urgent task. Recent research suggests that high-dimensional data usually distribute in several low-dimensional subspaces with low-dimensional structures [2]. Subspace clustering simultaneously clusters the data into their underlying subspaces and finds a low-dimensional subspace fitting each group of points [3]. Therefore, subspace clustering approaches have become powerful tools for clustering high-dimensional data. We can roughly divide subspace clustering methods into the four categories of iterative-based, statistical, algebraic geometry-based, and spectral clustering-based methods. Vidal [3] presented a detailed review of the above methods. The method proposed in this paper is spectral clustering-based.

* Corresponding author.

E-mail address: cmpun@um.edu.mo (C.-M. Pun).

¹ Guo Zhong and Chi-Man Pun are equally contributed to the manuscript.

Spectral clustering is a class of graph-based techniques that unravel the structural information of a graph using information conveyed by the spectral decomposition of an associated matrix [4]. Compared to the traditional K -means algorithm, it can achieve good clustering results even in with nonlinearly distributed data, and its implementation is not complicated. Spectral clustering can usually be implemented in two steps [5]: constructing an affinity matrix between data points and performing a low-dimension embedding of the affinity matrix. The affinity matrix can essentially be considered a graph. Thus, the grouping is accomplished by determining the optimal graph partitioning. Hence, clustering performance is sensitive to the quality of the learned graph [5,6]. It is difficult in practice to obtain the optimal graph for clustering, since the affinity matrix is usually computed via a predefined neighbor number to construct the graph, i.e., to empirically determine the neighborhood size of each vertex to learn a nearest neighbor graph. Moreover, the nearest measure based on Euclidean distance is sensitive to noise and outliers. To avoid those drawbacks of spectral clustering, sparse subspace clustering (SSC) [7] was proposed through utilizing the self-expressiveness property of the data, which inherits the merit of sparse representation (SR) [8] such that each data point can adaptively determine its neighbors and its similarity with other data points. Specifically, SSC uses the l_1 -norm to minimize the coefficient matrix Z of the linear representation of data, which encodes the subspace membership information from data. Moreover, adding an error term to the framework of SSC can make it robust to noise and outliers. As a result, a better affinity matrix for clustering can be constructed for Z as $\frac{|Z|+|Z^T|}{2}$, which reveals the subspace characteristics of data. The final clustering results can be attained by feeding the learned affinity matrix into the spectral grouping framework [7,9]. Elhamifar and Vidal [7] showed good performance of SSC in motion segmentation and face clustering. Along this line, Pham et al. [10] considered the spatial relationship between data points and proposed weighted sparse subspace clustering (WSSC) by embedding weights into the coefficient matrix Z . Desiring high efficiency and better clustering performance, Chen et al. [11] proposed sparse subspace clustering based on active orthogonal matching pursuit (AOMPSSC). However, SSC, WSSC, and AOMPSSC all suffer from the same handicap, i.e., the neglect of the global structure of data, since the solution to the problem of l_1 -norm minimization only brings data points close to the represented data point [7,12].

To alleviate the problem of losing global information of data in SCC, Liu et al. [9] introduced the idea of low rank to the subspace clustering problem and proposed a low-rank representation (LRR) algorithm. Assuming that data samples are roughly drawn from a union of multiple low-rank subspaces, LRR finds the lowest-rank representation of data using a given dictionary or the data itself. To optimize LRR is to solve a convex optimization problem of the minimum nuclear norm [13], which is considered an alternative to rank minimization. Compared to SCC, LRR better captures global information around each sample and shows promising application prospects. However, LRR ignores the local structure of data. The latest research [14] suggests that the learning performance of the algorithms can be significantly enhanced if the local and global structure of data are considered simultaneously. Therefore, to only depend on the coefficients of data representation to capture the local or global structure of data is not enough, as LRR and SSC cannot totally discover the intrinsic structure of the data. To enhance the performance of LRR, a manifold regularization is incorporated to characterize the local manifold structure of data [15,16], based on the assumption that close points in the original data space tend to have similar embedding [17]. However, the affinity matrix of manifold regularization is precomputed, which may not reflect the true similarities between data points, i.e., two data points lying in different subspaces could be very close to each other [18]. There is another dimension to the graph-based clustering problem. In an ideal case, the learned graph for clustering should have exactly c connected components [19], where c is the number of clusters. However, the above methods cannot obtain such a graph. As a result, the graphs obtained by these methods are not the optimal graphs for clustering.

Nonnegative representation (NR) [20] only allows nonnegative combinations of multiple data points to reconstruct a represented data point, which is compatible with the intuitive combination of parts into a whole [21]. Xu et al. [20] showed that to constrain the coding coefficients to be nonnegative can automatically boost the representational power of homogeneous data points while limiting the representation power of heterogeneous data points. Hence, NR often leads to better performance over SR for data representation [22]. However, NR is specifically designed for classification tasks and not for clustering. In this study, we address the limitations of current graph learning methods by integrating an adaptive distance regularization in the framework of the nonnegative constrained least squares model [23] to simultaneously capture the local and global structures of data. We use a fixed-rank constraint on the Laplacian matrix to inspire the affinity matrix to have the exact number of connected components required by the clustering task. Compared to state-of-the-art adaptive graph learning methods [19,24], the proposed nonnegative self-representation with a fixed-rank constraint (NSFRC) adjusts the graph based on least-squares regression to improve the robustness and clustering performance. Compared to recent improvements on LRR [15,16,25], which directly construct the Laplacian matrix for manifold regularization in advance, NSFRC simultaneously exploits the local distance information and global representation of data to learn an affinity matrix that captures the data's intrinsic structure. Extensive experiments have been conducted on nine real-world benchmark datasets. We show how parameters affect clustering performance in our method. Specifically, empirical results show that the proposed method is data-driven and consistently outperforms the compared clustering methods.

The main contributions of this paper are as follows:

- A novel subspace clustering approach is proposed, which simultaneously performs the global representations of data and local structure learning. Specifically, it adaptively learns the local manifold structure, and thus can greatly improve the quality of the affinity matrix.

Table 1
Notations.

Symbol	Meaning
c	number of clusters
n	number of samples
d	number of features
I	identity matrix
$M \geq 0$	nonnegative matrix
m_j	j th column of a matrix M
m^i	i th row of M
m_{ij}	i th row and j th column element of M
$\ m\ _2$	l_2 -norm of vector m
$\ M\ _F$	Frobenius norm of M , i.e., $\ M\ _F = \sqrt{\sum_{i,j} m_{ij}^2}$
$\ M\ _\cdot$	nuclear norm of M defined by the sum of its singular values
$\ M\ _1$	l_1 -norm of M , i.e., $\ M\ _1 = \sum_{i,j} m_{ij} $
$\ M\ _{2,1}$	$l_{2,1}$ -norm of M , i.e., $\ M\ _{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^d m_{ij}^2}$
M^T	transpose of M
$Tr(M)$	trace of M

- A reasonable rank constraint is introduced to NSFRC. The learned graph can more accurately reflect the true structure of data clusters.
- We derive a new augmented Lagrange multiplier-based algorithm to efficiently and effectively solve the associated optimization problem.
- Extensive experiments conducted on nine benchmark datasets demonstrate the effectiveness of the proposed approach. NSFRC consistently outperforms state-of-the-art one-step clustering methods in terms of accuracy.

The remainder of this paper is organized as follows. In [Section 2](#), related work in subspace clustering is briefly described. [Section 3](#) introduces the objective function of NSFRC. We present the augmented Lagrangian multiplier-based procedure to solve the proposed objective function in [Section 4](#). We perform experiments on nine benchmark datasets to validate NSFRC in [Section 5](#), and we present our conclusions in [Section 6](#).

2. Related work

Because this study is related to the category of spectral-clustering-based methods, we revisit some concepts of spectral clustering and then briefly describe current spectral-clustering-based subspace clustering methods. For convenience, we denote matrices by uppercase letters, and in particular, the input data matrix is X . We summarize the symbols used in the paper in [Table 1](#).

2.1. Spectral clustering

Given a data matrix $X = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}$ stacked by data with c clusters, whose columns are n data points drawn from a d -dimensional feature space, the main task of clustering is to partition the data into c clusters. A nice way to represent the relationships between data is in the form of a similarity graph $G = (V, E)$, where G is a set of vertices and E is a set of edges. Each vertex v_i in the graph G represents a data point x_i . Two vertices are connected if the similarity w_{ij} between the corresponding data points x_i and x_j is positive or larger than a certain threshold, and the edge is weighted by w_{ij} [5]. The problem of clustering can now be reformulated using the similarity graph to find a partition of the graph such that the edges between different groups have very low weights and the edges within a group have high weights. To this end, we need the so-called graph Laplacian (i.e., Laplacian matrix) [5],

$$L = D - W,$$

where the degree matrix D is a diagonal matrix with diagonal elements $d_{ii} = \sum_j w_{ij}$, and w_{ij} is the i th row and j th column entry of the affinity matrix W . The objective function of the famous normalized cuts (Ncuts) [6] spectral clustering method can be formulated as

$$\min_{Q^T Q = I} Tr(Q^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} Q), \quad (1)$$

where the optimal solution $Q \in \mathbb{R}^{n \times c}$ of [Eq. \(1\)](#) can be obtained by eigenvalue decomposition of the matrix $D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ [6].

Denote the cluster assignment matrix by $Y = [y^1, \dots, y^n]^T \in \mathbb{R}^{n \times c}$, where $y^i \in \mathbb{R}^{1 \times c}$ ($1 \leq i \leq n$) is the cluster assignment vector for the data point x_i . The j th element of y^i is 1 if the data point x_i is assigned to the j th cluster, it is 0 otherwise. Specifically, there is one and only one element equal to 1 in y^i . To obtain the final clustering result Y , we can input Q to the K -means clustering algorithm to output the discrete-valued cluster assignment for each data point [26]. Therefore, spectral clustering methods usually consist of two sequential steps, i.e., learning an affinity matrix from the data, and performing

Algorithm 1 Spectral clustering according to Shi and Malik [6].**Input:** Data $X \in \mathbb{R}^{d \times n}$, number c of clusters.**Output:** Cluster assignment matrix Y satisfying $y_{ij} = 1$ if $q^i \in C_i$, and $y_{ij} = 0$ otherwise.

- 1: Learn an affinity matrix W .
- 2: Use the learned affinity matrix W to compute the Laplacian matrix $L = D - W$.
- 3: Compute the first k eigenvectors $u_1, \dots, u_k$ of $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$.
- 4: Let $U \in \mathbb{R}^{n \times c}$ be the matrix containing the vectors $u_1, \dots, u_c$ as columns.
- 5: For $i = 1, \dots, n$, let $q^i \in \mathbb{R}^c$ be the vector corresponding to the i -th row of U .
- 6: Cluster the points $q^1, \dots, q^n$ in $\mathbb{R}^c$ with the K -means algorithm into clusters $C_1, \dots, C_c$.

spectral clustering on the resulting affinity matrix. The clustering method is outlined in Algorithm 1. For more details on spectral clustering, see an elaborate review [5].

2.2. Sparse subspace clustering

From the above, we can see that the most critical step of spectral clustering is to learn a high-quality similarity graph, which captures the intrinsic structure of data. Learning a similarity graph with different methods will yield different spectral clustering methods. SSC [7] is a class of spectral clustering methods based on SR. Benefiting from SR and the self-expressiveness property of data [7], SSC can simultaneously determine a few neighbors for each data point and the similarity between data instead of constructing a nearest neighborhood graph based on Euclidean distance. Let $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$ be a data matrix consisting of n data points, which are drawn from a union of multiple subspaces. By the self-expressiveness property of data, X can be represented by

$$X = XZ,$$

or

$$x_i = \sum_{j \neq i} z_{ji} x_j, \quad 1 \leq i \leq n,$$

where $Z \in \mathbb{R}^{n \times n}$ is the representation matrix whose each diagonal element is 0. Since each data point can be represented as a linear combination of all other data points, Elhamifar and Vidal [7] provided a new way to construct the similarity graph, i.e., the affinity matrix can be constructed as $W = \frac{|Z| + |Z^T|}{2}$. In principle, solving for Z is an ill-posed problem, since there are many possible solutions, and the data X are usually noisy. To resolve these issues, the objective function of SSC is described as

$$\min_Z \|Z\|_1 + \lambda \|E\|_F, \quad \text{s.t. } X = XZ + E, \quad \text{diag}(Z) = 0, \quad (2)$$

where $\lambda \geq 0$ is a parameter to balance the two parts of the objective function in Eq. (2), E is the observation noise, and $Z \in \mathbb{R}^{n \times n}$ is the self-expression coefficient matrix. $\|E\|_F$ is the Frobenius norm constraint on E , calculated as $\|E\|_F = \sqrt{\sum_{i,j} e_{ij}^2}$, where e_{ij} is the i -th row and j -th column entry of E . $\|Z\|_1$ is the l_1 -norm constraint on Z , calculated as $\|Z\|_1 = \sum_{i,j} |z_{ij}|$, where z_{ij} is the i -th row and j -th column entry of Z . By optimizing the objective function in Eq. (2), we can obtain that each column of Z is sparse [8]. More importantly, due to the competitiveness in joint representation, data points from the same subspace as the data point to be represented are more likely to be chosen for representation [27]. This means that the nonzero entry z_{ji} indicates that x_i can more possibly share the same class label as x_j . Therefore, SSC can capture the local structure of data, since those data points with nonzero representation value are often close to the represented data point, while losing global information in data.

2.3. Low-Rank representation and its variant

LRR [9,18] aims to jointly seek the lowest-rank representation of all data so that each data point can be represented as a linear combination of some dictionary or the data itself. The general formulation of LRR can be written as

$$\min_Z \text{rank}(Z) + \lambda \|X - XZ\|_F, \quad (3)$$

where $\text{rank}(Z)$ is the rank of matrix $Z \in \mathbb{R}^{n \times n}$. LRR finds a representation of the lowest rank instead of a sparse one like SSC, and it has been proved effective in preserving the global data structure.² The objective function in Eq. (3) is NP-hard and difficult to optimize [9].

² A space, which is measured by the l_2 -norm, is clearly a Euclidean space. The structure exploited by minimizing the approximate error of all data in the Frobenius norm (or equally in the l_2 -norm) is called the global Euclidean structure [15].

To efficiently optimize the objective function in Eq. (3), Liu et al. [9] reformulated the above problem of rank minimization as

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_{2,1}, \quad \text{s.t. } X = XZ + E,$$

where $\|Z\|_*$ is the nuclear norm constraint on the coefficient matrix Z . The nuclear norm [13] is a strong surrogate of the rank of a matrix, and $\|Z\|_*$ is defined as the sum of the singular values of the matrix Z . E represents the sample-specific corruptions and can be calculated as $\|E\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^d e_{ij}^2}$. The generated coefficient matrix Z of LRR may be too dense to capture the local structure of data [28].

A vast number of geometric surfaces are described as manifolds. A sphere is one of the simplest examples of a manifold in three-dimensional Euclidean space. Recent literature [29,30] shows that data lying in a high-dimensional ambient space commonly are found on a lower-dimensional manifold. To capture the manifold structure of data, in manifold learning theory [31], we naturally assume that if two data points are close in the original space, then the new representations of these two points in a new space must be close to each other [1]. Motivated by this basic idea, some researchers [15,16,25] incorporated a manifold regularization characterized by a graph Laplacian in LRR to preserve the manifold structure of data, as follows:

$$\min_{Z,E} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \text{Tr}(Z^T L Z), \quad \text{s.t. } X = XZ + E,$$

where $L = D - W$ is a precomputed Laplacian matrix and $\text{Tr}(Z^T L Z)$ is the manifold regularization. Note that $\min_Z \text{Tr}(Z^T L Z)$ can be expanded as

$$\min_Z \frac{1}{2} \sum_{i,j} \|z_i - z_j\|_2^2 w_{ij},$$

where w_{ij} is the i th row and j th column element of W . When $w_{ij} \rightarrow \infty$, $\|z_i - z_j\|_2^2 \rightarrow 0$. By incorporating manifold regularization in LRR, the learned Z captures the local structure of data, and the subspace clustering will benefit. The above method is called low-rank representation with manifold regularization (LRRMR).

Note the separation in LRRMR between constructing a Laplacian matrix from the original data and deriving a new representation of data. This two-step strategy separately achieves the optimal solution at each step, but may not obtain the globally optimal graph for clustering. Moreover, the precomputed W based on Euclidean distance may not reflect the true similarity between data points, i.e., two points x_i and x_j lying in different subspaces could be very close to each other, but they should not have a large w_{ij} [18].

3. Methodology

In this section, we formulate the objective function of NSFRC by several steps.

The clustering task is commonly expressed as partitioning a set of objects with c clusters into c groups according to the similarity among objects. The affinity matrix constructed by the similarity between adjacent data plays a key role in the process of clustering. From the local representation assumption that every data point should be connected with its nearest neighbors, close data points should have large similarity while distant data points have small or even zero similarity [32]. On the other hand, Lu et al. [33] showed that the Frobenius-norm constraint on the coefficient matrix can force highly-correlated data points to have consistent coefficients. Specifically, it is effective for discovering the global structures of the data [34]. Therefore, we integrate a distance regularization term in the framework of the least squares model [23], as follows:

$$\begin{aligned} \min_Z \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_1}{2} \|Z\|_F^2, \\ \text{s.t. } X = XZ, Z \geq 0, \text{diag}(Z) = \mathbf{0}, \end{aligned} \quad (4)$$

where $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$ is a data matrix, in which d and n are respectively the feature dimensionality of each sample and the number of data samples. $\lambda_1 > 0$ is a penalty parameter. Each data point x_i is expressed as a linear combination of all other data points, i.e., $x_i = \sum_{j \neq i} z_{ji} x_j$. We constrain each $z_{ij} \geq 0$ to directly reflect the similarity between x_i and x_j , and we avoid any two non-nearest neighbor data points being connected by a larger negative value. Naturally, z_{ji} should be larger if x_i is closer to x_j . Furthermore, $Z \in \mathbb{R}^{n \times n}$ can be considered an affinity matrix, in which z_{ij} measures the similarity between data points x_i and x_j . The global structure of the data is captured by the Frobenius-norm regularization term $\|Z\|_F^2$, which forces highly correlated samples to have consistent coefficients [34]. The local structure of data is exploited through the adaptive distance regularization term $\sum_{i,j} \|x_i - x_j\|_2^2 z_{ij}$ instead of precomputing a locality constraint. As a result, the objective function in Eq. (4) considers both global and local information of data. It follows that the learned Z should capture the intrinsic structure of data.

Data from real-world applications are usually contaminated by noise. We extend the above formulation to be robust to noise by introducing an error term E , as follows:

$$\begin{aligned} \min_{Z, E} \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_1}{2} \|Z\|_F^2 + \frac{\lambda_2}{2} \|E\|_F^2, \\ \text{s.t. } X = XZ + E, Z \geq 0, \text{diag}(Z) = \mathbf{0}, \end{aligned} \quad (5)$$

where $\lambda_2 > 0$ is a penalty parameter.

Ideally, the learned affinity matrix should contain exactly c connected components instead of learning a fully connected similarity graph [35]. Previous work usually neglects the number c of clusters in the process of clustering. From Proposition 2 [5], if the rank of the Laplacian matrix is $n - c$, i.e., $\text{rank}(L_Z) = n - c$, where L_Z is the Laplacian matrix with respect to the coefficient matrix Z , then the learned graph contains exactly c connected components [36]. To further improve the clustering performance, we integrate prior knowledge, i.e., the number of clusters of data, by adding the rank constraint on the Laplacian matrix to the objective function in Eq. (5), as follows:

$$\begin{aligned} \min_{Z, E} \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_1}{2} \|Z\|_F^2 + \frac{\lambda_2}{2} \|E\|_F^2, \\ \text{s.t. } X = XZ + E, Z \geq 0, \text{diag}(Z) = \mathbf{0}, \\ \text{rank}(L_Z) = n - c \end{aligned} \quad (6)$$

where

$$L_Z = D - \frac{(Z + Z^T)}{2}$$

is the Laplacian matrix, and D is the degree matrix, whose i th diagonal element is $\frac{\sum_j (z_{ij} + z_{ji})}{2}$.

However, it is difficult to solve the above optimization problem, since the constraint $\text{rank}(L_Z) = n - c$ is not easy to handle. Note that each eigenvalue of L_Z is nonnegative, since L_Z is positive semidefinite [5]. Consequently, $\text{rank}(L_Z) = n - c$ is equivalent to $\sum_{i=1}^c \sigma_i(L_Z) = 0$, where $\sigma_i(L_Z)$ denotes the i th smallest eigenvalue of L_Z . From Fan's theorem [37], we have

$$\sum_{i=1}^c \sigma_i(L_Z) = \min_{F \in \mathbb{R}^{n \times c}, F^T F = I} \text{Tr}(F^T L_Z F),$$

where F is a new variable introduced by the theorem [37]. As we will see, this makes the objective function in Eq. (6) much easier to optimize. It is natural to assume that a smaller distance $\|x_i - x_j\|_2^2$ should reflect that x_i and x_j have a larger probability of being in the same cluster. Therefore, we add an additional constraint on Z , i.e., $z^i \mathbf{1} = 1$, where z^i denotes the i -row of Z . Finally, the objective function of NSFRC is

$$\begin{aligned} \min_{Z, E, F} \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_1}{2} \|Z\|_F^2 + \frac{\lambda_2}{2} \|E\|_F^2 + 2\lambda_3 \text{Tr}(F^T L_Z F), \\ \text{s.t. } X = XZ + E, Z \geq 0, \text{diag}(Z) = \mathbf{0} \\ z^i \mathbf{1} = 1, F \in \mathbb{R}^{n \times c}, F^T F = I \end{aligned} \quad (7)$$

where $\lambda_3 > 0$ is also a penalty parameter, and if $\lambda_3 > 0$ is set large enough, then the optimal solution will force $\text{Tr}(F^T L_Z F)$ to be close to 0. It follows that the constraint $\text{rank}(L_Z) = n - c$ will be met. In our experiments, we did not set λ_3 too large, since the rank constraint is too strong and may result in the learned Z not capturing the local and global structure of data.

Similar to SCC, NSFRC is a spectral clustering-based method. We use $\frac{z_{ij} + z_{ji}}{2}$ to measure the similarity between x_i and x_j . The affinity matrix W is computed as $W = \frac{Z + Z^T}{2}$. Our ultimate clustering results will be produced by feeding W into Algorithm 1.

4. Optimization algorithm for NSFRC

In our objective function in Eq. (7), three variables must be optimized, and it is not convex for Z , E , and F . However, we still can optimize the objective function in Eq. (7) by decomposing it into several convex subproblems [38]. Specifically, the three variables can be optimized alternatively. We introduce two auxiliary variables, $Z = S$ and $Z = U$, as follows:

$$\begin{aligned} \min_{Z, E, F} \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|E\|_F^2 + 2\lambda_3 \text{Tr}(F^T L_S F), \\ \text{s.t. } X = XZ + E, Z = U, Z = S, S \geq 0, \\ \text{diag}(S) = \mathbf{0}, s^i \mathbf{1} = 1, F \in \mathbb{R}^{n \times c}, F^T F = I. \end{aligned} \quad (8)$$

The augmented Lagrangian function of the objective function in Eq. (8) is

$$\begin{aligned} \mathcal{L}(Z, E, F, S, U) = & \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij} + \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|E\|_F^2 + 2\lambda_3 \text{Tr}(F^T L_S F) + \langle \Lambda_1, X - XZ - E \rangle \\ & + \langle \Lambda_2, Z - U \rangle + \langle \Lambda_3, Z - S \rangle + \frac{\mu}{2} (\|X - XZ - E\|_F^2 + \|Z - U\|_F^2 + \|Z - S\|_F^2), \end{aligned} \quad (9)$$

where Λ_1 , Λ_2 , and Λ_3 are Lagrangian multipliers, and μ is a regularity coefficient to control the penalty for the violation of three equality constraints in the objective function in Eq. (8). Now, we optimize the objective function in Eq. (9) with respect to one variable while fixing the other variables. As we will see, the objective function in Eq. (9) is reduced to five manageable subproblems, each with a closed-form solution.

Step 1.

We fix E , F , S , and U , and Z is updated by minimizing

$$\mathcal{L}(Z) = \|X - XZ - E\|_F^2 + \frac{\Lambda_1}{\mu} \|Z\|_F^2 + \|Z - U\|_F^2 + \frac{\Lambda_2}{\mu} \|Z\|_F^2 + \|Z - S\|_F^2 + \frac{\Lambda_3}{\mu} \|Z\|_F^2.$$

To compute Z , we take the derivative of $\mathcal{L}(Z)$ with respect to Z , set it to zero, and obtain

$$Z = (X^T X + 2I)^{-1} (X^T Q_1 + Q_2 + Q_3), \quad (10)$$

where $Q_1 = X - E + \frac{\Lambda_1}{\mu}$, $Q_2 = U - \frac{\Lambda_2}{\mu}$, and $Q_3 = S - \frac{\Lambda_3}{\mu}$.

Step 2.

When Z , E , F , and U are fixed, S can be obtained by solving the following problem:

$$\begin{aligned} \min_S \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij} + 2\lambda_3 \text{Tr}(F^T L_S F) + \frac{\mu}{2} \|Z - S\|_F^2 + \frac{\Lambda_3}{\mu} \|Z\|_F^2, \\ \text{s.t. } S \geq 0, \text{diag}(S) = \mathbf{0}, s^i \mathbf{1} = 1. \end{aligned} \quad (11)$$

Note that

$$\frac{1}{2} \sum_{i,j=1}^n \|f^i - f^j\|_2^2 s_{ij} = \text{Tr}(F^T L_S F),$$

and the objective function in Eq. (11) can be directly decoupled into rows, enabling us to tackle the following objective function individually for every i :

$$\begin{aligned} \min_{s^i} \sum_{j=1}^n (\|x_i - x_j\|_2^2 s_{ij} + \lambda_3 \|f^i - f^j\|_2^2 s_{ij} + \frac{\mu}{2} s_{ij}^2 - \mu s_{ij} h_{ij}), \\ \text{s.t. } s^i \geq 0, s_{ii} = 0, s^i \mathbf{1} = 1, \end{aligned} \quad (12)$$

where h_{ij} is the (i, j) th element of $H = Z + \frac{\Lambda_3}{\mu}$. Let

$$d_{ij} = \|x_i - x_j\|_2^2 + \lambda_3 \|f^i - f^j\|_2^2 - \mu h_{ij}$$

be the j th element of $d_i \in \mathbb{R}^{1 \times n}$. With some algebra, we can state Eq. (12) as

$$\min_{s^i \geq 0, s_{ii}=0, s^i \mathbf{1}=1} \|s^i + \frac{d_i}{\mu}\|_2^2. \quad (13)$$

The Lagrangian function of the objective function in Eq. (13) is

$$\mathcal{L}(s^i, \eta, \beta) = \|s^i + \frac{d_i}{\mu}\|_2^2 + \eta(1 - s^i \mathbf{1}) + \beta^T (-s^i),$$

where η and $\beta \geq \mathbf{0}$ are the Lagrangian multipliers. According to the KKT condition [39], it can be verified that the optimal solution s^i should be

$$s^i = \left(-\frac{d_i}{\mu} + \eta \mathbf{1}^T \right)_+, \quad (14)$$

where $(\cdot)_+ = \max(\cdot, 0)$.

Step 3.

With fixed Z , S , U , and E , we can obtain F by solving the following problem:

$$\min_{F \in \mathbb{R}^{n \times c}, F^T F = I} 2\lambda_3 \text{Tr}(F^T L_S F). \quad (15)$$

The optimal solution F consists of the c eigenvectors of L_S with respect to the c smallest eigenvalues.

Step 4.

We fix E , F , S , and Z , and update U by minimizing

$$\mathcal{L}(U) = \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\mu}{2} \|Z - U\|_F^2 + \frac{\Lambda_2}{\mu} \|U\|_F^2.$$

We take the derivative of $\mathcal{L}(U)$ with respect to U and set it to zero to obtain

$$U = \frac{\Lambda_2 + \mu Z}{\lambda_1 + \mu}. \quad (16)$$

Step 5.

Similar to step 4, we fix all variables except E to obtain

$$\mathcal{L}(E) = \frac{\lambda_2}{2} \|E\|_F^2 + \frac{\mu}{2} \|X - XZ - E\|_F^2 + \frac{\Lambda_1}{\mu} \|E\|_F^2.$$

By setting the derivative of $\mathcal{L}(E)$ with respect to E equal to zero, it is obvious that

$$E = \frac{\Lambda_1 + \mu(X - XZ)}{\lambda_2 + \mu}. \quad (17)$$

To characterize the random corruptions, the l_1 -norm constraint on E can be adopted, and E is also easy to compute by the method of Lin et al. [40].

Step 6.

From Boyd [38], we update Λ_1 , Λ_2 , Λ_3 , and μ as follows:

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu(X - XZ - E), \\ \Lambda_2 &= \Lambda_2 + \mu(Z - U), \\ \Lambda_3 &= \Lambda_3 + \mu(Z - S), \\ \mu &= \min(\rho\mu, \mu_{\max}), \end{aligned} \quad (18)$$

where ρ and $\mu_{\max}$ are prefixed constants, and ρ controls the convergence speed.

After Z is learned, we can perform clustering by feeding Z into the spectral clustering framework, such as Algorithm 1, to obtain the clustering results. The details of the optimization algorithm of NSFRC are presented in Algorithm 2. As we have

Algorithm 2 ADMM for NSFRC.

Input: Data $X \in \mathbb{R}^{d \times n}$, and regularization parameters λ_1 , λ_2 and λ_3 .

Output: Affinity matrix $W = \frac{Z+Z^T}{2}$.

- 1: initialization: $\mu > 0$, $\rho > 1$, $\Lambda_1 = \Lambda_2 = \Lambda_3 = \mathbf{0}$, $Z = S$, $U = Z$, $E = \mathbf{0}$ and Z is an affinity matrix based on k nearest neighbor graph.
- 2: **repeat**
- 3: Update Z using Eq. (10).
- 4: Update S line by line using Eq. (14).
- 5: Update F by computing the first c eigenvectors of L_S in Eq. (15).
- 6: Update U using Eq. (16).
- 7: Update E using Eq. (17).
- 8: Update Λ_1 , Λ_2 , Λ_3 and μ according to Eq. (18).
- 9: **until** convergence

seen, the alternating direction method with multipliers (ADMM) [38] is adopted to optimize NSFRC. For a theoretical analysis of convergence for the proposed algorithm, the reader is referred to Zhang et al. [41]. We will provide empirical evidence in the experimental section to show the convergence of the method. Note that the objective function in Eq. (7) is not convex, so there may be many local minima. Table 2 shows a computational complexity analysis for the proposed method. Its total computational complexity can be improved by using some numerical packages [42].

Table 2
Complexity analysis for the proposed method.

	Addition	Multiplication	Complexity
Z	$dn + n^2$	$dn^2 + n^{2.373}$	$O(dn^2)$
S	n^2	$dn + n^2$	$O(n^2)$
F	n^2	cn^2	$O(cn^2)$
U	n^2	n^2	$O(n^2)$
E	dn^2	dn^2	$O(dn^2)$

Table 3
Statistics of the benchmark datasets.

Data set	#dimensionality	#instances	#clusters
JAFFE	1024	213	10
Dermatology	34	366	6
COIL20	1024	1440	20
MSRA	256	1799	12
PalmData	256	2000	100
Extended YaleB	1024	2414	38
Glass	9	214	6
USPS	256	9298	10
COIL100	1024	7200	100

5. Connection between SSC and NSFRC

If we set λ_1 and λ_3 to zero, the objective function in Eq. (7) can be considered as the following weighted SSC problem:

$$\min_{Z, E} \sum_{i,j} \|x_i - x_j\|_2^2 z_{ij} + \frac{\lambda_2}{2} \|E\|_F^2$$

$$\text{s.t. } X = XZ + E, \text{diag}(Z) = \mathbf{0}, Z \geq \mathbf{0}, Z^T \mathbf{1} = \mathbf{1}.$$

In this way, NSFRC can help each data point to adaptively choose a few nearest neighbor data points to represent itself. Moreover, the learned Z is sparse, hence Z can capture the local structure of data. Therefore, NSFRC can also be regarded as an extension of SSC.

6. Experimental analysis

In this section, we demonstrate the effectiveness of the proposed method on nine benchmark datasets. Specifically, we compare NSFRC to state-of-the-art clustering algorithms.

6.1. Data sets

We conducted clustering experiments on nine benchmark datasets, including seven image datasets: JAFFE [43], COIL20 [44], MSRA [45], PalmData [46], COIL100 [44], USPS [47], and the Extended YaleB [48], where Extended YaleB, JAFFE, and MSRA are facial datasets, COIL20 and COIL100 are object image datasets, USPS is a handwritten digit dataset, and PalmData is a palmprint image dataset. The non-image datasets Dermatology and Glass are from the UCI machine learning repository [49]. The important statistics of these datasets are summarized in Table 3.

6.2. Evaluation metric

Following Wen et al. [27], we evaluated the clustering performance by two standard metrics: accuracy (ACC) and normalized mutual information (NMI). Assume we have N samples. Then, ACC can be defined as

$$\text{ACC} = \frac{\sum_{i=1}^N f(o_i, \text{map}(r_i))}{N},$$

where $f(\cdot, \cdot)$ is a function of two variables satisfying $f(x, y) = 1$ if $x = y$, and $f(x, y) = 0$ otherwise; $\text{map}(r_i)$ is the permutation mapping function that maps each ground-truth label r_i to the equivalent label from the dataset; and o_i and r_i are respectively the cluster label and ground-truth label.

Mutual information (MI) is generally used to measure the similarity of two sets of clusters. Assume that C is the ground-truth set of clusters for the original dataset, and that C' is the set of clusters generated by an algorithm. Then, the mutual information of C and C' is defined as

$$\text{MI}(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)},$$

where $p(c_i, c'_j)$ is the joint probability that this arbitrarily selected image belongs to both the cluster c_i and c'_j , and $p(c_i)$ and $p(c'_j)$ are the probabilities that a sample arbitrarily selected from the dataset belongs to the clusters c_i and c'_j , respectively. In the experiment, we use

$$\text{NMI} = \frac{\text{MI}(C, C')}{\max(H(C), H(C'))},$$

where $H(C)$ and $H(C')$ are the entropies of C and C' , respectively. The range of $\text{NMI}(C, C')$ is $[0, 1]$. The bigger the NMI is, the more similar C and C' are.

Table 4Clustering mean ACC (%) compared with different methods, $ACC = \frac{\sum_{i=1}^N f(o_i, map(r_i))}{N}$.

Dataset/Method	K-means	NCuts	SSC	LRR	LSR	LRRMR	CLR	LSS	NSFRC	Avg.
JAFFE	80.28	95.89	98.12	98.12	98.60	98.53	96.71	95.86	100.00	95.79
Dermatology	76.31	82.43	89.07	93.44	94.73	94.15	95.35	94.81	97.92	90.91
COIL20	60.49	81.78	75.87	71.04	51.08	81.04	93.95	87.02	99.44	77.97
MSRA	50.25	54.62	60.42	62.46	57.36	61.02	53.81	57.32	69.48	58.53
PalmData	74.53	78.89	83.78	77.08	87.34	75.82	80.40	85.27	90.25	81.48
Extended YaleB	11.38	49.85	67.73	70.56	72.36	79.35	28.63	64.23	85.43	58.84
Glass	54.21	53.27	51.94	55.61	54.67	55.08	46.26	52.69	56.54	53.36
USPS	63.85	65.72	68.93	66.25	70.03	73.12	81.66	75.64	85.97	72.35
COIL100	50.45	62.37	40.57	49.87	49.32	52.36	70.72	72.13	81.70	58.53
Avg.	57.97	69.42	70.71	71.60	70.61	74.50	71.94	76.11	85.19	–

Table 5Clustering mean NMI (%) compared with different methods, $NMI = \frac{MI(C,C')}{\max(H(C), H(C'))}$.

Dataset/Method	K-means	NCuts	SSC	LRR	LSR	LRRMR	CLR	LSS	NSFRC	Avg.
JAFFE	84.15	96.17	97.76	97.27	97.52	97.80	91.13	96.83	100.00	95.40
Dermatology	84.23	85.04	76.14	86.99	87.69	88.08	91.30	87.35	96.35	87.02
COIL20	73.86	82.96	86.38	81.06	70.01	88.12	97.68	87.64	99.40	85.23
MSRA	57.28	60.10	68.98	64.35	62.32	62.71	69.20	65.88	77.69	65.39
PalmData	90.03	93.14	92.65	88.89	94.27	90.26	94.62	90.36	97.57	92.42
Extended YaleB	13.60	52.36	78.48	71.36	72.86	76.33	32.06	69.02	82.44	60.95
Glass	39.35	35.25	34.71	42.08	38.74	40.47	37.53	43.53	42.80	39.38
USPS	60.24	64.03	68.51	64.39	69.81	73.45	83.68	76.39	84.53	71.67
COIL100	73.91	74.93	71.63	46.84	74.20	54.01	87.88	88.45	93.69	73.95
Avg.	64.07	71.55	75.03	71.47	74.16	74.58	76.12	78.38	86.05	–

6.3. Comparison methods

In our experiments, we compared NSFRC to the following eight algorithms, including a baseline method:

- NCuts [6] is a classic spectral clustering method that treats the grouping problem as a graph partitioning problem. In particular, a generalized eigenvalue system provides a real-valued solution to the graph partitioning problem.
- SSC [7] solves a sparse optimization program whose solution is used in a spectral clustering framework to infer the clustering of the data into subspaces.
- LRR [9] seeks the lowest-rank representation among all the candidates that can represent the data samples as linear combinations of the bases in a given dictionary. It has been shown that LRR can efficiently and effectively perform robust subspace clustering.
- LSR [33] encourages a grouping effect that tends to group highly correlated data together.
- LRRMR [15,16,25] takes into account the nonlinear geometric structures within data by introducing a Laplacian regularizer to the low-rank representation framework.
- CLR [19] allows the data graph itself to be adjusted as part of the clustering procedure, and then the learned graph is exactly c , the number of clusters, connected components.
- LSS [50] dynamically learns an intrinsic affinity matrix from the low-dimensional space of the original data.
- K-means as a baseline method performs clustering in the original feature space of the data.

6.4. Clustering results

(1) Parameter Settings: Except K-means, where clustering is performed on the original features, the algorithms were conducted on the original data to learn a graph for clustering. For NCuts, we constructed the adjacency graph by setting the nearest neighbor number of k from a candidate domain $\{3, 5, 7, 11, 13, 15\}$. The Gaussian kernel $K(x, y) = \exp(-\frac{\|x-y\|_2^2}{td_{\max}^2})$ was adopted, where $d_{\max}$ is the maximal distance between data points, x and y are two data points, and the tuning parameter t is in the range $\{0.01, 0.1, 1, 10, 50, 100\}$. CLR and LSS directly yield clustering results from the learned graphs without spectral clustering. The NCuts uses K-means to segment data into respective groups. Due to its sensitivity to the initialization, we performed K-means 20 times and then reported the mean scores. The regularization parameters of terms for SSC, LRR, LSR, LRRMR, LSS, and NSFRC seriously impact the clustering results. Hence, we conducted experiments when tuning all the parameters in a wide range.

(2) Performance Comparison: Tables 4 and 5 summarize the clustering mean performance for each method on nine datasets. We can see that the performance of K-means lags behind most other algorithms. So, to directly perform clustering in the original data should be avoided, since the data usually contain redundant features and are even contaminated by noise. LRRMR integrates a Laplacian regularizer in the low-rank representation framework to enhance the performance of

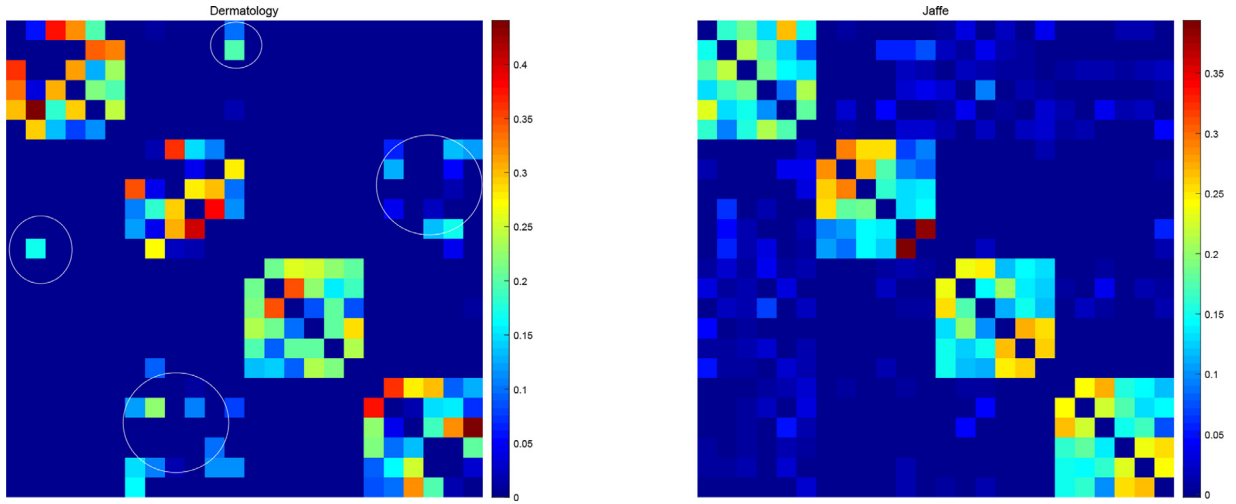


Fig. 1. Affinity matrices learned by the proposed method. Each block corresponds to a class of samples.

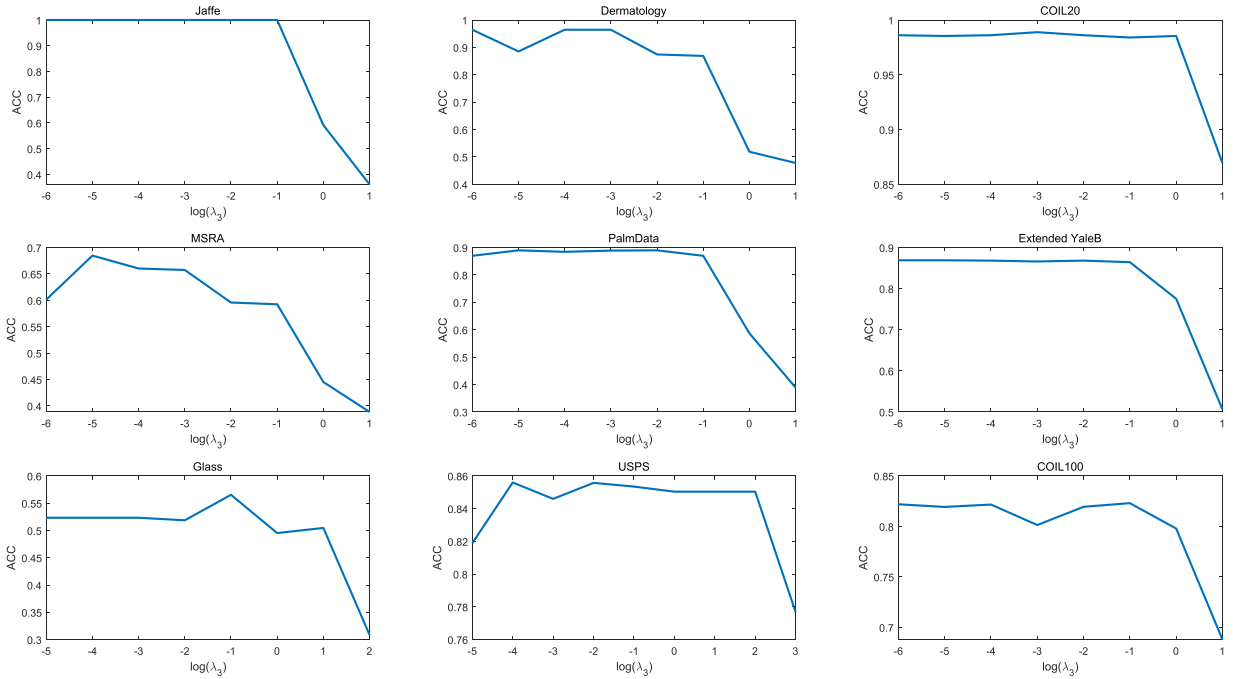


Fig. 2. Clustering ACC versus different values of parameter λ_3 on nine benchmark datasets.

LRR. We can see that LRRMR did not consistently outperform LRR. That is to say, the precomputed affinity matrix in LRRMR may lead to a mismatch of the true similarity between data points. As the state-of-the-art one-step clustering methods, CLR and LSS achieve results comparable to SSC, LRR, LSR, and LRRMR. In spite of the one-step clustering methods that can directly obtain the clustering results, i.e., the learned graph contains exactly c connected components and each component corresponds to a cluster, the graph still may not be the optimal graph for clustering. Because of noise and outliers in reality, data points belonging to the same component may not share the same class label. The last lines of Tables 4 and 5 show that our method achieves the best clustering performance and significantly outperforms the other methods in terms of the average scores. On all datasets except for the Glass dataset, our method outperforms other methods on all measures. On the COIL100 dataset, NSFRC exceeds the second-best LSS by nearly 11% and 5% in terms of ACC and NMI, respectively. On the COIL20, MSRA, Extended YaleB, and USPS datasets, the ACC values of NSFRC are on average nearly 5% higher than those of other methods. In addition, the proposed method still shows encouraging clustering performance on the JAFFE, Dermatology, Glass, and PalmData datasets. Specifically, our NSFRC achieves 100% mean ACC and 100% mean NMI. To see this, in Fig. 1, we show the learned affinity matrices constructed by NSFRC on the Dermatology and JAFFE datasets. The left figure is the

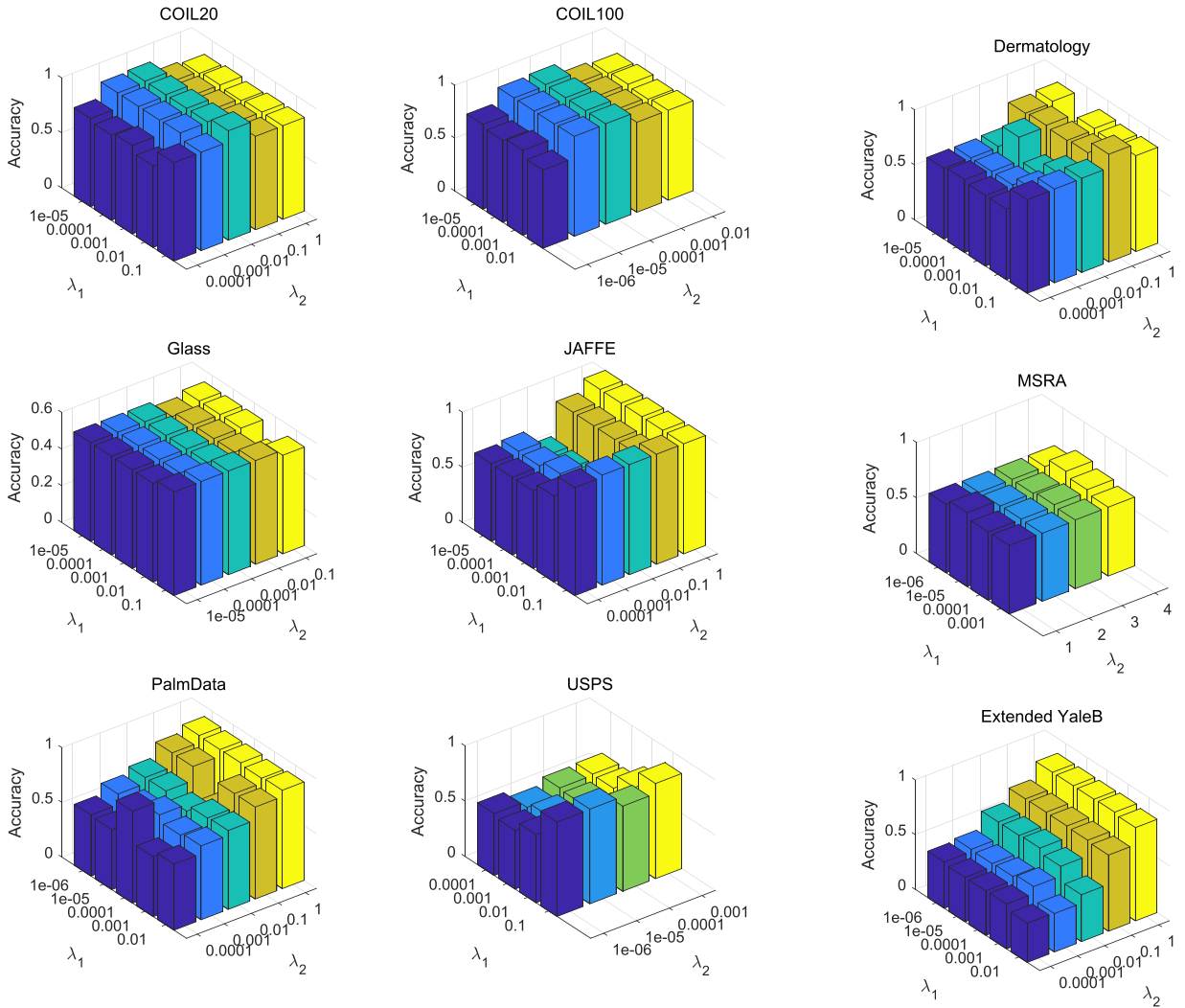


Fig. 3. Clustering ACC versus different values of parameters λ_1 and λ_2 on nine real datasets.

affinity matrix on the Dermatology dataset. It is obvious that there are many messy points, which are circled in white in the affinity matrix. These messy points confuse the intrinsic subspace structure of data and affect the final accuracy. Compared to the affinity matrix of the Dermatology dataset, that of JAFFE has a clearer block diagonal structure, which enables more easy and accurate partitioning of data points. Each entry in the last column of Tables 4 and 5 is an average of all methods in terms of ACC and NMI with respect to a specific dataset. From this, we can see that it is relatively difficult to obtain better clustering results from the MSRA, Extended YaleB, Glass, and COIL100 datasets compared to the JAFFE, COIL20, Dermatology, and PalmData datasets. Our method also shows outstanding performance on the Extended YaleB and COIL100 datasets. In summary, the proposed method benefits from simultaneously considering the global and local information of data, as well as the rank constraint on the Laplacian matrix. Thus, our method shows better data clustering performance than previous work.

6.5. Parameter sensitivity analysis and convergence study

(1) Parameter Sensitivity Analysis: Note that there are three regularization parameters in the objective function in Eq. (7). These parameters can balance the weight of a term to avoid overfitting, error terms, and rank constraint terms. We report the sensitivity of the regular parameters λ_1 , λ_2 , and λ_3 in Figs. 2 and 3. First, we fixed both λ_1 and λ_2 to analyze λ_3 using Fig. 2. As we will see, the range of λ_3 is easy to determine since the best ACC is always obtained when it is less than 1. Then, we fixed λ_3 and selected λ_1 and λ_2 from the range $\{1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10\}$.

λ_3 should not be too large, since the rank constraint is quite strong and may limit the ability of data to represent itself, such that Z cannot learn the intrinsic structure of the data. The highest ACC is always achieved when λ_3 is less than 1. In

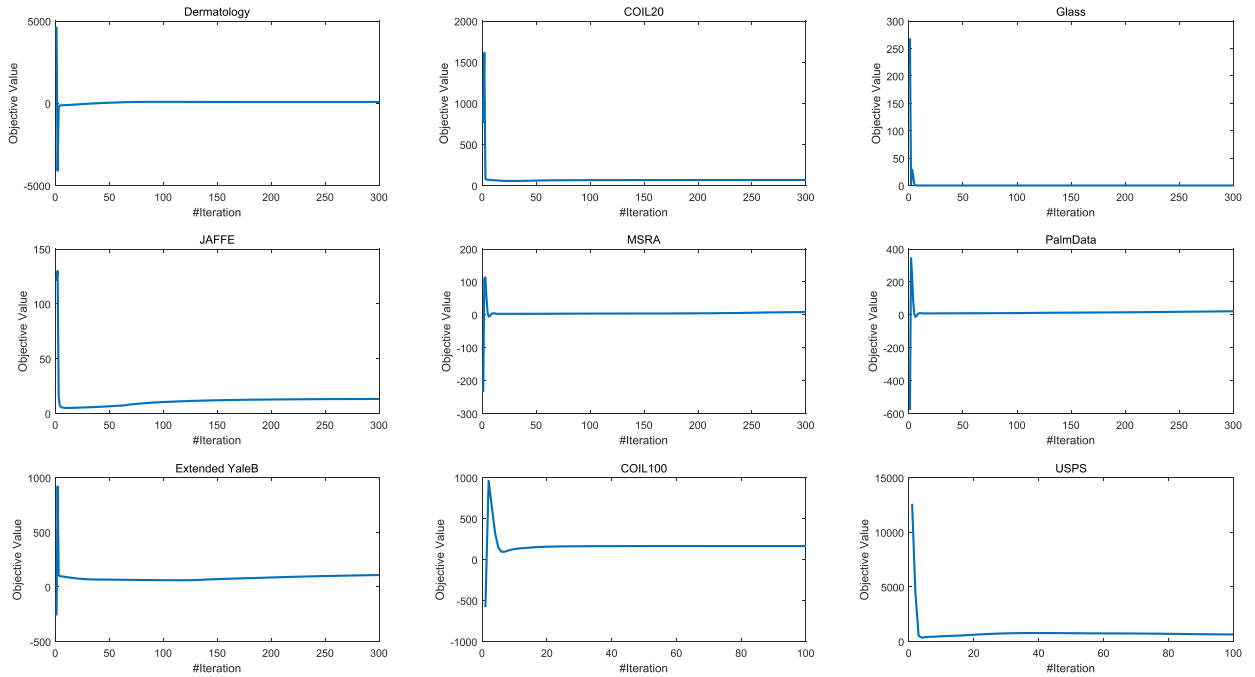


Fig. 4. The convergence curve of our proposed algorithm on nine benchmark datasets. Note that in early iterations, variables are not feasible.

practice, we can tune λ_3 in the range $\{1e-6, 1e-5, 1e-2, 1e-1\}$, since Fig. 2 shows that our NSFRC basically achieves the best clustering results when λ_3 falls in this range. Obviously, the larger λ_3 is, the more important the rank constraint term is in the objective function in Eq. (7). The rank constraint can improve ACC and NMI, but Fig. 2 shows that ACC decreases sharply with the increase of λ_3 . Although the ideal affinity matrix should surely have exactly c connected components, there are many possibilities to choose c connected components from the process of optimization. Consequently, to only depend on the learned graph with exactly c connected components cannot guarantee that the graph is optimal for clustering, since a large λ_3 can cause the rank constraint term to play the dominant role in the objective function in Eq. (7), such that its local and global regularization terms are weakened. This may result in the learned affinity matrix failing to preserve the local and global structures of the data. In Fig. 3, we investigate how λ_1 and λ_2 impact ACC when λ_3 is fixed. There are some important observations. First, the proposed NSFRC is sensitive to λ_1 and λ_2 with fixed λ_3 . Different λ_2 values remove different levels of noise. In practice, different datasets have different degrees of contamination. For example, we can get good performance on Extended YaleB with $\lambda_2 = 1$, but only with $\lambda_2 = 0.01$ on COIL20. Second, our NSFRC is data-driven, since the best performance is achieved by setting different parameter ranges for different datasets. As we can see, the best clustering results can be obtained by tuning λ_1 and λ_2 in a feasible range with fixed λ_3 .

(2) Convergence Study: We now verify the convergence speed of the proposed algorithm through experiments. Fig. 4 shows the trends of objective values with respect to iterations on the nine datasets. Note that variables are not feasible in earlier iterations until the algorithm converges, although they have lower costs. In earlier iterations, before the optimization algorithm converges, the difference between X and $XZ + E$ is large, i.e., $X \neq XZ + E$, hence data points cannot be reconstructed using the current Z . The similarity reflected by Z is not the true similarity between data points.

7. Conclusions

In this paper, we presented a novel subspace clustering approach, called nonnegative self-representation with a fixed-rank constraint (NSFRC) by integrating an adaptive distance regularization term and a fixed-rank constraint on the Laplacian matrix into nonnegative least squares regression to simultaneously discover the local and global structure of data. Specifically, the fixed-rank constraint in our model encourages the affinity matrix to have the exact number of connected components. It results in a better graph for clustering. Furthermore, an error term was introduced in NSFRC such that the obtained solution was robust against noise and outliers. Therefore, NSFRC can be applied to many real-world problems. Experimental results conducted on nine datasets demonstrate that NSFRC can achieve superior clustering performance against state-of-the-art methods. In future work, we will extend NSFRC to semi-supervised learning and multi-view clustering.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partly supported by the University of Macau under Grants: MYRG2018-00035-FST and MYRG2019-00086-FST, and the Science and Technology Development Fund, Macau SAR (File no. 041/2017/A1, 0019/2019/A).

References

- [1] C. Turchetti, L. Falaschetti, A manifold learning approach to dimensionality reduction for modeling data, *Inf. Sci.* 491 (2019) 16–29.
- [2] R. Basri, D.W. Jacobs, Lambertian reflectance and linear subspaces, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2) (2003) 218–233, doi:10.1109/TPAMI.2003.1177153.
- [3] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2011) 52–68, doi:10.1109/MSP.2010.939739.
- [4] S. Theodoridis, K. Koutroumbas, *Spectral Clustering*, Springer New York, New York, NY, pp. 1–5. 10.1007/978-1-4899-7993-3_606-2
- [5] U. Von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* 17 (4) (2007) 395–416.
- [6] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (8) (2000) 888–905, doi:10.1109/34.868688.
- [7] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (11) (2013) 2765–2781, doi:10.1109/TPAMI.2013.57.
- [8] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2) (2009) 210–227.
- [9] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1) (2012) 171–184.
- [10] D.-S. Pham, S. Budhaditya, D. Phung, S. Venkatesh, Improved subspace clustering via exploitation of spatial constraints, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 550–557.
- [11] Y. Chen, G. Li, Y. Gu, Active orthogonal matching pursuit for sparse subspace clustering, *IEEE Signal Process. Lett.* 25 (2) (2018) 164–168.
- [12] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, *J. Comput. Graph. Stat.* 15 (2) (2006) 265–286.
- [13] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? *J. ACM (JACM)* 58 (3) (2011) 11.
- [14] H. Yin, W. Huang, Adaptive nonlinear manifolds and their applications to pattern recognition, *Inf. Sci.* 180 (14) (2010) 2649–2662.
- [15] J. Liu, Y. Chen, J. Zhang, Z. Xu, Enhancing low-rank subspace clustering by manifold regularization, *IEEE Trans. Image Process.* 23 (9) (2014) 4022–4030.
- [16] M. Yin, J. Gao, Z. Lin, Laplacian regularized low-rank representation and its applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (3) (2016) 504–517, doi:10.1109/TPAMI.2015.2462360.
- [17] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: *Advances in neural information processing systems*, 2002, pp. 585–591.
- [18] M. Yin, S. Xie, Z. Wu, Y. Zhang, J. Gao, Subspace clustering via learning an adaptive low-rank graph, *IEEE Trans. Image Process.* 27 (8) (2018) 3716–3728, doi:10.1109/TIP.2018.2825647.
- [19] F. Nie, X. Wang, M.I. Jordan, H. Huang, The constrained laplacian rank algorithm for graph-based clustering (2016) 1969–1976.
- [20] J. Xu, W. An, L. Zhang, D. Zhang, Sparse, collaborative, or nonnegative representation: which helps pattern classification? *Pattern Recognit.* 88 (2019) 679–688.
- [21] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: *Advances in neural information processing systems*, 2001, pp. 556–562.
- [22] R. He, W. Zheng, B. Hu, X. Kong, Nonnegative sparse coding for discriminative semi-supervised learning, in: *CVPR 2011*, 2011, pp. 2849–2856, doi:10.1109/CVPR.2011.5995487.
- [23] M. Slawski, M. Hein, et al., Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization, *Electron. J. Stat.* 7 (2013) 3004–3056.
- [24] G.-F. Lu, Y. Wang, J. Zou, Low-rank matrix factorization with adaptive graph regularizer, *IEEE Trans. Image Process.* 25 (5) (2016) 2196–2205.
- [25] X. Lu, Y. Wang, Y. Yuan, Graph-regularized low-rank representation for destriping of hyperspectral images, *IEEE Trans. Geosci. Remote Sens.* 51 (7) (2013) 4009–4018, doi:10.1109/TGRS.2012.2226730.
- [26] F. Nie, Z. Zeng, I.W. Tsang, D. Xu, C. Zhang, Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering, *IEEE Trans. Neural Netw.* 22 (11) (2011) 1796–1808.
- [27] J. Wen, X. Fang, Y. Xu, C. Tian, L. Fei, Low-rank representation with adaptive graph regularization, *Neural Networks* 108 (2018) 83–96.
- [28] J. Yang, J. Liang, K. Wang, P. Rosin, M. Yang, Subspace clustering via good neighbors, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), doi:10.1109/TPAMI.2019.2913863. 1–1
- [29] X.-J. Shen, S.-X. Liu, B.-K. Bao, C.-H. Pan, Z.-J. Zha, J. Fan, A generalized least-squares approach regularized with graph embedding for dimensionality reduction, *Pattern Recognit.* 98 (2020) 107023, doi:10.1016/j.patcog.2019.107023.
- [30] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8) (2011) 1548–1560, doi:10.1109/TPAMI.2010.231.
- [31] C. Turchetti, L. Falaschetti, A manifold learning approach to dimensionality reduction for modeling data, *Inf. Sci.* 491 (2019) 16–29, doi:10.1016/j.ins.2019.04.005.
- [32] X. Zhu, W. He, Y. Li, Y. Yang, S. Zhang, R. Hu, Y. Zhu, One-step spectral clustering via dynamically learning affinity matrix and subspace, 2017.
- [33] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), *Computer Vision – ECCV 2012*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 347–360.
- [34] T. Jin, R. Ji, Y. Gao, X. Sun, X. Zhao, D. Tao, Correntropy-induced robust low-rank hypergraph, *IEEE Trans. Image Process.* 28 (6) (2019) 2755–2769, doi:10.1109/TIP.2018.2889960.
- [35] L. Zhang, Q. Zhang, B. Du, J. You, D. Tao, Adaptive manifold regularized matrix factorization for data clustering (2017) 3399–3405.
- [36] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in: KDD '14, ACM, New York, NY, USA, 2014, pp. 977–986, doi:10.1145/2623330.2623726.
- [37] K. Fan, On a theorem of weyl concerning eigenvalues of linear transformations i, *Proc. Natl. Acad. Sci.* 35 (11) (1949) 652–655.
- [38] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, now, 2011.
- [39] S. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge University Press, New York, NY, USA, 2004.
- [40] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv:1009.5055 (2010).
- [41] H. Zhang, J. Yang, F. Shang, C. Gong, Z. Zhang, Lrr for subspace segmentation via tractable Schatten- p norm minimization and factorization, *IEEE Trans. Cybern.* 49 (5) (2019) 1722–1734, doi:10.1109/TCYB.2018.2811764.
- [42] G.J. Borse, *Numerical methods with MATLAB: a resource for scientists and engineers*, International Thomson Publishing, 1996.
- [43] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205, doi:10.1109/AFGR.1998.670949.
- [44] S.A. Nene, S.K. Nayar, H. Murase, et al., Columbia object image library (1996).
- [45] X. Zhang, L. Zhang, X. Wang, H. Shum, Finding celebrities in billions of web images, *IEEE Trans. Multimed.* 14 (4) (2012) 995–1007, doi:10.1109/TMM.2012.2186121.
- [46] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, *BMC Bioinform.* 7 (1) (2006) 3, doi:10.1186/1471-2105-7-3.

- [47] A.K. Seewald, Digits-a dataset for handwritten digit recognition.
- [48] A.S. Georgiades, P.N. Belhumeur, D.J. Kriegman, From few to many: generative models for recognition under variable pose and illumination, in: *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (Cat. No. PR00580), 2000, pp. 277–284, doi:[10.1109/AFGR.2000.840647](https://doi.org/10.1109/AFGR.2000.840647).
- [49] D. Dua, E. Karra Taniskidou, UCI machine learning repository, 2017.
- [50] X. Zhu, S. Zhang, Y. Li, J. Zhang, L. Yang, Y. Fang, Low-rank sparse subspace for spectral clustering, *IEEE Transactions on Knowledge and Data Engineering* (2018) 1, doi:[10.1109/TKDE.2018.2858782](https://doi.org/10.1109/TKDE.2018.2858782).