World Wide Web (2020) 23:2107–2127
https://doi.org/10.1007/s11280-020-00802-1

Exploring nonnegative and low-rank correlation for noise-resistant spectral clustering

Zheng Wang[1,2] · Lin Zuo[1] · Jing Ma[3] · Si Chen[4] · Jingjing Li[1] · Zhao Kang[1] · Lei Zhang[5]

Received: 17 September 2019 / Revised: 30 December 2019 / Accepted: 17 February 2020 /
Published online: 12 March 2020

Abstract
Clustering has been extensively explored in pattern recognition and data mining in order to facilitate various applications. Due to the presence of data noise, traditional clustering approaches may become vulnerable and unreliable, thereby degrading clustering performance. In this paper, we propose a robust spectral clustering approach, termed *Non-negative Low-rank Self-reconstruction* (NLS), which simultaneously a) explores the nonnegative low-rank properties of data correlation as well as b) adaptively models the structural sparsity of data noise. Specifically, in order to discover the intrinsic correlation among data, we devise a self-reconstruction approach to jointly consider the nonnegativity and low-rank property of data correlation matrix. Meanwhile, we propose to model data noise via a structural norm, i.e., $\ell_{p,2}$-norm, which not only naturally conforms to genuine patterns of data noise in real-world situations, but also provides more adaptivity and flexibility to different noise levels. Extensive experiments on various real-world datasets illustrate the advantage of the proposed robust spectral clustering approach compared to existing clustering methods.

Keywords Non-negative · Low-rank · Structural sparsity

1 Introduction

In the field of machine learning and data mining, massive research is devoted to clustering technology [21] and its application, such as image segmentation [13, 35] gene expression analysis [23, 28], document analysis [17], content based image retrieval [15, 46], image annotation [22, 24, 37], similarity searches [20, 25, 45, 47, 48].

k-means is one of the most classical clustering models, which has been applied in reality due to its effectiveness and simplicity. The typical process of traditional k-means (TKM)

Lin Zuo and Zheng Wang contributed equally to this work.

This article belongs to the Topical Collection: *Special Issue on Web and Big Data 2019*
Guest Editors: Jie Shao, Man Lung Yiu, and Toyoda Masashi

✉ Lin Zuo
linzuo@uestc.edu.cn

Extended author information available on the last page of the article.

clustering algorithm iteratively assigns each data point to the nearest cluster and computes a new clustering center. However, the "curse of dimensionality" may significantly reduce the performance of TKM [11]. In order to solve this problem, some research works have been done to find low-dimensional projections by reducing dimensions, e.g., PCA, and then performing TKM. In order to further improve the clustering performance, discriminative analysis [9, 11, 49, 50] has been injected into TKM. The work [9, 11], which employed TKM and LDA to obtain cluster labels and learn the most discriminative subspace in an alternating way, has shown the strength of integrating TKM and LDA into a joint framework. Ye et al. [50] proposed a joint framework, namely discriminative k-means (DKM) algorithm, which formalizes the clustering problem into the tracking maximization problem.

In recent years, spectral analysis [5, 8] has been proven to be effective in many applications, especially spectral clustering (SC) [14]. The spectral clustering pays more attention on mining the intrinsic data geometric structures [3, 4, 26, 27, 32, 33, 40, 42–44], which makes it become one of the most successful clustering methods and show more capability in partitioning data with more complicated structures compared to traditional clustering approaches. Therefore, the spectral clustering has been widely applied and shown their effectiveness in various real-world applications, such as image segmentation [35, 51]. The basic idea of spectral clustering is to use different similarity graphs of data points to predict clustering labels. Besides NCut and k-way NCut, a new SC algorithm, i.e., local learning based clustering (LLC) [40], was developed according to the assumption that the cluster label of a data point can be determined by its neighbors, and a kernel regression model was used for label prediction. In [44], discriminative information is used to improve clustering performance by injecting into the construction of the similarity matrix. Most of the existing methods heavily rely on such parametric similarity (or correlation) estimation.

Recently an explosion of emerging Web data, caused by mass storage, fast networks and the widespread availability of media-sharing websites, has been posing more challenges on traditional clustering techniques. On the one hand, as a result of the rapid development of Web data, traditional approaches may require expensive cost of parameter tuning process for calculating a proper data similarity matrix, so they become inapplicable in the face of different data types, different data distributions and so on. Moreover, existing methods mostly focus on using local structure rather than global feature, and data correlation is usually calculated independently. The intrinsic nature of data correlation matrix and global structure of data have not been well explored to facilitate the subsequent spectral clustering process. On the other hand, in real-world scenarios, data may be usually contaminated by unpredictable noise and outliers, which can easily make existing method vulnerable.

In this work, in order to jointly address the aforementioned issues, a novel approach, termed *Non-negative Low-rank Self-reconstruction* (NLS), is proposed for robust spectral clustering. Specifically, the goal of NLS is to collectively (self-)reconstruct a set of data by linearly combining all the data points in the dataset itself. Linear model is possibly the most commonly chosen one due to its ease for use and effectiveness in practice. We first propose to enforce the reconstruction coefficient matrix (i.e., data correlation matrix) to exhibit low-rank property, which not only provides data with a more interpretable representation but also integrates valuable global structural information to identify data correlation. Different from our previous work [39], we have expanded on experiments and analysis, and the competent experiments and analysis show the effectiveness of our method.

In addition, a nonnegative constraint is added purposely on the correlation matrix in order to promote the interpretability (i.e., zero presents no relevance and positive value connotes the degree of relevance). The original motivation of posing the nonnegative constraint on data correlation matrix is to meet the nonnegative property of data similarity, such as the

ones based on Euclidean distance and cosine similarity. By this means, the nonnegative would probably help to characterize the data correlation in a more accurate and interpretable manner, thereby further boosting the clustering performance. Unlike our previous work [44], which posed nonnegative constraint on cluster labels, in this work we utilize nonnegative constraint to describe the inherent and actual correlation among data.

Moreover, on account of that only a (small) part of data in a dataset may be corrupted and different sources of data may have different noise levels, we design a novel noise model by utilizing an effective $\ell_{p,2}$-norm over noise matrix to characterize noise in a more precise way. The $\ell_{p,2}$-norm is able to produce sample-wise sparsity over noise matrix, thereby leading to automatic identification and modelling of noisy samples. At the same time, by changing the value of p, $\ell_{p,2}$-norm can provide greater flexibility on controlling levels of noise and also expand the scope of our method.

The **contributions** of this paper are summarized as follows:

- We propose a novel approach, named as *Non-negative Low-rank Self-reconstruction* (NLS), to facilitate robust spectral clustering. NLS jointly reconstructs data samples in a dataset from themselves, i.e., self-reconstruction, by exploring the intrinsic low-rank nature and nonnegativity of data correlation matrix and precisely modelling sample-wise data noise.
- We devise a nonnegative low-rank approach, which provides data with a more interpretable representation as well as incorporates precious global structural information for identifying data correlation.
- We incorporate an effective $\ell_{p,2}$-norm for characterizing data noise in a more precise way. The $\ell_{p,2}$-norm injects more flexibility to our approach for adapting to different levels of noise and expands applicable range.
- Extensive experiments on multiple real-world datasets illustrate that our proposal outperforms the existing clustering algorithms.

2 Related work

2.1 Data clustering

Data clustering has been a fundamental research topic in the machine learning and data mining communities. k-means based clustering is the most widely used technology because of its simplicity and mathematical tractability, and various methods [9, 11, 18, 19, 49, 50, 53] based on k-means have been proposed successively to improve the severely unaffordable computing time and storage requirements. However, these methods cannot completely overcome the limitations of high complexity and cumbersome memory load. Recently, spectral clustering (SC) [14] raises to prominence and becomes the most successful method available, which uses different similarity graphs of data points to predict clustering labels. Wu et al. [40] proposed a new SC algorithm, termed local learning based clustering (LLC), according to the hypothesis that the cluster label of a data point can be determined by its neighbors. Yang et al. [44] injected discriminative information into the construction of the similarity matrix to improve clustering performance. Deng et al. [10] proposed a novel distributed Policy Decision Point (PDP) model based on SC, called XPDP, to improve the PDP evaluation performance, which combine two-stage clustering and reordering to eliminate the limitation of computational performance of a single PDP.

2.2 A revisit of spectral clustering

We preliminarily review details of spectral clustering. Suppose we are given n data points $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ where d is the dimensionality of data space. The objective of clustering is to partition X into c groups $\{C_j\}_{j=1}^{c}$ such that data points within the same group are close while those in different groups are far from each other. Let us define $Y = [y_1, y_2, \ldots, y_n]^T \in \{0,1\}^{n\times c}$, where $y_i \in \{0,1\}^c$ $(i = 1, 2, \ldots, n)$ is the x_i's cluster indicator vector with the j^{th} entry $y_{ij} = 1$ if $x_i \in C_j$ and $y_{ij} = 0$ otherwise. By following [50], we further define the scaled cluster indicator matrix F as below:

$$F = [F_1, F_2, \ldots, F_n]^T = Y(Y^TY)^{-1/2}$$

where F_i is the scaled cluster indicator of x_i. Note that the j^{th} column of F indicates which data points belong to the j^{th} cluster C_j, and it is in the following form:

$$f_j = [\underbrace{0, \ldots, 0}_{\sum_{k=1}^{j-1} n_k}, \underbrace{1/\sqrt{n_j}, \ldots, 1/\sqrt{n_j}}_{n_j}, \underbrace{0, \ldots, 0}_{\sum_{k=j+1}^{c} n_k}]$$

where n_j is the number of data points in the j^{th} cluster.

Below is a general objective function of spectral clustering:

$$\begin{aligned} &\min_F Tr(F^TLF) \\ &\text{s.t. } F = Y(Y^TY)^{-1/2} \end{aligned} \tag{1}$$

where $Tr(\cdot)$ is the trace operator and L denotes a graph Laplacian matrix computed according to the data local structure using different strategies. Given the dataset $X = [x_1, x_2, \ldots, x_n]$, we can construct an undirected graph which can be represented by the weighted adjacency matrix $S = (s_{ij})_{i,j=1,2,\ldots,n}$. Here $s_{ij} > 0$ indicates that x_i and x_j are connected, and $s_{ij} = 0$ means they are not connected.

A common way to compute the edge weight is defined as follows:

$$s_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right), & \text{if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{otherwise.} \end{cases}$$

where $\mathcal{N}_k(\cdot)$ is the function for searching for the k nearest neighbors and σ is the bandwidth parameter. Denote D as a diagonal matrix with its diagonal $d_{ii} = \sum_j D_{ij}$, then the graph Laplacian can be calculated as

$$L = D - S$$

If we instead use the normalized graph Laplacian in (1),

$$L_n = D^{-1/2}LD^{-1/2} = I_n - D^{-1/2}SD^{-1/2}$$

then the objective function turns out to be the well-known spectral clustering algorithm, namely normalized cut [35]. Similarly, if we replace L with L_l which is the graph Laplacian matrix obtained by the local learning [40], then the objective function in (1) becomes the local learning clustering (LLC).

Note that the discretization constraint on F makes (1) difficult to solve. A practical way to handle this problem is to make a relaxation to allow F to be of continuous values, and then use eigenvalue decomposition on the corresponding graph Laplacian matrix.

3 The proposed approach

In this section, we elaborate the details of the proposed robust spectral clustering approach, including a nonnegative low-rank self-reconstruction process for learning data correlation matrix and a structural noise modelling component for handling noisy data.

Given a set of data points $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, where each column vector $x_i \in \mathbb{R}^d$ represents a datum and d is the dimensionality of feature space. Ideally, the data in X should not contain any noise. Nonetheless, in real-world scenarios, data would be inevitably contaminated by various unpredictable factors, such as distortion, transmission error, malicious tempering, etc. Intuitively, a reasonable assumption is that only a (small) proportion of data are influenced by noise, i.e., the noise should be sparse. Furthermore, the noise levels in different sources of data may vary significantly, which poses great challenges for precise noise control using a unified model. In this case, the major objective of this work is to devise an effective spectral clustering approach, which is able to capture the genuine correlation among data, identify noisy samples as well as suppress influence of different levels of noise effectively.

3.1 Nonnegative low-rank self-reconstruction

As aforementioned in Section 1, we employ linear model for reconstructing data due to its ease for use and effectiveness in practice. Given the data matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, the i-th datum x_i can be represented as a linear combination of m basis vectors in a dictionary $B = [b_1, b_2, \ldots, b_m] \in \mathbb{R}^{d\times m}$:

$$x_i = Bw_i + \varepsilon_i,$$

where $w_i \in \mathbb{R}^m$ is the reconstruction coefficient of x_i and $\varepsilon_i \in \mathbb{R}^d$ is the noise on x_i. By denoting the linear model in concise matrix form, we have:

$$X = BW + \mathcal{E}, \tag{2}$$

where $W = [w_1, w_2, \ldots, w_n] \in \mathbb{R}^{m\times n}$ is the reconstruction coefficient matrix, which can be regarded as either the new representation of data or the correlation of data in X and the basis in B. $\mathcal{E} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n] \in \mathbb{R}^{d\times n}$ indicates the noise matrix of all data in X.

In order to infer the data correlation within X, a reasonable way is to exploit X itself as the dictionary to perform (self-)reconstruction as follows:

$$X = XW + \mathcal{E}. \tag{3}$$

In this way, we can regard W as the new representation of X or correlation between data in X and themselves. The i-th column of W, i.e., $w_i = [w_{1i}, w_{2i}, \ldots, w_{ni}]^T \in \mathbb{R}^n$, is the reconstruction coefficient vector of the i-th datum x_i. The coefficient w_{ji} measures the contribution of the j-th datum x_j on the reconstruction of x_i.

In order to model data correlation, one may use sparse constraint (e.g., $\|W\|_1$) [12] for obtain an optimized W. Indeed, it may uncover the local structure of X and achieve the denoising purpose to some extent; nevertheless, such sparse constraint may easily cause the data correlation W ignoring the precious global structural information. Based on this analysis, we propose to employ low-rank constraint, which has been proven to be more

proper for characterize the data correlation as well as explore the global information. The general optimization problem is stated as below:

$$\begin{aligned} &\min_{W,\mathcal{E}} \text{rank}\ (W) + \lambda\Omega(\mathcal{E}), \\ &\text{s.t. } X = XW + \mathcal{E}, \end{aligned} \tag{4}$$

where the first term calculates the rank of W, the second term $\Omega(\mathcal{E})$ enables certain forms of sparsity for modelling data noise, and λ is a balance parameter determining the contribution of these two terms.

It is known that optimizing the rank function is difficult due to its discreteness. To handle this problem, a common way is to relax the rank optimization to a nuclear norm optimization, which is convex. Thus, the problem in (4) is transformed to

$$\begin{aligned} &\min_{W,\mathcal{E}} \|W\|_* + \lambda\Omega(\mathcal{E}), \\ &\text{s.t. } X = XW + \mathcal{E}. \end{aligned} \tag{5}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, i.e., the sum of the matrix's singular values.

If we solve the problem in (5), the optimized W^* would probably be mixing-signed, which makes it difficult to describe data correlation in an interpretable way. Intuitively, given two data points, if they are relevant, in order to quantify the degree of their correlation, we may use a positive value as measurement; otherwise, we use value zero to indicate the fact that they are not relevant. In other words, this intuition implies us that W^* should be nonnegative. It has been shown that nonnegative analysis would probably boost performance [41]. Accordingly, we impose an explicit nonnegative constraint over data correlation matrix W, and the problem is reformulated as

$$\begin{aligned} &\min_{W,\mathcal{E}} \|W\|_* + \lambda\Omega(\mathcal{E}), \\ &\text{s.t. } X = XW + \mathcal{E} \wedge W \geq 0. \end{aligned} \tag{6}$$

In the next part, we will introduce how to precisely characterize data noise $\mathcal{E}$, i.e., specify $\Omega(\mathcal{E})$ to further reinforce the establishment of data correlation.

3.2 Modelling data noise

In real-world cases, data would be inevitably contaminated by various types of noise, such as distortion, transmission error, etc. Normally, it is reasonable to assume that only a (small) proportion of data are actually corrupted and the rest are clean. Suppose we use ℓ_1-norm to model data noise as below:

$$\|\mathcal{E}\|_1 = \sum_{j=1}^{d}\sum_{i=1}^{n} |\varepsilon_{ji}|, \tag{7}$$

where ε_{ji} indicates the j-th element of ε_i. Such modelling would probably cause that the identified noise propagated to all the data, thereby negatively influencing other clean samples and further degrading the performance. In order to avoid such noise propagation problem, a more effective way is to intentionally shape noise according to certain reason-

able assumption. As aforementioned, noise may only occur in a (small) proportion of data, which inspires us to exploit structural modelling approach, such as $\ell_{1,2}$-norm:

$$\|\mathcal{E}\|_{1,2} = \sum_{i=1}^{n} \|\varepsilon_i\| . \tag{8}$$

As we can see from the definition, $\ell_{1,2}$-norm of $\mathcal{E}$ actually accounts to the ℓ_1-norm of the vector $[\|\varepsilon_1\|, \|\varepsilon_2\|, \ldots, \|\varepsilon_n\|]$, which implies that it helps to induce sample-wise sparsity. In other words, some columns of $\mathcal{E}$ shrink to zero. For better understand, we use a visual example to illustrate difference of $\ell_{1,2}$-norm and ℓ_1-norm in Figure 1. As we can see, $\ell_{1,2}$-norm enforces sample-wise sparsity on $\mathcal{E}$ to achieve accurate identification of noisy samples (i.e., $\{x_2, x_4, x_6, x_{10}\}$ in red), while the uncontrolled ℓ_1-norm tends to propagate data noise to the whole dataset, thereby contaminating more samples (i.e., $\{x_1, x_3, x_5, x_7, x_8, x_{11}\}$ in blue).

In order to further increase the flexibility for handling different corrupt levels of data, we propose to generalize of $\ell_{1,2}$-norm to $\ell_{p,2}$-norm:

$$\|\mathcal{E}\|_{p,2} = \sum_{i=1}^{n} \|\varepsilon_i\|^p , \tag{9}$$

where $0 < p < 2$. Note that when p is set to 1, (9) is identically equivalent to (8). As p varies, the $\ell_{p,2}$ norm may help to induce different levels of sparsity, which corresponds to different levels of noise intended to be recognized. For instance, when $p \to 2$, the $\ell_{p,2}$ norm tends to become ℓ_2 norm, which will not induce any sparsity, thereby disabling the ability of NLS identifying noisy samples. In contrast, small p would probably induce too much unnecessary sample-wise sparsity, which may force NLS to "over-identify" noisy samples, thereby degrading clustering performance.

Thus, by substituting (9) into (6), we have

$$\begin{aligned} &\min_{W,\mathcal{E}} \|W\|_* + \lambda\|\mathcal{E}\|_{p,2}, \\ &\text{s.t. } X = XW + \mathcal{E} \wedge W \geq 0. \end{aligned} \tag{10}$$

In the next part, we will introduce the optimization details of (10).

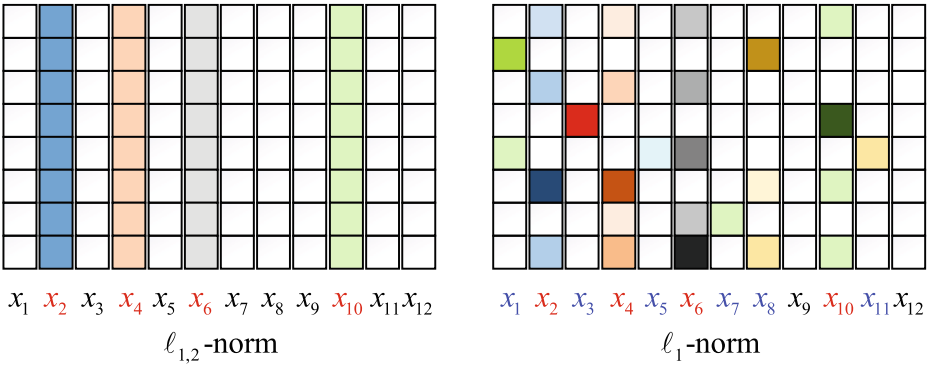


Figure 1 Illustration of difference between $\ell_{1,2}$-norm and ℓ_1-norm

3.3 Optimization

In this subsection, we present an alternating algorithm for optimizing the problem in (10). We first transform the original problem (10) by adding an additional variable V for facilitating the optimization:

$$\begin{aligned} &\min_{W,\mathcal{E},V} \|V\|_* + \lambda\|\mathcal{E}\|_{p,2}, \\ &\text{s.t.} \begin{cases} X = XW + \mathcal{E}, \\ W \geq 0, \\ V = W. \end{cases} \end{aligned} \tag{11}$$

By utilizing Augmented Lagrange Multiplier (ALM), the above constrained problem can be further changed to the following form:

$$\begin{aligned} &\min_{W,\mathcal{E},V,P,Q} \|V\|_* + \lambda\|\mathcal{E}\|_{p,2} \\ &\quad + Tr\left(P^T(X - XW - \mathcal{E})\right) + Tr\left(Q^T(W - V)\right) \\ &\quad + \frac{\alpha}{2}\left(\|X - XW - \mathcal{E}\|_F^2 + \|W - V\|_F^2,\right) \\ &\text{s.t. } W \geq 0, \end{aligned} \tag{12}$$

where $Tr(\cdot)$ is the trace of a matrix. $P \in \mathbb{R}^{d\times n}$ and $Q \in \mathbb{R}^{n\times n}$ are Lagrange multipliers for the two equality constraints, and $\alpha > 0$ is a trade-off parameter. In order to solve the problem (12), we alternatingly update W, $\mathcal{E}$, V, P, Q.

Update V By fixing W, $\mathcal{E}$, P, Q, we have the following sub-problem:

$$\min_V \|V\|_* - Tr\left(Q^T V\right) + \frac{\alpha}{2}\|W - V\|_F^2, \tag{13}$$

which is equivalent to

$$\min_V \frac{1}{2}\left\|V - \left(W + \frac{Q}{\alpha}\right)\right\|_F^2 + \frac{1}{\alpha}\|V\|_*. \tag{14}$$

The above optimization problem with nuclear norm regularization can be efficiently solved by singular value thresholding [6].

Update W By fixing $\mathcal{E}$, V, P, Q, the problem (12) is reduced to

$$\begin{aligned} &\min_W Tr\left(-P^T XW + Q^T W\right) \\ &+ \frac{\alpha}{2}\left(\|X - XW - \mathcal{E}\|_F^2 + \|W - V\|_F^2\right), \\ &\text{s.t. } W \geq 0, \end{aligned} \tag{15}$$

which can be solved by applying the following multiplicative update rule:

$$w_{ij} \leftarrow w_{ij} \times \frac{H_{ij}}{(U\tilde{W})_{ij}}, \tag{16}$$

where $\tilde{W}$ is the outcome in the previous iteration. $U = X^T X + I$ and I is identity matrix of size $n \times n$. $H = (X^T X - X^T\mathcal{E} + V + \frac{1}{\alpha}(X^T P - Q))$.

Update $\mathcal{E}$ Now let us fix W, V, P, Q, then the problem can be transformed to

$$\begin{aligned} &\min_{\mathcal{E}} \frac{\alpha}{2} \|X - XW - \mathcal{E}\|_F^2 - Tr\left(P^T\mathcal{E}\right) + \lambda \|\mathcal{E}\|_{p,2}, \\ \Leftrightarrow\ &\min_{\mathcal{E}} \frac{1}{2} \left\|\mathcal{E} - \left(X - XW + \frac{P}{\alpha}\right)\right\|_F^2 + \frac{\lambda}{\alpha}\|\mathcal{E}\|_{p,2}. \end{aligned} \tag{17}$$

In order to solve the sub-problem in (17), we first consider the following alternative problem:

$$\min_{\mathcal{E}} \frac{1}{2} \left\|\mathcal{E} - \left(X - XW + \frac{P}{\alpha}\right)\right\|_F^2 + \frac{\lambda}{\alpha} Tr(\mathcal{E}^T Z \mathcal{E}), \tag{18}$$

where Z is a diagonal matrix, whose i-th diagonal element is computed as

$$Z_{ii} = \frac{p}{2\left\|\varepsilon_i\right\|^{2-p}} \tag{19}$$

Note that Z is derived from $\mathcal{E}$ which makes it difficult to directly optimize (18). Hence, we devise an iterative algorithm to handle the problem. To be more specific, in each iteration we alternatingly update Z and $\mathcal{E}$. We first calculate Z with the obtained $\mathcal{E}$ in the previous iteration, then $\mathcal{E}$ is updated via a close-form solution. By fixing Z and setting the derivative of (18) w.r.t. $\mathcal{E}$ to zero, we arrive at

$$\mathcal{E} = \left(I + \frac{2\lambda}{\alpha} Z\right)^{-1} \left(X - XW + \frac{1}{\alpha} P\right). \tag{20}$$

We can show that by iteratively solving the problem (18), the optimal solution can be obtained for the problem (17). To this end, we present the following lemmas and theorem.

Lemma 1 *Let ε_i be the i^{th} column of the updated $\mathcal{E}$ in previous iteration and $\tilde{\varepsilon}_i$ be the i^{th} column of the variable $\tilde{\mathcal{E}}$ in current iteration, then the following inequality holds:*

$$\left\|\tilde{\varepsilon}_i\right\|^p - \frac{p\left\|\tilde{\varepsilon}_i\right\|^2}{2\left\|\varepsilon_i\right\|^{2-p}} \leq \left\|\varepsilon_i\right\|^p - \frac{p\left\|\varepsilon_i\right\|^2}{2\left\|\varepsilon_i\right\|^{2-p}} \tag{21}$$

Proof Please refer to Appendices A and B for more details. □

Lemma 2 *Given $\mathcal{E} = [\varepsilon_1, \varepsilon_2 \ldots, \varepsilon_n]$, where ε_i is the i^{th} column of $\mathcal{E}$, then we have the following conclusion:*

$$\sum_{i=1}^{n} \left\|\tilde{\varepsilon}_i\right\|^p - \sum_{i=1}^{n} \frac{p\|\tilde{\varepsilon}_i\|^2}{2\left\|\varepsilon_i\right\|^{2-p}} \leq \sum_{i=1}^{n} \left\|\varepsilon_i\right\|^p - \sum_{i=1}^{n} \frac{p\left\|\varepsilon_i\right\|^p}{2\left\|\varepsilon_i\right\|^{2-p}} \tag{22}$$

Proof It can be easily seen that by summing up the inequalities over all the columns in $\mathcal{E}$ in Lemma 1 we are able to obtain the conclusion of Lemma 2. □

Theorem 1 *At each iteration (line 3-4) of Algorithm 1, the value of the objective function in (17) monotonically decreases.*

Proof Please refer to Appendices A and B for more details. □

Theorem 1 guarantees the convergence of Algorithm 1.
The algorithm is shown Algorithm 1.

Algorithm 1 Algorithm for optimizing the $\ell_{p,2}$-norm regularized problem in (18).

Input: Data matrix X, correlation matrix W, Lagrange multiplier P, parameters λ, α and p;
Output: $\mathcal{E}$;
1: Initialize $\mathcal{E}$;
2: **repeat**
3: Compute the diagonal matrix Z according to (19);
4: Update $\mathcal{E} = \left(I + \frac{2\lambda}{\alpha} Z\right)^{-1} \left(X - XW + \frac{1}{\alpha} P\right)$;
5: **until** convergence
6: **return** $\mathcal{E}$;

Update P and Q Given $W, \mathcal{E}, V$, we may update the Lagrange multipliers P and Q as follows:

$$\begin{cases} P \leftarrow P + \alpha(X - XW - \mathcal{E}) \\ Q \leftarrow Q + \alpha(W - V) \end{cases} \tag{23}$$

We summarize the optimization for the problem (12) in Algorithm 2.

Algorithm 2 Algorithm for optimizing the problem in (12).

Input: Data matrix X and parameters λ, p;
Output: W;
1: Initialize $\mathcal{E}$, P, Q, W, $max_\alpha = 10^{10}$, $\rho = 1.1$;
2: **repeat**
3: Update V by solving the problem in (13);
4: Update W by solving the problem in (15) with the multiplicative update rule in (16);
5: Update $\mathcal{E}$ by running Algorithm 1;
6: Update P and Q according to (23);
7: Update $\alpha = \min(\rho\alpha, \max_\alpha)$;
8: **until** convergence
9: **return** W;

Next, we briefly analyze the computational complexity of Algorithm 2. The computation of V is dominated by the SVD operation in the optimization of the problem (13), which requires the cost of $O(n^3)$. The update of W mainly relies on the update rule in (16), which costs $O(n^3)$. As to the calculation of $\mathcal{E}$ in Algorithm 1, the time cost of (20) is $O(n^3)$. Considering that the numbers of iterations is far smaller than n, the total time cost of Algorithm 2 is $O(n^3)$.

3.4 Overall spectral clustering

Given the optimized data correlation matrix W^*, where the element w^*_{ji} indicates the directed relation from the j-th datum to the i-th datum, i.e., the contribution of the j-th datum in the reconstruction process of the i-th datum. Intuitively, it is reasonable to assume that a given datum is only related to a few samples. To this end, we choose to reserve k nearest neighbors in terms of the data correlation and construct the sparse data correlation matrix $\hat{W}^*$, where k is an empirical parameter.

Note that NLS model does not guarantee that $\hat{W}^*$ is symmetric, which implies that in most cases $\hat{w}^*_{ij} \neq \hat{w}^*_{ji}$. In general, most spectral clustering algorithms use symmetric affinity matrix to partition data. Following this convention, we practically add $\hat{W}^*$ and its transpose to guarantee the constructed graph is undirected and the affinity matrix is symmetric, which will facilitate the subsequent typical spectral clustering procedure:

$$A = \frac{\hat{W}^* + (\hat{W}^*)^T}{2}. \tag{24}$$

Finally, we perform spectral clustering by applying eigen-value decomposition on the Laplacian matrix of A and discretizing clustering labels (e.g., spectral rotation or k-means). We summarize the overall clustering procedure in Algorithm 3.

Algorithm 3 Spectral clustering based on NLS algorithm.

Input: Data matrix X;
Output: Clustering label matrix F;
1: Compute data correlation matrix W^* by performing Algorithm 1;
2: Find k nearest neighbors for each datum and construct sparse correlation matrix $\hat{W}^*$;
3: Symmetrize $\hat{W}^*$ as follows:

$$A = \frac{\hat{W}^* + (\hat{W}^*)^T}{2};$$

4: Compute Laplacian matrix L of A;
5: Perform eigen-decomposition on L and obtain F;
6: Discretize F by performing spectral rotation;
7: **return** F;

4 Experiments

In this section, we evaluate the effectiveness of the proposed NLS spectral clustering algorithm by comparing it to the existing approaches on various datasets.

4.1 Datasets

In the following experiments, we evaluate on five datasets, including Jaffe, Umist, Yale, Lenses and Auto.

Jaffe [31] The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. Images are 256×256 gray level, in .tiff format, with no compression.

Umist [16] (now: The Sheffield Face Database). It consists of 564 images of 20 people. Each covering a range of poses from profile to frontal views. Subjects cover a range of race/sex/appearance. The files are all in PGM format, approximately 220×220 pixels in 256 shades of grey.

Yale [2] The Yale Face Database contains 165 grayscale images in GIf format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

Lenses [7] The lenses data set is a data set that tries to predict whether people will need soft contact lenses, hard contact lenses or no contacts.

Auto [1] Auto data set is from the machine learning repositor of UCI, donated by Jeffrey C. Schlimmer. It contains 205 instances of 1985 model import car and truck specifications, 3 types of entities, 26 number of attributes.

4.2 Experimental settings

We compare our algorithm to six existing clustering approaches, including k-means clustering (TKM), discriminative k-means (DKM) clustering [50], Spectral Clustering (SC), Normalized Cuts (NCuts) [51], Local Learning Clustering (LLC) [40], CLGR [38] and LRR [29, 30]. Besides, we also evaluate three variants of our approach, i.e., LS, LS_1 and NLS_1. LS is the version of our approach NLS without nonnegative constraint. LS_1 and NLS_1 are the corresponding versions of LS and NLS using ℓ_1-norm instead of $\ell_{1,2}$-norm.

For spectral clustering algorithms which need to specify the number of neighbors, we always set it to $k = 5$. We perform the self-tuning algorithm [52] to determine an adaptive bandwidth. For fair comparison, the trade-off parameters in all the comparison algorithms are consistently tuned from the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$; for the parameter p in our approach, we set it in the range of $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75\}$; and the best results are reported. To reduce statistical variation, each clustering algorithm is run repeatedly for 10 times and the average results are reported.

4.3 Evaluation metrics

Following conventional clustering study, we use Accuracy (ACC) and Normalized Mutual Information (NMI) as our evaluation metrics in the subsequent parts.

Denote q_i as the clustering label result from a clustering algorithm and p_i as the corresponding ground truth label of x_i, then we define ACC as

$$ACC = \frac{\sum_i \delta(p_i, map(q_i))}{n}, \tag{25}$$

where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise, and $map(q_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithms. A larger ACC indicates a better clustering performance.

For any two arbitrary variable P and Q, NMI is defined as follows [36]:

$$NMI = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}} \tag{26}$$

where $I(P, Q)$ computes the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q. Denote t_l as the number of data in the cluster $\mathcal{C}_l (1 \leq l \leq c)$

Table 1 Overall ACC performance (%) comparison to the existing algorithms

	Jaffe	Umist	Yale	Lenses	Auto
TKM	73.1 ± 4.6	38.8 ± 1.0	38.4 ± 1.1	54.6 ± 3.6	35.1 ± 1.6
DKM	84.5 ± 0.6	44.2 ± 0.2	42.4 ± 0.7	66.7 ± 2.0	39.1 ± 0.1
SC	78.5 ± 6.8	53.3 ± 2.0	48.9 ± 2.3	54.4 ± 5.7	31.7 ± 1.1
NCuts	84.1 ± 2.0	51.9 ± 1.3	47.4 ± 1.8	55.8 ± 7.1	31.6 ± 0.8
LLC	83.6 ± 5.5	43.1 ± 0.7	51.9 ± 2.4	55.0 ± 6.7	37.0 ± 1.1
CLGR	82.6 ± 3.8	55.0 ± 1.4	48.3 ± 2.0	48.6 ± 7.4	39.7 ± 0.4
LRR	91.2 ± 3.9	68.8 ± 0.5	46.1 ± 1.4	56.0 ± 2.4	37.7 ± 0.5
NLS_ℓ_1	97.5 ± 0.3	76.6 ± 0.3	50.2 ± 0.3	60.8 ± 1.9	38.5 ± 0.2
NLS	99.2±0.8	77.8±0.5	53.0±0.7	68.3±2.9	41.5±0.4

generated by a clustering algorithm and $\tilde{t}_h (1 \leq l \leq c)$ as the number of data points from the h^{th} ground truth class. NMI metric is then defined as below [36]:

$$NMI = \frac{\sum_{l=1}^{c}\sum_{h=1}^{c} t_{l,h} \log\left(\frac{n \times t_{l,h}}{t_l \tilde{t}_h}\right)}{\sqrt{\left(\sum_{l=1}^{c} t_l \log\frac{t_l}{n}\right)\left(\sum_{h=1}^{c} \tilde{t}_h \log\frac{\tilde{t}_h}{n}\right)}} \tag{27}$$

where $t_{l,h}$ is the number of data samples that lie in the intersection between $\mathcal{C}_l$ and h^{th} ground truth class. Likewise, a larger NMI indicates a better clustering performance.

4.4 Comparison

In this subsection, we conduct empirical studies on five datasets to show the performance comparison of existing algorithms and our proposed method. The comparison results of ACC and NMI are listed in Tables 1 and 2, respectively. From these results, we can derive the following observations and analysis.

Table 2 Overall NMI performance (%) comparison to the existing algorithms

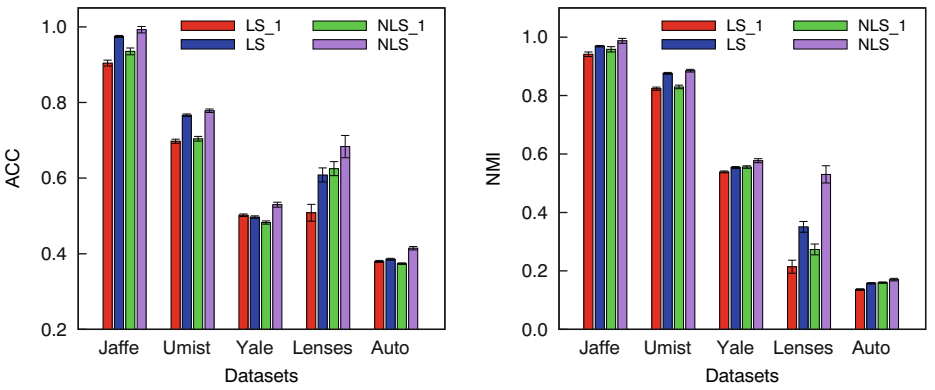
	Jaffe	Umist	Yale	Lenses	Auto
TKM	81.3 ± 2.2	58.3 ± 0.6	46.7 ± 0.9	26.2 ± 8.9	15.9 ± 0.6
DKM	90.1 ± 0.6	61.8 ± 0.5	47.7 ± 0.5	4.7 ± 0.7	16.8 ± 0.5
SC	84.1 ± 2.7	74.4 ± 0.8	54.8 ± 1.2	22.1 ± 7.5	13.5 ± 1.1
NCuts	93.9 ± 1.6	73.3 ± 0.8	55.2 ± 1.3	32.1 ± 8.0	13.2 ± 1.2
LLC	89.5 ± 2.5	65.8 ± 0.7	54.9 ± 1.9	20.9 ± 7.7	15.0 ± 0.4
CLGR	92.1 ± 1.1	76.6 ± 0.8	53.2 ± 1.4	23.1 ± 7.7	16.9 ± 0.9
LRR	94.6 ± 2.3	82.5 ± 0.4	53.3 ± 1.1	29.9 ± 3.1	15.7 ± 0.4
NLS_ℓ_1	96.9 ± 0.1	87.6 ± 0.4	55.4 ± 0.2	35.1 ± 3.9	15.8 ± 0.3
NLS	98.7±0.7	88.5±0.4	57.7±0.5	53.0±3.7	17.0±0.2

– In most cases, NLS_1 achieves better performances than TKM, DKM, SC and NCuts. This phenomenon indicates that jointly exploring nonnegativity and low rank properties of data correlation as well as suppressing data noise can be of benefit for achieving satifactory clustering performance.
– NLS consistently outperforms NLS_1. This observation implies that the structural modelling using $\ell_{p,2}$-norm is able to better capture the genuine distribution of data noise than ℓ_1-norm. NLS exploits $\ell_{p,2}$-norm to "actively" enforce sample-wise sparsity on data noise, thereby accurately identifying and quantifying noisy samples; nonetheless, NLS_1 models data noise using ℓ_1-norm, which tends to "accidentally" propagate noise across all data samples and cause inventible contamination of clean samples.
– NLS always achieves the best performance comparing to other comparison algorithms on all five datasets. Both LLC and CLGR exploit additional knowledge, e.g., discriminative information, to enhance the exploration of data correlation; hence, in most cases they achieve better performance than TKM, DKM, SC, NCuts and LRR. However, compared to our method, they do not fully take any consideration into intrinsic properties of data correlation, namely nonnegativity and low rankness, as well as structural modelling of data noise, which together guarantees a reliable and robust process for data self-reconstruction and the subsequent spectral clustering.

4.5 Nonnegativity and noise modelling

In this part, we evaluate efficacy of the nonnegative constraint and $\ell_{p,2}$-norm for modelling data noise. Specifically, we compare NLS with its variant that uses ℓ_1-norm, denoted as NLS_1. To the end of illustrating the effect of the nonnegative consideration, we also compare NLS and NLS_1 to their counterparts (i.e., LS and LS_1, respectively) that do not pose nonnegative constraint.

The experimental results are reported in Figure 2. Figures 2a and b illustrate, respectively, ACC performance and NMI performance of the four comparison algorithms on five datasets. We can attain the following observations and conclusions:



(a) ACC (b) NMI

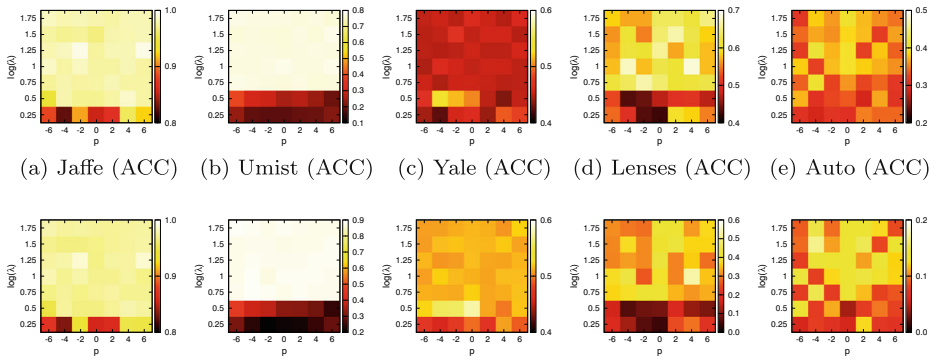
Figure 2 Effects of nonnegative constraint and data noise modelling on five datasets. **a** and **b** gives ACC performance and NMI performance, respectively

– NLS and NLS_1 consistently outperform their counterparts, i.e., LS and LS_1, respectively. This fact clearly indicates that the nonnegativity consideration helps to achieve performance improvement in terms of both ACC and NMI metrics. As analyzed before, by explicitly pose nonnegative constraint on data correlation matrix W, we can formulate a better process for jointly reconstructing all data samples, thereby characterizing correlation among data in a more interpretable manner (i.e., value 0 indicates that two data points are not related, and a positive value measures the degree of relationship of two data samples.)
– NLS and LS always gain better performance than NLS_1 and LS_1, respectively. Similar to the analysis in Section 4.4, such observation reveals the superior efficacy of $\ell_{p,2}$-norm for structurally modelling data noise. Compared to ℓ_1-norm, $\ell_{p,2}$-norm not only captures the genuine noise distribution but also provides sufficient flexible control on different noise levels.
– By comparing LS and NLS_1, we can see that in some cases, the former performs better than the latter one; while in other cases, we observe that LS achieves slightly worse results that NLS_1 or they gain comparable performance. Although $\ell_{p,2}$-norm may contribute more than nonnegative constraint in more cases, it is not easy to draw any conclusion that which component is more important in our approach. In practice, we should suggest integrating both of them to achieve better performance.

4.6 Sensitivity analysis

In this subsection, we analyze the sensitivity of parameters in our approaches on five datasets. Specifically, we evaluate the joint effects of λ and p on NLS as well as the effects of λ on NLS_1. As aforementioned, we set λ in the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$ and p in the range of $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75\}$.

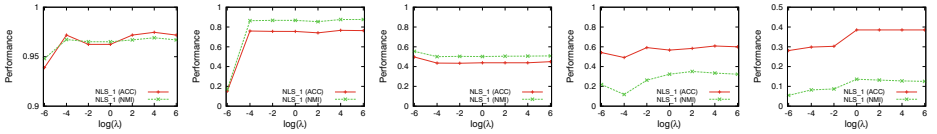
While Figure 3a–e report ACC performance of NLS w.r.t. to λ and p on the five evaluated datasets, respectively; Figure 3f–j illustrate NMI performance of NLS. For different datasets, the distributional patterns of parameter combinations vary. In most cases, both λ and p are neither too large nor too small when NLS achieves the best performance. If λ is



(a) Jaffe (ACC) (b) Umist (ACC) (c) Yale (ACC) (d) Lenses (ACC) (e) Auto (ACC)

(f) Jaffe (NMI) (g) Umist (NMI) (h) Yale (NMI) (i) Lenses (NMI) (j) Auto (NMI)

Figure 3 Joint Effects of λ and p on our approach NLS



(a) Jaffe (b) Umist (c) Yale (d) Lenses (e) Auto

Figure 4 Effects of λ on NLS_1. **a–e** corresponds to five evaluated datasets, respectively

too large, the contribution of the noise modelling term ℓ_1-norm will be weakened, thereby leading the self-reconstruction and clustering into failure. If p is close to 2, then the $\ell_{p,2}$-norm tends to be closer to ℓ_2-norm, which implies the ability of NLS identifying noisy samples may significantly shrink; in contrast, small p tends to force NLS to "over-identify" noisy samples, which may lead to uncontrolled contamination of clean samples, thereby degrading clustering performance.

We also test the effects of λ on NLS_1. The experimental results corresponding to five datasets are listed in Figure 4a–e, from which we can see that as λ becomes larger, both ACC and NMI performance tends to be stable or slightly decreases. This fact implies that λ should be not be set either too large or too small. Similar to the explanation of NLS, λ greatly affects the noise modelling term ℓ_1-norm, thus it should be carefully chosen.

We summarize the optimal parameter settings for different variants of our approaches in Table 3.

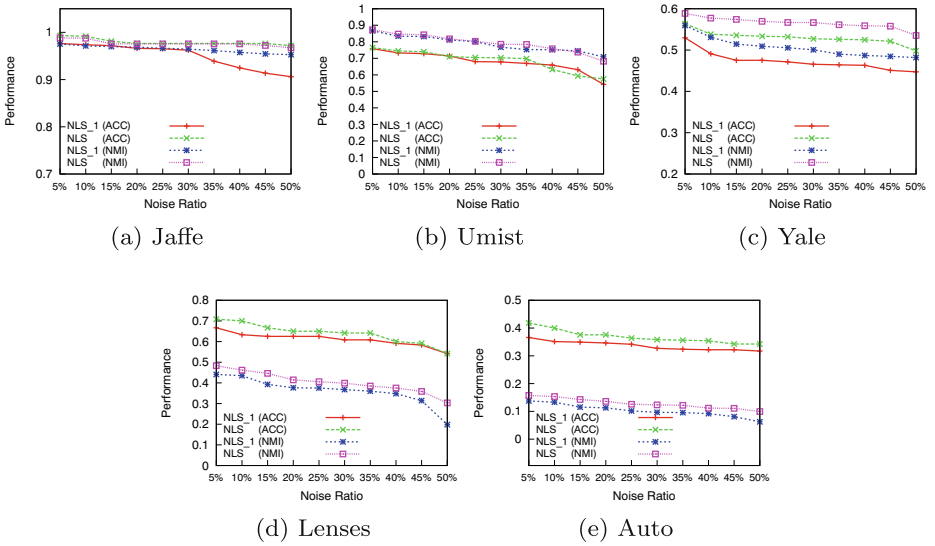
According to our observation, in most cases, the optimal parameters are neither too large nor too small, which suggests that it is possible to narrow down the search space of parameters. In practical task, one possible way of choosing these parameters is to combine cross-validation and shrinking the search space.

4.7 Robustness

In this subsection, we test the robustness of the proposed approaches NLS and NLS_1. To this end, we randomly add Gaussian noise to {5%, 10%, . . . , 50%} of data samples and use

Table 3 Optimal Parameters for NLS and NLS_1 on five datasets

Method	Dataset	ACC		NMI	
		λ	p	λ	p
NLS	Jaffe	10^2	0.75	10^2	1.75
	Umist	10^0	1.75	10^{-2}	0.5
	Yale	10^{-4}	0.5	10^{-4}	1
	Lenses	10^0	1.5	10^0	1.5
	Auto	10^2	0.5	10^4	0.5
NLS_1	Jaffe	10^4	–	10^0	–
	Umist	10^4	–	10^6	–
	Yale	10^{-6}	–	10^{-6}	–
	Lenses	10^4	–	10^2	–
	Auto	10^0	–	10^0	–

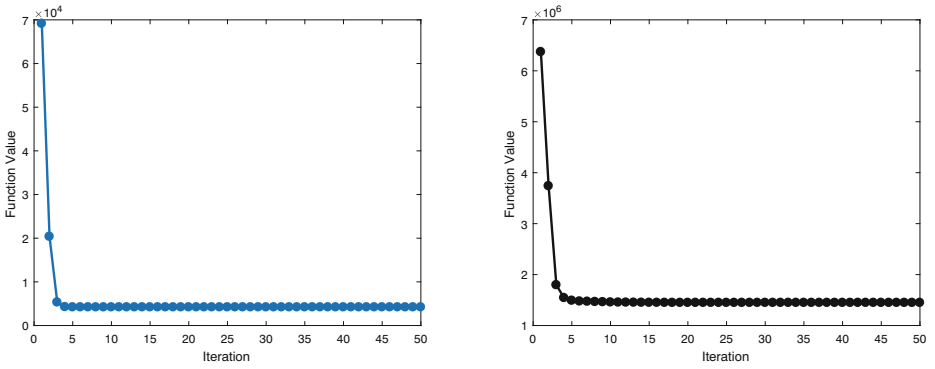


(a) Jaffe (b) Umist (c) Yale

(d) Lenses (e) Auto

Figure 5 Robustness of NLS and NLS_1 w.r.t. different ratios of noisy data samples in five datasets

the optimal parameter settings obtained in Section 4.6 for each algorithm. The experimental results on five datasets are listed in Figure 5. As we can see, in general, as the ratio increases from 5 to 50%, the performance (ACC and NMI) of NLS slightly decreases, which implies that NLS can be tolerant to noise and provide robust clustering ability. Nonetheless, compared to NLS, the stability of NLS_1 resisting the added noise tends to become worse as the proportion of noise goes up. For example, in Figure 5a, the ACC performance of NLS_1 drops significantly when the noise ratio is larger than 30%. The possible reason is that as the ratio increases, it becomes easier for ℓ_1-norm to propagate noise to all data samples and cause the whole self-reconstruction process more vulnerable. In contrast, $\ell_{p,2}$-norm is able to capture the global data structure and accurately identify noisy samples, which makes the final clustering stable and robust.



(a) Algorithm 1 (b) Algorithm 2

Figure 6 Convergence curves of two algorithms on Yale [2] dataset

4.8 Convergence study

As shown in Algorithms 1 and 2, the alternative updates of different variables make the objective function value decrease monotonously. An empirical study on convergence property is conducted (see Figure 6). it can be shown that the objective function value of our NLS descends dramatically within only 5 iterations and then converges.

5 Conclusion

In this work, we proposed a new spectral clustering method, termed *Non-negative Low-rank Self-reconstruction* (NLS), which jointly explores the nonnegative low-rank properties of data correlation and adaptively models the structural sparsity of data noise. We developed a self-reconstruction method by taking the nonnegativity and low-rank property of the data correlation matrix into consideration. Furthermore, we employed $\ell_{p,2}$-norm to model data noise, which conforms to the nature of data noise in real-world situation, as well as provides more adaptivity to different noise levels. We reported extensive experiments on various real-world datasets to show the superiority of the proposal. In the future, we intend to explore more reasonable properties of data to better characterize data correlation and enhance the performance of the current proposal.

Appendix A: Proof of Lemma 1

Proof Inspired by [34], we consider the following function

$$f(a) = pa^2 - 2a^p + (2-p), \tag{28}$$

where $p \in (0, 2)$. We expect to show that when $a > 0$, $f(a) \geq 0$. The first and second order derivatives of the function in (28) are $f'(a) = 2pa - 2pa^{p-1}$ and $f''(a) = 2p - 2p(p-1)a^{p-2}$, respectively. We can see that $a = 1$ is the only point that satisfies $f'(a) = 0$. Also, when $0 < a < 1$, $f'(a) < 0$ and when $a > 1$, $f'(a) > 0$. This means that $f(a)$ is monotonically decreasing when $0 < a < 1$ and monotonically increasing when $a > 1$. Moreover, we have $f''(1) = 2p(2-p) > 0$. Therefore, for $\forall a > 0$, $f(a) \geq f(1) = 0$.

Then, by substituting $a = \frac{\|\tilde{\varepsilon}_i\|}{\|\varepsilon_i\|}$ into (28), we obtain the conclusion

$$\begin{aligned}
& p\frac{\|\tilde{\varepsilon}_i\|^2}{\|\varepsilon_i\|^2} - 2\frac{\|\tilde{\varepsilon}_i\|^p}{\|\varepsilon_i\|^p} + (2-p) \geq 0, \\
\Leftrightarrow\ & p\|\tilde{\varepsilon}_i\|^2 - 2\|\tilde{\varepsilon}_i\|^p\|\varepsilon_i\|^{2-p} + (2-p)\|\varepsilon_i\|^2 \geq 0, \\
\Leftrightarrow\ & p\|\tilde{\varepsilon}_i\|^2\|\varepsilon_i\|^{p-2} - 2\|\tilde{\varepsilon}_i\|^p + (2-p)\|\varepsilon_i\|^p \geq 0, \\
\Leftrightarrow\ & 2\|\tilde{\varepsilon}_i\|^p - p\|\tilde{\varepsilon}_i\|^2\|\varepsilon_i\|^{p-2} \leq (2-p)\|\varepsilon_i\|^p, \\
\Leftrightarrow\ & \|\tilde{\varepsilon}_i\|^p - \frac{p\|\tilde{\varepsilon}_i\|^2}{2\|\varepsilon_i\|^{2-p}} \leq \|\varepsilon_i\|^p - \frac{p\|\varepsilon_i\|^p}{2\|\varepsilon_i\|^{2-p}}.
\end{aligned}$$

□

Appendix B: Proof of Theorem 1

Proof Denote $\mathcal{L}(\mathcal{E}) = \frac{1}{2}\left\|\mathcal{E} - \left(X - XW + \frac{P}{\alpha}\right)\right\|_F^2$ and $\lambda' = \lambda/\alpha$. Suppose $\tilde{\mathcal{E}}$ is the optimized solution of the alternative problem (18), then we obtain the following conclusion:

$$\begin{aligned}
&\mathcal{L}(\tilde{\mathcal{E}}) + \lambda' Tr(\tilde{\mathcal{E}}^T Z \tilde{\mathcal{E}}) \le \mathcal{L}(\mathcal{E}) + \lambda' Tr(\mathcal{E}^T Z \mathcal{E}) \\
\Rightarrow\ &\mathcal{L}(\tilde{\mathcal{E}}) + \lambda' \sum_{i=1}^{n} \frac{p\|\tilde{\varepsilon}_i\|^2}{2\|\varepsilon_i\|^{2-p}} \le \mathcal{L}(\mathcal{E}) + \lambda' \sum_{i=1}^{n} \frac{p\|\varepsilon_i\|^2}{2\|\varepsilon_i\|^{2-p}} \\
\Rightarrow\ &\mathcal{L}(\tilde{\mathcal{E}}) + \lambda' \sum_{i=1}^{n} \|\tilde{\varepsilon}_i\|_p^2 - \lambda' \left(\sum_{i=1}^{n} \|\tilde{\varepsilon}_i\|_p^2 - \sum_{i=1}^{n} \frac{p\|\tilde{\varepsilon}_i\|^2}{2\|\varepsilon_i\|^{2-p}} \right) \\
\le\ &\mathcal{L}(\mathcal{E}) + \lambda' \sum_{i=1}^{n} \|\varepsilon_i\|_p^2 - \lambda' \left(\sum_{i=1}^{n} \|\varepsilon_i\|_p^2 - \sum_{i=1}^{n} \frac{p\|\varepsilon_i\|^2}{2\|\varepsilon_i\|^{2-p}} \right).
\end{aligned}$$

Given the conclusion of Lemma 2, we finally arrive at

$$\mathcal{L}(\tilde{\mathcal{E}}) + \lambda' \sum_{i=1}^{n} \|\tilde{\varepsilon}_i\|_p^2 \le \mathcal{L}(\mathcal{E}) + \lambda' \sum_{i=1}^{n} \|\varepsilon_i\|_p^2.$$

Hence, the value of the objective function in (17) monotonically decreases in each iteration. □

References

1. Aha, D.W., Kibler, D.F., Albert, M.K.: Instance-based prediction of real-valued attributes, vol. 5 (1989)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE T PAMI **19**(7), 711–720 (1997)
3. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6), 1373–1396 (2003)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. JMLR **7**, 2399–2434 (2006)
5. Borjigin, S., Guo, C.: Perturbation analysis for the normalized Laplacian matrices in the multiway spectral clustering method. Sci. China Inf. Sci. **57**(11), 112102 (2014). https://doi.org/10.1007/s11432-014-5156-y
6. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J Optim **20**(4), 1956–1982 (2010)
7. Cendrowska, J.: Prism: an algorithm for inducing modular rules. International Journal of Man-Machine Studies **27**(4), 349–370 (1987)
8. Chen, L., Xu, D., Tsang, I.-H., Li, X.: Spectral embedded hashing for scalable image retrieval. IEEE TCYB **44**(7), 1180–1190 (2014)
9. De la Torre, F., Kanade, T.: Discriminative cluster analysis. In: ICML, pp. 241–248 (2006)
10. Deng, F., Lu, J., Wang, S., Pan, J., Zhang, L.: A distributed PDP model based on spectral clustering for improving evaluation performance. World Wide Web **22**(4), 1555–1576 (2019)
11. Ding, C., Li, T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In: ICML, pp. 521–528 (2007)
12. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR, pp. 2790–2797 (2009)
13. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV **59**(2), 167–181 (2004)
14. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. PR **41**(1), 176–190 (2008)
15. Gordon, S., Greenspan, H., Goldberger, J.: Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In: CVPR, pp. 370–377 (2003)

16. Graham, D., Allinson, N.: Characterising virtual eigensignatures for general purpose face recognition. NATO ASI series. Series F: Computer and System Sciences **163**, 446–456 (1998)
17. Hammouda, K., Kamel, M.: Efficient phrase-based document indexing for web document clustering. TKDE **16**(10), 1279–1296 (2004)
18. He, S., Wang, B., Wang, Z., Yang, Y., Shen, F., Huang, Z., Shen, H.T.: Bidirectional discrete matrix factorization hashing for image search. IEEE Transactions on Cybernetics (2019)
19. Hu, M., Yang, Y., Shen, F., Xie, N., Hong, R., Shen, H.T.: Collective reconstructive embeddings for cross-modal hashing. IEEE Trans. Image Process. **28**(6), 2770–2784 (2019)
20. Huang, Q., Wang, T., Tao, D., Li, X.: Biclustering learning of trading rules. IEEE TCYB **45**(10), 2287–2298 (2015)
21. Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
22. Jia, J., Yu, N., Hua, X.: Annotating personal albums via web mining. In: ACM Multimedia, pp. 459–468 (2008)
23. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. TKDE **16**(11), 1370–1386 (2004)
24. Li, J., Wang, J.: Real-time computerized annotation of pictures. TPAMI **30**(6), 985–1002 (2008)
25. Li, C., Chang, E., Garcia-Molina, H., Wiederhold, G.: Clustering for approximate similarity search in high-dimensional spaces. TKDE **14**(4), 792–808 (2002)
26. Li, J., Lu, K., Huang, Z., Zhu, L., Shen, H.T.: Heterogeneous domain adaptation through progressive alignment. IEEE TNNLS **30**(5), 1381–1391 (2018)
27. Li, J., Lu, K., Huang, Z., Zhu, L., Shen, H.T.: Transfer independently together: a generalized framework for domain adaptation. IEEE TCYB **49**(6), 2144–2155 (2018)
28. Li, C., Xu, Z., Quiao, C., Luo, T.: Hierarchical clustering driven by cognitive features. Sci. China Inf. Sci. **57**(1), 12109 (2014). https://doi.org/10.1007/s11432-013-4858-x
29. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE TPAMI **35**(1), 171–184 (2013)
30. Liu, G., Xu, H., Tang, J., Liu, Q., Yan, S.: A deterministic analysis for lrr. IEEE TPAMI **38**(3), 417–430 (2016)
31. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
32. Meng, D., Leung, Y., Xu, Z.: Detecting intrinsic loops underlying data manifold. IEEE Trans. Knowl. Data Eng. **25**(2), 337–347 (2013)
33. Nie, F., Xu, D., Tsang, I., Zhang, C.: Spectral embedded clustering. In: IJCAI, pp. 1181–1186 (2009)
34. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. In: NIPS, pp. 1813–1821 (2010)
35. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI **22**(8), 888–905 (2000)
36. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. JMLR **3**, 583–617 (2003)
37. Wang, X., Zhang, L., Li, X., Ma, W.: Annotating images by mining image search results. TPAMI **30**(11), 1919–1932 (2008)
38. Wang, F., Zhang, C., Li, T.: Clustering with local and global regularization. TKDE **21**(12), 1665–1678 (2009)
39. Wang, Z., Na, C., Ma, Z., Chen, S., Song, L., Yang, Y.: Exploring nonnegative and low-rank correlation for noise-resistant spectral clustering. In: Apweb-Waim, pp. 12–26 (2019)
40. Wu, M., Scholkopf, B.: A local learning approach for clustering. NIPS **19**, 1529–1536 (2007)
41. Xiao, Y., Zhu, Z., Zhao, Y., Wei, Y., Wei, S., Li, X.: Topographic nmf for data representation. IEEE TCYB **44**(10), 1762–1771 (2014)
42. Yang, Y., Xu, D., Nie, F., Yan, S., Zhuang, Y.: Image clustering using local discriminant models and global integration. TIP **19**(10), 2761–2773 (2010)
43. Yang, Y., Shen, H., Nie, F., Ji, R., Zhou, X.: Nonnegative spectral clustering with discriminative regularization. In: AAAI, pp. 555–560 (2011)
44. Yang, Y., Yang, Y., Shen, H.T., Zhang, Y., Du, X., Zhou, X.: Discriminative nonnegative spectral clustering with out-of-sample extension. IEEE TKDE **25**(8), 1760–1771 (2013)
45. Yang, Y., Ma, Z., Yang, Y., Nie, F., Shen, H.T.: Multitask spectral clustering by exploring intertask correlation. IEEE Trans. Cyber. **45**(5), 1083–1094 (2015)
46. Yang, Y., Shen, F., Shen, H.T., Li, H., Li, X.: Robust discrete spectral hashing for large-scale image semantic indexing. IEEE Transactions on Big Data **1**(4), 162–171 (2015)
47. Yang, Y., Shen, F., Huang, Z., Shen, H.T.: A unified framework for discrete spectral clustering. In: IJCAI, pp. 2273–2279 (2016)

48. Yang, Y., Shen, F., Huang, Z., Shen, H.T., Li, X.: Discrete nonnegative spectral clustering. IEEE Trans. Knowl. Data Eng. **29**(9), 1834–1845 (2017)
49. Ye, J., Zhao, Z., Liu, H.: Adaptive distance metric learning for clustering. In: CVPR, pp. 1–7 (2007)
50. Ye, J., Zhao, Z., Wu, M.: Discriminative k-means for clustering. NIPS **20**, 1649–1656 (2007)
51. Yu, S., Shi, J.: Multiclass Spectral Clustering. In: ICCV, pp. 313–319 (2003)
52. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. NIPS **17**, 1601–1608 (2004)
53. Zhang, Z., Liu, L., Shen, F., Shen, H.T., Shao, L.: Binary multi-view clustering. IEEE TPAMI **41**(7), 1774–1782 (July 2019)

Affiliations

Zheng Wang[1,2] · Lin Zuo[1] · Jing Ma[3] · Si Chen[4] · Jingjing Li[1] · Zhao Kang[1] · Lei Zhang[5]

Zheng Wang
zh_wang@hotmail.com

Jing Ma
majing@se.cuhk.edu.hk

Si Chen
chensi@cetc.com.cn

Jingjing Li
lijin117@yeah.net

Zhao Kang
zkang@uestc.edu.cn

Lei Zhang
leizhang@cqu.edu.cn

1 School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

2 School of Computer Science and Engineering, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan, Guangdong Province, China

3 The Chinese University of Hongkong, Hong Kong, China

4 Information Science Academy, China Electronics Technology Group Corporation, Beijing, China

5 School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China