

Robust subspace clustering based on non-convex low-rank approximation and adaptive kernel

Xuqian Xue^a, Xiaoqian Zhang^{b,c,*}, Xinghua Feng^b, Huaijiang Sun^c, Wei Chen^a, Zhigui Liu^{a,b}

^a School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, PR China

^b School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, PR China

^c School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Article history:

Received 18 September 2019

Revised 16 October 2019

Accepted 25 October 2019

Available online 6 November 2019

Keywords:

Subspace clustering

Low-rank approximation

Adaptive kernel

Correntropy

ABSTRACT

As a relatively advanced method, the low-rank kernel space clustering method shows good performance in dealing with nonlinear structure of high-dimensional data. Unfortunately, this method is sensitive to large corruptions and doesn't balance the contribution of all singular values. To solve the above problems, the low-rank kernel method is modified, and a robust subspace clustering method (LAKRSC) based on non-convex low-rank approximation and adaptive kernel is proposed. In our model, the weighted Schatten p -norm is introduced to balance the importance of different singular values, which can more accurately approximate the rank function and be more flexible in practical applications. Therefore, applying weighted Schatten p -norm to adaptive kernel can approximate the original low rank hypothesis better when the data is mapped into the feature space. In addition, our model uses correntropy to handle complex noise which enhances the robustness of the model. A new algorithm HQ&ADMM, combined by Half-Quadratic technique (HQ) and ADMM, is studied to solve our model. Experiments on four real-world datasets show that the clustering performance of LAKRSC is significantly better than that of several more advanced methods.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

The task objectives of subspace clustering are as follows, finding suitable low-dimensional subspaces for each set of data points, and clustering multiple sets of data into multiple subspaces. Due to its excellent performance, subspace clustering techniques are widely used and rapidly developed in the fields of pattern recognition, signal processing and so on [3,39]. Existing methods are mainly divided into four categories [4]. Among them, spectral clustering-based methods have been very popular recently. Because the optimization problems of these methods are solved easily, and they achieve competitive results in real-world applications, such as data classification [46], face clustering [30], image recognition [40], motion segmentation [11].

The theoretical basis of spectral clustering is Graph Theory [1]. Usually, this kind of methods are divided into two stages in subspace clustering: the first step is to build the affinity matrix between the data samples; the second step is to apply

* Corresponding author.

E-mail addresses: braveq313@163.com (X. Xue), zhxq0528@163.com (X. Zhang), fengxh@swust.edu.cn (X. Feng), sunhuaijiang@njnu.edu.cn (H. Sun), chenwei9307@163.com (W. Chen), liuzhigui@swust.edu.cn (Z. Liu).

a specialized algorithm to the constructed affinity matrix and get the final results. Therefore, a major link in the clustering process of spectral clustering-based algorithms is to construct a good affinity matrix. Inspired by this, based on L_1 -graph, sparse subspace clustering algorithm (SSC) [5] is proposed, which makes full use of the characteristics of self-expression and sparsity of data. However, SSC ignores the global characteristics and only considers the local characteristics. Based on this, low-rank representation method (LRR) [16] is proposed, which is a two-dimensional sparse method that represents data vectors as linear combinations of other vectors. It is worth mentioning that the LRR solves NP-hard problem caused by the rank function with the applying of nuclear norm to get the lowest rank of matrix. For the problem of rank minimization, Xue et al. [35] studied the non-convex tensor rank minimization method and proposed an optimization scheme, which provides a new idea for the rank minimization problem. These methods show good performance in dealing with the linear structure of data. Unfortunately, in practice, many datasets are more suitable for modeling by nonlinear models.

Some scholars have studied subspace clustering methods of nonlinear structural data. Deng et al. [2] proposed a low-rank local tangent space embedded subspace clustering model. In addition, some kernel subspace clustering methods are designed to implement nonlinear structural data modeling. In particular, Patel et al. [27] proposed a subspace clustering method that implemented the nonlinear extension of SSC by using the Mercer kernel. Similarly, by extending LRR nonlinearly, Xiao et al. [33] proposed a robust kernel LRR method. Yin et al. [37] used a kernel to embed the Riemannian manifold into a high dimension feature space, and named this kernel subspace clustering method as KSSCR. On the basis of self-representation of subspace, Ji et al. [10] proposed a robust method (LRKSC) by adding low-rank constraints to the mapping kernel. The low-rank kernel mapping (in an unsupervised manner) makes the coefficient matrix low-rank. In [19], the Multiple Kernel Learning (MKL) is applied to solving nonlinear structural data modeling problem, which makes the algorithms to select or combine kernels suitable for the datasets from a set of kernels. However, if the original dataset itself is corrupted in some cases, the induced kernel may not be optimal. Therefore, the performance of the MKL methods will be worse than the single kernel methods. Although the above methods can handle nonlinear structure, the importance of different rank components of matrix is not considered in the kernel mapping process. Finally, recently, subspace clustering methods based on deep learning [9,42] have attracted some scholars' attention. In general, the deep learning based methods are better than the ordinary subspace clustering methods, but the deep learning-based methods have higher requirements on the hardware of the computer, which makes them not universally used, and the focus of this paper differs from the research focus of these methods.

Instead of processing singular values equally, handling each rank component differently can obtain the low-rank of the matrix more accurately, and thereby improve the clustering performance of the model. In [41], a novel matrix completion method (TNNR) is proposed, which only minimized the smallest r singular values. However, r is difficult to estimate. A targeted approach (PSM) is proposed [26]. It is a pity that PSM doesn't consider the different rank components. To make up for this deficiency, Weighted Nuclear Norm Minimization method (WNNM) [6] assigned suitable weights to each singular values. In [17], to improve the performance of signal recovery, a Schatten p -norm based method is proposed. Many current research results verify the superiority of the Schatten p -norm [44,45], but it still has some shortcomings. For example, the Schatten p -norm gives them the same importance when dealing with the rank components. The flexibility of these methods is clearly unsatisfactory. Motivated by WNNM, Xie et al. [34] introduced the weighting method in the model (WSNM) that can fully exploit the effects of all rank components. In this model, the optimization is mainly for coefficient matrix, but noise in data is not well modeled. On the basis of WSNM, Zhang et al. [43] used ℓ_q norm to model various noise and proposed a robust subspace clustering method WSPQ. However, when dealing with nonlinear structural data, clustering performance of the method is affected.

Since noise can disturb the subspace structure and affect the subspace clustering performance, how to deal with noise is also a major challenge for subspace clustering. In [5], data items used ℓ_1 norm to process arbitrary sparse outliers. In [15], the sample-specific corruption was handled by using the $\ell_{1,2}$ norm. In the LSR [21], Frobenius norm was used to model Gaussian noise. However, data terms for these methods all used relatively simple specifications to describe noise, which may cause models to generate inaccurate affinity matrices resulting in lower clustering accuracy. In practice, the complex noise caused by errors in actual problems usually comes from many types of corruption and its statistical structure is extremely complex [8]. For this trouble, some scholars proposed to use subspace clustering models based on correntropy to resist complex noise. In [7,20], correntropy is introduced to improve the robustness of model. Above methods inspire us to use correntropy to deal with complex noise. Unfortunately, these models do not pay enough attention to the nonlinear structural data. In [36], a robust subspace clustering model is proposed, which combines correntropy and multi-kernel learning. Xia et al. [32] learned a nonlinear structural data modeling method for human motion capture data, using correntropy to measure reconstruction errors. However, this method still produces convex relaxation in the modeling process, and doesn't consider the importance of different rank components.

In a word, so far, there is not a non-convex low-rank subspace clustering method that takes into consideration the importance of different rank components in the kernel feature mapping process. If this problem is not well solved, it is difficult to get the expected clustering results. Therefore, we have optimized the low rank kernel subspace clustering method (LRKSC) [10], including the following two points: (1) the low rank term of the LRKSC model has been changed from the nuclear norm to the weighted Schatten p -norm, which can better approximate the original low rank hypothesis; (2) the data item of the LRKSC model has been changed from ℓ_1 norm to correntropy, which can better handle complex noise and large corruption in the data. Based on these optimizations, we propose a robust subspace clustering method (LAKRSC), which effectively combines non-convex low rank approximation and adaptive kernel and correntropy. We have extensively

evaluated our model on four image clustering datasets, and obtain good clustering results. Our contributions are summarized below:

- To fully exploit the effects of all rank components of the data, the weighted Schatten p -norm is applied to the adaptive kernel. And a robust subspace clustering method is proposed.
- To improve the robustness of LAKRSC, this paper uses the most advanced correntropy to process non-Gaussian noise and large corruption.
- To improve the efficiency of solving the proposed model, we design a new algorithm HQ&ADMM by combing HQ and ADMM.

The remaining of this paper: (1) a brief introduction of the related work is described in [Section 2](#); (2) in [Section 3](#), the model LARSC and its detailed optimization are proposed; (3) [Section 4](#) contains relevant experiments and results analysis of LARSC; (4) [Section 5](#) describes summary of this work.

2. Related work

Next, some works related to our model are briefly introduced, including weighted Schatten p -norm, Low-rank kernel subspace self-expressiveness, correntropy.

2.1. Weighted Schatten p -norm

LRR replaces the rank function with the kernel norm to solve the rank of the matrix $\mathbf{Y}$. However, the kernel norm is a convex agent of the rank function, which leads to a suboptimal solution. The weighted Schatten p -norm is a better non-convex agent for low-rank approximation, defined as

$$\|\mathbf{Y}\|_{\mathbf{w},S_p} = \left(\sum_{i=1}^n w_i \sigma_i^p \right)^{\frac{1}{p}} \quad (0 < p \leq 1) \quad (1)$$

where σ_i is the i th largest singular value of $\mathbf{Y}$, and $\mathbf{w} = [w_1, w_2, \dots, w_n]$ is the weight vector. The [Eq. \(1\)](#) can be rewritten as

$$\|\mathbf{Y}\|_{\mathbf{w},S_p}^p = \sum_{i=1}^n w_i \sigma_i^p \quad (0 < p \leq 1) \quad (2)$$

Obviously, when $p = 1$ and $w_i = 1$, the [Eq. \(2\)](#) is equivalent to the nuclear norm, while it is the rank function when $p = 0$. Further, when $0 < p < 1$ and $w_i = 1$, the weighted Schatten p -norm is closer to rank function than nuclear norm. Considering that the main properties of matrix are determined by the large singular values of matrix, assigning different weights to different singular values is a more reasonable way to approximate the rank of matrix. Therefore, the weighted Schatten p -norm minimization is more able to maintain the low rank property of matrix.

2.2. Low-rank kernel subspace self-expressiveness

The SSC algorithm takes advantage of the “self-expression” of data, i.e., data points in the same subspace can be expressed linearly with each other. Given a data matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \mathbb{R}^{D \times N}$, to be more accurate, each data sample point $\mathbf{y}_i \in \cup_{l=1}^n \mathbf{S}_l$ can be expressed as

$$\mathbf{y}_i = \mathbf{Y}\mathbf{G}, \mathbf{G}_{ii} = 0 \quad (3)$$

where $\mathbf{G}$ is the self-expressive coefficient matrix. The data matrix $\mathbf{Y}$ is a self-representation dictionary. The optimal solution for $\mathbf{G}$ can be solved by

$$\min_{\mathbf{G}} \|\mathbf{G}\|_1 \quad \text{s.t. } \mathbf{Y} = \mathbf{Y}\mathbf{G}, \mathbf{G}_{ii} = 0 \quad (4)$$

where $\|\mathbf{G}\|_1$ is the ℓ_1 norm of matrix $\mathbf{G}$, i.e., $\|\mathbf{G}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |g_{ij}|$. Unfortunately, [\(4\)](#) is only applicable to linear (or affine) subspaces. Using kernel trick [\[28\]](#), nonlinear structural data can be mapped to linear feature space (see [Fig. 1](#)). The nonlinear extension of SSC is given by

$$\min_{\mathbf{K}, \mathbf{G}} \|\phi(\mathbf{Y})\|_* + \lambda \|\mathbf{G}\|_1 \quad \text{s.t. } \phi(\mathbf{Y}) = \phi(\mathbf{Y})\mathbf{G}, \mathbf{G}_{ii} = 0 \quad (5)$$

where $\mathbf{K} = \phi(\mathbf{Y})^T \phi(\mathbf{Y})$ is a Gram matrix, $\phi(\mathbf{Y}) = [\phi(\mathbf{y}_1), \phi(\mathbf{y}_2), \dots, \phi(\mathbf{y}_n)]$ represents the mapped feature space, and λ is a trade-off parameter. $\|\phi(\mathbf{Y})\|_*$ is a convex surrogate of $\text{rank}(\phi(\mathbf{Y}))$, which ensures that the mapped data is contained in multiple linear subspaces. However, the above low-rank kernel mapping doesn't take into account the importance of different rank components.

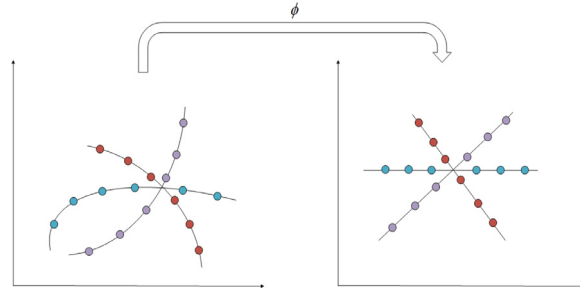


Fig. 1. The subspace structure retains the feature map: the nonlinear subspace maps to the linear subspace in the high dimensional Hilbert space.

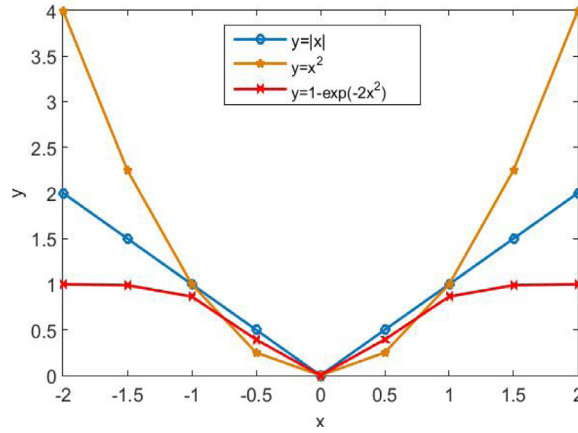


Fig. 2. Comparison of different loss functions.

2.3. Correntropy

Correntropy was first defined in [18] to measure the similarity of A and B . It is expressed in a mathematical formula as

$$V_{\sigma}(A, B) = E[k_{\sigma}(A - B)] \quad (6)$$

where $E[\bullet]$ stands for the mathematical expectation, $k(\bullet)$ represents kernel function, and σ is the width of $k(\bullet)$. In practice, usually only limited amount of data samples $\{(a_i, b_i)\}_{i=1}^N$ can be obtained. So we use

$$\hat{V}_{N, \sigma}(A, B) = \frac{1}{N} \sum_{i=1}^N k_{\sigma}(a_i - b_i) \quad (7)$$

to estimate correntropy of A and B . The large value indicates that the two variables are more similar. Conversely, the two variables have low similarity. We select Gaussian kernel $k_{\sigma}(a_i - b_i) = \exp(-\frac{\|a_i - b_i\|_2^2}{2\sigma^2})$ as the kernel function. As an extension of the concept of correntropy criteria, Correntropy Induced Metric (CIM) [18] measures the similarity of A and B as

$$\text{CIM}(A, B) = \{k(0) - V(A, B)\}^{1/2} = \left(g(0) - \frac{1}{N} \sum_{i=1}^N g(A_i - B_i)\right)^{1/2} \quad (8)$$

We do an experiment on the comparison of different loss functions, which can help readers understand better that correntropy can deal with large corruption. Here, the different loss functions include the absolute error, mean squared error (MSE) and CIM. The experimental result is shown in Fig. 2. We can see that compared to MSE and the absolute error, CIM can handle large errors better (values close to 1). We will use correntropy in the proposed model.

3. Robust subspace clustering based on non-convex low-rank approximation and adaptive kernel

In this section, a robust subspace clustering model (LAKRSC) is proposed, which joint non-convex low-rank approximation and adaptive kernel. Firstly, we formulate the objective function of LAKRSC. Second, an effective optimization algorithm (HQ&ADMM) is proposed.

3.1. Model of LAKRSC

Inspired by LRKSC [10], as a non-convex method for solving $\text{rank}(\phi(\mathbf{Y}))$, the weighted Schatten p -norm is used to constrain adaptive kernel, which can fully exploit the role of rank components. The detailed optimization problem is described as below

$$\min_{\mathbf{K}, \mathbf{G}} \|\phi(\mathbf{Y})\|_{\mathbf{w}, S_p}^p + \lambda \|\mathbf{G}\|_1 \quad \text{s.t.} \quad \phi(\mathbf{Y}) = \phi(\mathbf{Y})\mathbf{G}, \quad \text{diag}(\mathbf{G}) = 0 \quad (9)$$

where λ is a tradeoff parameter and $\mathbf{K} = \phi(\mathbf{Y})^T \phi(\mathbf{Y})$ is an unknown kernel. Therefore, kernel trick is applied to determine the specific form of $\phi(\mathbf{Y})$. Since data points are generally contaminated by corruption or occlusion rather than clean, $\phi(\mathbf{Y}) = \phi(\mathbf{Y})\mathbf{G}$ is joined to (9) in the form of $\|\phi(\mathbf{Y}) - \phi(\mathbf{Y})\mathbf{G}\|_F^2 = \text{tr}((\mathbf{I} - 2\mathbf{G} + \mathbf{G}\mathbf{G}^T)\phi(\mathbf{Y})^T \phi(\mathbf{Y}))$. Besides, the process of optimizing (9) contains $\|\phi(\mathbf{Y})\|_{\mathbf{w}, S_p}^p$. So it depends on $\phi(\mathbf{Y})$. In order to optimize the objective function conveniently, one suitable solution for solving $\phi(\mathbf{Y})$ is to use reparameterization. According to the characteristics of $\mathbf{K}$, we can factorize it as $\mathbf{K} = \mathbf{Z}^T \mathbf{Z}$, where $\mathbf{Z}$ is a square matrix. Therefore, we are easy to understand that $\|\phi(\mathbf{Y})\|_{\mathbf{w}, S_p}^p = \|\mathbf{Z}\|_{\mathbf{w}, S_p}^p$. Adding affine constraint, (9) can be changed to

$$\min_{\mathbf{Z}, \mathbf{G}} \|\mathbf{Z}\|_{\mathbf{w}, S_p}^p + \lambda_1 \|\mathbf{G}\|_1 + \frac{\lambda_2}{2} \|\phi(\mathbf{Y}) - \phi(\mathbf{Y})\mathbf{G}\|_F^2 + \frac{\lambda_3}{2} \|\mathbf{K} - \mathbf{Z}^T \mathbf{Z}\|_F^2 \quad \text{s.t.} \quad \mathbf{1}^T \mathbf{G} = \mathbf{1}^T, \mathbf{G}_{ii} = 0 \quad (10)$$

where $\phi(\mathbf{Y})^T \phi(\mathbf{Y}) = \mathbf{Z}^T \mathbf{Z}$, and the elements of column vector $\mathbf{1}$ are all one. As show in [5], $\mathbf{H}$ as an auxiliary variable is designed to deal with diagonal constraints on $\mathbf{G}$, and $\mathbf{H} = \mathbf{G} - \text{diag}(\mathbf{G})$. Introducing the auxiliary variable $\mathbf{H}$, the new optimization problem is shown below

$$\min_{\mathbf{H}, \mathbf{Z}, \mathbf{G}} \|\mathbf{Z}\|_{\mathbf{w}, S_p}^p + \lambda_1 \|\mathbf{G}\|_1 + \frac{\lambda_2}{2} \text{tr}((\mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H}^T)\mathbf{Z}^T \mathbf{Z}) + \frac{\lambda_3}{2} \|\mathbf{K} - \mathbf{Z}^T \mathbf{Z}\|_F^2 \quad \text{s.t.} \quad \mathbf{1}^T \mathbf{G} = \mathbf{1}^T, \mathbf{H} = \mathbf{G} - \text{diag}(\mathbf{G}) \quad (11)$$

The weighting strategy in (11) is designed to protect the major low-rank structure of $\mathbf{Z}$. In order to balance the role of all singular values, when assigning weights, larger singular values should have smaller weights. So we can let

$$w_i = C/\delta_i(\mathbf{Z}) \quad (12)$$

where C is a constant. Considering the effect of the size of the data matrix $\mathbf{Y} \in \mathbb{R}^{D \times N}$, we rewrite (12) as

$$w_i = C\sqrt{DN}/[\delta_i(\mathbf{Z}) + \varepsilon] \quad (13)$$

where ε is set to 10^{-6} .

3.2. General solution of LAKRSC

Due to the bilinear terms in (11), the optimization of (11) is non-convex. Recently, some scholars have used the ADMM to solve the non-convex problem [13], especially the bilinear problems. Hence, we propose to solve the above optimization via the ADMM. Convergence analysis of its non-convex problem was proposed in [14]. In order to derive the ADMM solution of (11), we first get its augmented Lagrangian function

$$\begin{aligned} L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = & \|\mathbf{Z}\|_{\mathbf{w}, S_p}^p + \lambda_1 \|\mathbf{G}\|_1 + \frac{\lambda_2}{2} \text{tr}((\mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H}^T)\mathbf{Z}^T \mathbf{Z}) + \frac{\lambda_3}{2} \|\mathbf{K} - \mathbf{Z}^T \mathbf{Z}\|_F^2 \\ & + \text{tr}(\mathbf{\Lambda}_1^T (\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G}))) + \text{tr}(\mathbf{\Lambda}_2^T (\mathbf{1}^T \mathbf{G} - \mathbf{1}^T)) + \frac{\rho}{2} (\|\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G})\|_F^2) \\ & + \|\mathbf{1}^T \mathbf{G} - \mathbf{1}^T\|_2^2 \end{aligned} \quad (14)$$

where $\mathbf{\Lambda}_1 \in \mathbb{R}^{N \times N}$, $\mathbf{\Lambda}_2 \in \mathbb{R}^{1 \times N}$ are the Lagrange multipliers, $\rho \geq 0$ is a tradeoff parameter. The variables $\mathbf{Z}$, $\mathbf{G}$, $\mathbf{H}$ can be updated alternately by fixing others. The updating rules are as follows

$$\begin{aligned} \mathbf{Z}^{k+1} &= \arg \min_{\mathbf{Z}} L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \\ \mathbf{G}^{k+1} &= \arg \min_{\mathbf{G}} L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \\ \mathbf{H}^{k+1} &= \arg \min_{\mathbf{H}} L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \end{aligned} \quad (15)$$

and then updating the Lagrange multipliers $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$.

(1) Updating $\mathbf{G}$

To update $\mathbf{G}$, we solve the following problem

$$\min_{\mathbf{G}} \frac{\lambda_1}{\rho} \|\mathbf{G}\|_1 + \frac{1}{2} \left\| \mathbf{G} - \text{diag}(\mathbf{G}) - \left(\mathbf{H} + \frac{\mathbf{\Lambda}_1}{\rho} \right) \right\|_F^2 \quad (16)$$

The problem has a closed-form solution as

$$\mathbf{G}^{k+1} = \mathbf{v} - \text{diag}(\mathbf{v}) \quad (17)$$

where $\mathbf{v} = \Upsilon_{\frac{\lambda_1}{\rho}}\left(\mathbf{H} + \frac{\mathbf{A}_1}{\rho}\right)$. Υ is a soft-threshold operator by element, and $\Upsilon_{\tau}(\theta) = \text{sign}(\theta) \cdot \max(|\theta| - \tau, 0)$.

(2) Updating $\mathbf{H}$

It can be solved by

$$\min_{\mathbf{H}} \frac{\lambda_2}{2} \text{tr}((-2\mathbf{H} - \mathbf{H}\mathbf{H}^T)\mathbf{Z}^T\mathbf{Z}) + \text{tr}(\mathbf{A}_1^T\mathbf{H}) + \text{tr}(\mathbf{A}_2^T\mathbf{1}^T\mathbf{G}) + \frac{\rho}{2} \left(\|\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G})\|_F^2 + \|\mathbf{1}^T\mathbf{G} - \mathbf{1}^T\|_2^2 \right) \quad (18)$$

We can get

$$\mathbf{H}^{k+1} = (\lambda_2\mathbf{Z}^T\mathbf{Z} + \rho(\mathbf{I} + \mathbf{1}\mathbf{1}^T))^{-1} (\lambda_2\mathbf{Z}^T\mathbf{Z} - \mathbf{A}_1 - \mathbf{1}\mathbf{A}_2 + \rho(\mathbf{G} - \text{diag}(\mathbf{G})) + \mathbf{1}\mathbf{1}^T) \quad (19)$$

where $\mathbf{1}$ is an identity matrix.

(3) Updating $\mathbf{Z}$

The optimization of $\mathbf{Z}$ can be solved by

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{w, S_p}^p + \frac{\lambda_3}{2} \|\bar{\mathbf{K}} - \mathbf{Z}^T\mathbf{Z}\|_F^2 \quad (20)$$

where $\bar{\mathbf{K}} = \mathbf{K} - \frac{\lambda_2}{2\lambda_3}(\mathbf{I} - 2\mathbf{H}^T + \mathbf{H}\mathbf{H}^T)$.

The optimization problem of (12) is a major contribution of this paper. In order to clearly describe its solution, we first give the **Theorem**.

Theorem (Von-Neumann [22]). Given matrices $\mathbf{S} \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\sigma(\mathbf{S}) = [\sigma_1(\mathbf{S}), \dots, \sigma_r(\mathbf{S})]^T$ and $\sigma(\mathbf{Q}) = [\sigma_1(\mathbf{Q}), \dots, \sigma_r(\mathbf{Q})]^T$ are the singular values of $\mathbf{S}$ and $\mathbf{Q}$, respectively, where $r = \min(m, n)$, then $\text{tr}(\mathbf{S}^T\mathbf{Q}) \leq \text{tr}(\sigma(\mathbf{S})^T\sigma(\mathbf{Q}))$. The condition that the equation is established is to find $\mathbf{U}$ and $\mathbf{V}$ that satisfy the following equation

$$\mathbf{S} = \mathbf{U}\Sigma_S\mathbf{V}^T, \text{ and } \mathbf{Q} = \mathbf{U}\Sigma_Q\mathbf{V}^T \quad (21)$$

where Σ_S and Σ_Q represent ordered singular value matrices.

Lemma 1. Let matrices $\mathbf{S}$ and $\mathbf{Q}$ have the SVD of $\mathbf{S} = \mathbf{M}\mathbf{\Gamma}\mathbf{L}^T$ and $\mathbf{Q} = \mathbf{U}\Sigma\mathbf{V}^T$, respectively, where both $\mathbf{\Gamma} = \text{diag}(\delta_1, \dots, \delta_N)$ and $\Sigma = \text{diag}(\gamma_1, \dots, \gamma_N)$ are diagonal matrices. We can get the following problem

$$\min_{\mathbf{S}} \|\mathbf{S}\|_{w, S_p}^p + \frac{\mu}{2} \|\mathbf{Q} - \mathbf{S}^T\mathbf{S}\|_F^2 \quad (22)$$

Its closed-form solution is

$$\mathbf{S}^* = \mathbf{\Gamma}^*\mathbf{V}^T \quad (23)$$

where $\mathbf{\Gamma}^* = \text{diag}(\delta_1^*, \dots, \delta_N^*)$, with $\delta_i^* = \arg \min_{\delta_i} \frac{\rho}{2}(\gamma_i - \delta_i^2)^2 + w_i\delta_i^p$ and $\delta_i \in \{x \in \mathbb{R}_+ \mid x^3 - \delta_i x + \frac{w_i p}{2\rho} x^{p-1} = 0\} \cup \{0\}$.

From Lemma 1, the closed-form solution of (20) is

$$\mathbf{Z}^{k+1} = \mathbf{\Gamma}^*\mathbf{V}^T \quad (24)$$

where $\mathbf{\Gamma}^*$ and $\mathbf{V}^T$ are obtained from $\bar{\mathbf{K}}$.

An extend proof to the Lemma 1 is shown in the Appendix.

3.3. Handling gross corruptions of LAKRSC

In this section, correntropy, as a robust measurement method for large corruptions, is introduced into our model. The detailed optimization problem can be formulated as

$$\min_{\mathbf{Z}, \mathbf{G}, \mathbf{E}} \|\mathbf{Z}\|_{w, S_p}^p + \lambda_1 \|\mathbf{G}\|_1 + \frac{\lambda_2}{2} \text{tr}((\mathbf{I} - 2\mathbf{H} + \mathbf{H}^T\mathbf{H})\mathbf{Z}^T\mathbf{Z}) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij})$$

$$\text{s.t } \mathbf{H} = \mathbf{G} - \text{diag}(\mathbf{G}), \mathbf{1}^T\mathbf{G} = \mathbf{1}^T, \mathbf{K} = \mathbf{Z}^T\mathbf{Z} + \mathbf{E} \quad (25)$$

where $\mathbf{E}$ models noise and corruptions in data, $\varphi(\mathbf{E}_{ij}) = 1 - \exp(-\frac{\mathbf{E}_{ij}^2}{2\sigma^2})$, σ is the size of the Gaussian kernel. For the optimization problem of the objective function, we add the half-quadratic (HQ) optimization method to solve the data item based on the ADMM framework. This new algorithm is named HQ&ADMM. We derive the formula augmented Lagrangian function as

$$L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{E}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3) = \|\mathbf{Z}\|_{w, S_p}^p + \lambda_1 \|\mathbf{G}\|_1 + \frac{\lambda_2}{2} \text{tr}((\mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H}^T)\mathbf{Z}^T\mathbf{Z}) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij})$$

$$+ \text{tr}(\mathbf{A}_1^T(\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G}))) + \text{tr}(\mathbf{A}_2^T(\mathbf{1}^T\mathbf{G} - \mathbf{1}^T)) + \text{tr}(\mathbf{A}_3^T(\mathbf{K} - \mathbf{Z}^T\mathbf{Z} - \mathbf{E}))$$

$$+ \frac{\rho}{2} (\|\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G})\|_F^2 + \|\mathbf{1}^T\mathbf{G} - \mathbf{1}^T\|_2^2 + \|\mathbf{K} - \mathbf{Z}^T\mathbf{Z} - \mathbf{E}\|_F^2) \quad (26)$$

where $\Lambda_1 \in \mathbb{R}^{N \times N}$, $\Lambda_2 \in \mathbb{R}^{1 \times N}$ and $\Lambda_3 \in \mathbb{R}^{N \times N}$. The variables can be updated alternately by minimizing $L(\mathbf{Z}, \mathbf{G}, \mathbf{H}, \mathbf{E}, \Lambda_1, \Lambda_2, \Lambda_3)$.

(1) Updating $\mathbf{G}$ and $\mathbf{H}$

The update of $\mathbf{G}$ and $\mathbf{H}$ are identical to (16) and (18) by comparing the two Eq. (14) and (26). Correspondingly, the solutions are also the same as that in (17) and (19).

(2) Updating $\mathbf{Z}$

Except for $\bar{\mathbf{K}}$, the subproblem to update $\mathbf{Z}$ is similar to Eq. (24). The $\bar{\mathbf{K}}$ is now defined as

$$\bar{\mathbf{K}} = \mathbf{K} - \frac{1}{\rho} \left(\frac{\lambda_2}{2} (\mathbf{I} - 2\mathbf{H}^T + \mathbf{H}\mathbf{H}^T) - \Lambda_3 \right) - \mathbf{E} \quad (27)$$

(3) Updating $\mathbf{E}$

$$\min_{\mathbf{E}} \frac{\lambda_3}{\rho} \sum_{ij} \varphi(\mathbf{E}_{ij}) + \frac{1}{2} \left\| \mathbf{E} - \left(\mathbf{K} - \mathbf{Z}^T \mathbf{Z} + \frac{\Lambda_3}{\rho} \right) \right\|_F^2 \quad (28)$$

This problem is difficult to be optimized directly. Refer to Nikolova and Ng [25], this paper use the HQ technique to optimize this function. Nikolova and Ng [25] showed the calculation speed of the HQ minimization problem obviously faster than the gradient based method. We update $\mathbf{E}_{ij}$ with HQ theory as follows

$$\varphi(\mathbf{E}_{ij}) = \min_{\mathbf{W}_{ij} \in \mathbb{R}} \frac{1}{2} \mathbf{W}_{ij} \mathbf{E}_{ij}^2 + \psi(\mathbf{W}_{ij}) \quad (29)$$

where $\psi(\bullet)$ is dual potential function of $\varphi(\bullet)$, and $\mathbf{W}_{ij}$ is a auxiliary variable. So, (28) can be converted to the augmentation function as follows

$$\min_{\mathbf{E}, \mathbf{W}} \frac{1}{2} \left\| \mathbf{E} - \left(\mathbf{K}_G - \mathbf{Z}^T \mathbf{Z} + \frac{\Lambda_3}{\rho} \right) \right\|_F^2 + \frac{\lambda_3}{2\rho} \left\| \mathbf{W}^{\frac{1}{2}} \otimes \mathbf{E} \right\|_F^2 + \sum_{ij} \Psi(\mathbf{W}_{ij}) \quad (30)$$

where $\otimes$ denotes the Hadamard product. Supposing the current solutions are $(\mathbf{E}^t, \mathbf{W}^t)$, the results of the next iteration is calculated as

a when $\mathbf{E}$ is fixed, $\mathbf{W}_{ij}$ is solved as

$$\mathbf{W}_{ij}^{t+1} = \delta(\mathbf{E}_{ij}^t) \quad (31)$$

where $\delta(x) = \exp(-x^2/\sigma^2)$.

b when $\mathbf{W}$ is fixed, eliminating the last part of (30), (30) is downgraded as

$$\begin{aligned} \mathbf{E}^{t+1} &= \arg \min_{\mathbf{E}} \left\| \mathbf{E} - \left(\mathbf{K} - \mathbf{Z}^T \mathbf{Z} + \frac{\Lambda_3}{\mu} \right) \right\|_F^2 + \frac{\lambda_3}{\mu} \left\| (\mathbf{W}^{t+1})^{\frac{1}{2}} \otimes \mathbf{E} \right\|_F^2 \\ &= \left(\mathbf{K} - \mathbf{Z}^T \mathbf{Z} + \frac{\Lambda_3}{\mu} \right) ./ \left(\frac{\lambda_3}{\mu} \mathbf{W}^{t+1} + \mathbf{1} \right) \end{aligned} \quad (32)$$

where $./$ means the division of the corresponding elements.

3.4. The complete algorithm

Getting data $\mathbf{Y}$, we choose the corresponding Algorithm 1 or 2 to solve (11) or (25) depending on whether the data points

Algorithm 1 Solving Eq. (11).

Input: $\mathbf{Y}, \mathbf{K}, \lambda_1, \lambda_2, \lambda_3$.

1: Initialize: $\mathbf{Z} = \sqrt{\mathbf{K}}, \mathbf{G} = 0, \mathbf{H} = 0, \Lambda_1 = 0, \Lambda_2 = 0, \rho = 10^{-8}, \rho_m = 10^8, \eta = 11, \varepsilon = 10^{-6}$.

2: **while** not converged **do**

3: step 1: Update $\mathbf{G}, \mathbf{H}, \mathbf{Z}$ in the order of (16), (18), (20), respectively;

4: step 2: Update $\Lambda_1, \Lambda_2, \rho$, according to the following equation

5: $\Lambda_1 := \Lambda_1 + \rho(\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G}))$,

6: $\Lambda_2 := \Lambda_2 + \rho(\mathbf{1}^T \mathbf{G} - \mathbf{1}^T)$,

7: $\rho := \min(\eta\rho, \rho_m)$;

8: step3: Check the convergence conditions

9: $\|\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G})\|_\infty \leq \varepsilon$, and $\|\mathbf{1}^T \mathbf{G} - \mathbf{1}^T\|_\infty \leq \varepsilon$;

10: **end while**

Output: Coefficient matrix $\mathbf{G}^*$.

Algorithm 2 Solving Eq. (25).**Input:** $\mathbf{Y}, \mathbf{K}, \lambda_1, \lambda_2, \lambda_3$.

```

1: Initialize:  $\mathbf{Z} = \sqrt{\mathbf{K}}, \mathbf{G} = 0, \mathbf{H} = 0, \mathbf{E} = 0, \mathbf{\Lambda}_1 = 0, \mathbf{\Lambda}_2 = 0, \mathbf{\Lambda}_3 = 0, \rho = 10^{-8}, \rho_m = 10^8, \eta = 11, \varepsilon = 10^{-6}$ .
2: while not converged do
3:   step 1: Update  $\mathbf{G}, \mathbf{H}, \mathbf{Z}, \mathbf{E}$  in the order of (16), (18), (20), (29), respectively;
4:   step 2: Update  $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{\Lambda}_3, \rho$ , according to the following equation
5:      $\mathbf{\Lambda}_1 := \mathbf{\Lambda}_1 + \rho(\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G}))$ ,
6:      $\mathbf{\Lambda}_2 := \mathbf{\Lambda}_2 + \rho(\mathbf{1}^T \mathbf{G} - \mathbf{1}^T)$ ,
7:      $\mathbf{\Lambda}_3 := \mathbf{\Lambda}_3 + \rho(\mathbf{K} - \mathbf{Z}^T \mathbf{Z} - \mathbf{E})$ ,
8:      $\rho := \min(\eta \rho, \rho_m)$ ;
9:   step3: Check the convergence conditions
10:     $\|\mathbf{H} - \mathbf{G} + \text{diag}(\mathbf{G})\|_\infty \leq \varepsilon$ , and  $\|\mathbf{1}^T \mathbf{G} - \mathbf{1}^T\|_\infty \leq \varepsilon$ , and  $\|\mathbf{K} - \mathbf{Z}^T \mathbf{Z} - \mathbf{E}\|_\infty \leq \varepsilon$ ;
11: end while
Output: Coefficient matrix  $\mathbf{G}^*$ .

```

are heavily polluted. The coefficient matrix $\mathbf{G}^*$ solved in the proposed model is substituted into the constructed affinity matrix $\mathbf{A} = \frac{1}{2}(|\mathbf{G}^*| + |\mathbf{G}^*|^T)$. The overall framework for the robust subspace clustering method (LAKRSC) is summarized in Algorithm 3.

Algorithm 3 The complete algorithm of LAKRSC. The reference cited in this algorithm is [24].**Input:** $\mathbf{Y}, \mathbf{K}, \lambda_1, \lambda_2, \lambda_3$.

```

1: Get  $\mathbf{G}^*$  by Algorithm 1 or 2, respectively;
2: Construct the affinity matrix by  $\mathbf{A} = \frac{1}{2}(|\mathbf{G}^*| + |\mathbf{G}^*|^T)$ ;
3: Apply clustering algorithm [24] on  $\mathbf{A}$ ;
Output: The clustering results.

```

4. Experiments

Four datasets are used to verify the effectiveness of LAKRSC. Firstly, we introduce the datasets, the compared algorithms, and experimental parameter settings. Then, we analyze the clustering performance, the corresponding parameter sensitivity, and the convergence of our algorithm in detail. Finally, it is tested that the running time of four algorithms.

4.1. Datasets

- **Extended Yale B dataset** [12]: As shown in Fig. 3(a), the face images in this dataset are collected from 38 subjects, each of whom contains 64 facial images exposed to different lights. To facilitate the experiment, the pixels of each image are put into a vector of 2016-D (48×42). Due to the presence of specular reflections, most images are damaged by shadow and noise. This dataset is useful for verifying the clustering performance of LAKRSC in handling large corruptions. Besides, when taking an image for the same object, the facial postures and expressions are not exactly the same, which results in nonlinear structural data.
- **ORL dataset** [29]: As shown in Fig. 3(b), the dataset contains 400 facial images from 40 different subjects. Each image is taken under different environments, including light, expression and so on. Similar to the extended Yale B dataset, the pixels of each image are put into a vector of 1024-D (32×32). Because of changes in facial expressions and details, this dataset is more challenging.
- **COIL-20 dataset** [23]: As shown in Fig. 3(c), this dataset contains object images of 20 subjects, which is made up of images taken every 5 degrees against a black background. The pixels of each image are vectorized to 1024-D. COIL-20 dataset has a variety of object images. And object images of the same subject are also different due to the change of perspective. Therefore, the COIL-20 dataset is challenging for testing the performance of subspace clustering methods.
- **Binary Alphadigits dataset**¹: As shown in Fig. 3(d), this dataset contains symbol images of 36 subjects, including A through Z and 0 through 9. There are 39 images in each subject, and the pixels of each image are put into a vector of 320-D. Due to the similarity of shapes between different numbers and letters, such as: the letters O, Z seem to be similar to the numbers 0, 2, respectively, this dataset is challenging.

¹ <http://www.cs.toronto.edu/roweis/data.html>.

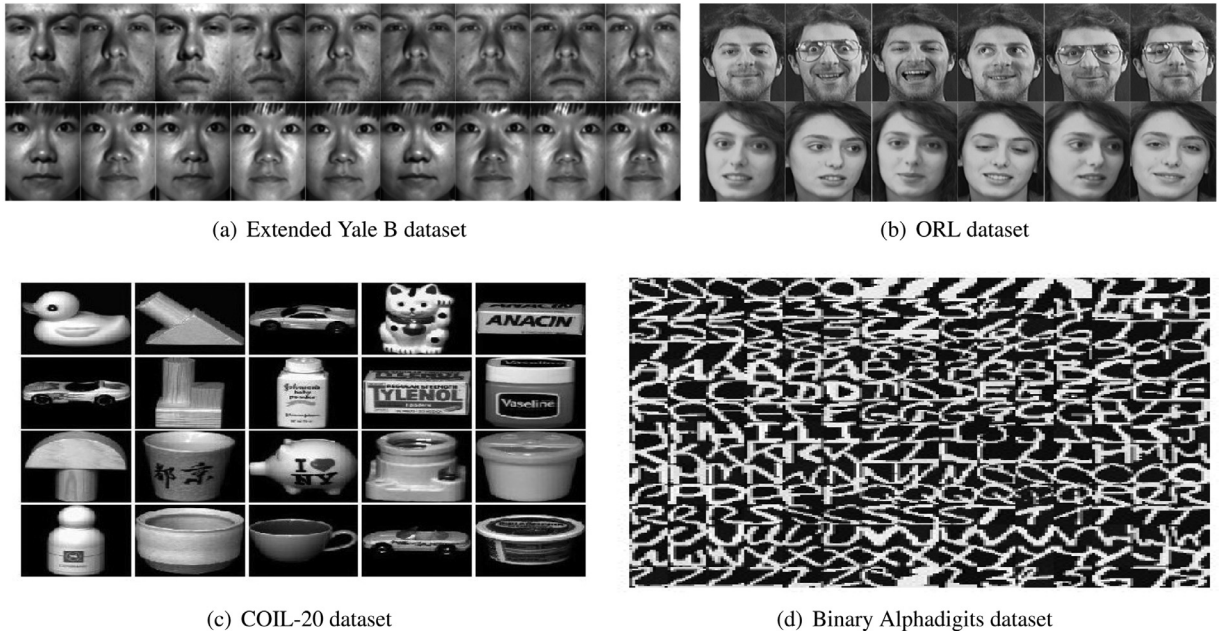


Fig. 3. Sample images of four datasets.

Table 1

Comparison of all algorithms on the Extended Yale B Dataset.

Subjects	LRR	SSC	KSSC	LRSC	SSC-OMP	WSPQ	SCHQ	LRKSC	LAKRSC
10	23.52	12.25	15.51	30.38	13.08	10.85	8.64	8.91	6.50
15	23.98	13.98	16.32	33.45	15.06	14.32	11.68	11.39	9.81
20	32.03	21.7	16.54	28.74	14.39	15.81	13.01	12.56	11.39
25	27.78	28.21	18.52	29.01	19.91	17.73	13.76	13.58	12.28
30	36.98	27.75	20.51	30.69	21.05	19.52	14.54	15.23	13.30
35	42.05	27.95	27.01	33.3	21.31	22.16	14.59	15.31	13.76
Avg	31.06	21.97	19.07	30.93	17.47	16.73	12.7	12.83	11.17

4.2. Contrast models and settings

To broadly measure the clustering effect of the LAKRSC, we compare it with the following most advanced subspace clustering methods, including LRR [16], SSC [5], KSSC [28], LRSC [31], SSC-OMP [38], LRKSC [10], WSPQ [43], SCHQ [7]. Particularly, KSSC and LRKSC are kernel-based models. WSPQ is a weighted Schatten p -norm based model, and SCHQ is a robust subspace method based on correntropy.

We use the source code of the compared methods as far as possible. The algorithm settings can be obtained from publications or adjusted to an optimal level. In our model, we use the same polynomial kernel $\mathbf{K} = (\mathbf{x}_1^T \mathbf{x}_2 + a)^b$ as LRKSC, where a is set from 1 to 12 (interval is 1), and b is set from 2 to 3 (interval is 1). The parameters $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\lambda_i (i = 1, 2, 3) \in \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$, respectively.

Clustering error is used as an evaluation index of clustering performance, which shows the ratio of wrongly clustered points, i.e.,

$$\text{Err}(\%) = \frac{N(\text{wrongly clustered points})}{N(\text{total of clustered points})} \times 100\% \quad (33)$$

Small clustering error indicates good clustering performance. Conversely, large clustering error indicates poor clustering performance.

4.3. Results and discussion

Experimental results on four datasets are organized in Table 1 to Table 4. We repeated each experiment 20 times and averaged the results. The results are shown by clustering errors (%) and only two decimal places are left. The values marked in bold in Table 1 to Table 4 are the best results for the same dataset under the different algorithms. The detailed analysis of all datasets are described as follows:



(a) Corrupted with pixel corruptions (ratio: 10% to 40%). (b) Corrupted with random block occlusions (size: 5×5 to 20×20).

Fig. 4. Sample images of the extended Yale B dataset.

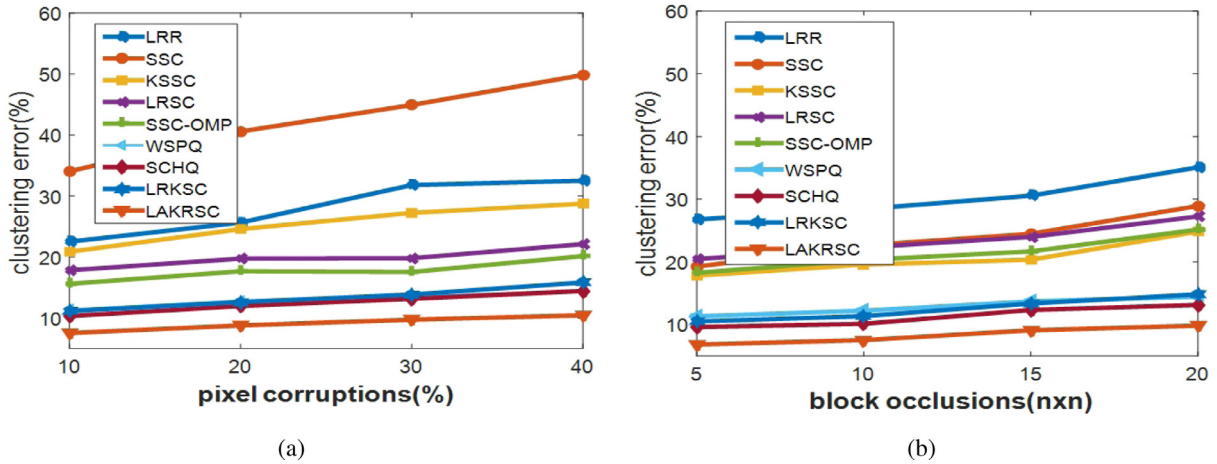


Fig. 5. Clustering Error(%) of extended Yale B dataset added to noise.

- **Extended Yale B Dataset:** On this dataset, we use Algorithm 2 and design three experiments to verify the effectiveness of LAKRSC.

(1) Experiment with different illuminations:

We arbitrarily choose N subjects from 38 subjects, $N \in \{10, 15, 20, 25, 30, 35, 38\}$. We describe the experimental process with $N = 10$. Firstly, we conduct the first N subjects, and then we carry out N subjects headed by 2 until we have completed all subjects (1–10, 2–11, or 29–38). The results of this experiment are shown in Table 1. LAKRSC outperforms other models, and its average clustering error reaches 11.17%, which shows that our model improves the clustering performance compared with other algorithms. Next, we analyze the pros and cons of several other algorithms to summarize the reasons why our model performs best on this dataset. KSSC improves the clustering accuracies over SSC. This is because nonlinear mapping of KSSC can increase the angle between two subspaces, which can improve the clustering performance. LRKSC outperforms KSSC because LRKSC makes the data in mapped feature space is not only self-expressive but also low-rank. It also is observed that WSPQ performs better than LRR, which indicates that weighted Schatten p -norm is closer to the rank of matrix than nuclear norm. Obviously, SCHQ outperforms other algorithms except our algorithm. Because SCHQ can handle corruption in the data by correntropy. The reason why the clustering effect of our model is the best is that our model uses the non-convex low-rank approximation and adaptive kernel based on weighted Schatten p -norm to process nonlinear structural data. In addition, we use correntropy to deal with noise data, which improves the robustness of the model.

(2) Experiment with non-Gaussian noise:

To evaluate the clustering performance of LAKRSC for face images with non-Gaussian noise, we select the first 10 subjects in the dataset and randomly replace some pixels in the image with evenly distributed values (even over $[0:255]$). As shown in Fig. 4(a), the four percentages of damaged facial images are 10%, 20%, 30%, 40%, respectively. All the clustering results are revealed in Fig. 5(a) for face images with non-Gaussian noise. Obviously, as the number of damaged pixels increases, all the clustering results decrease to varying degrees. Our model and SCHQ always maintain the lowest two clustering errors because both models use correntropy to deal with large corruptions. Our model demonstrates better clustering performance than SCHQ because our model uses the non-convex low-rank approximation and adaptive kernel to process the data. In addition, we also observe poor performance of SSC and KSSC, indicating that the SSC based methods are sensitive to non-Gaussian noise.

(3) Experiment with block occlusion:

The robustness of our model is further validated for block blocking. Here, we select the first 10 subjects of the extended Yale B dataset for experimentation. It is used that black blocks of different size pixels (from 5×5 to 20×20) to block the original pixels to destroy the images. The position of the black occlusion block is chosen at random. Fig. 4(b)

Table 2

Comparison of all algorithms on the ORL Dataset.

Subjects	LRR	SSC	KSSC	LRSC	SSC-OMP	WSPQ	SCHQ	LRKSC	LAKRSC
10	26.48	27.22	23.43	21.31	22.98	18.32	15.98	16.67	14.00
15	29.59	29.32	26.84	24.35	27.24	20.65	18.69	19.61	17.20
20	32.20	28.49	31.21	26.32	30.19	22.03	21.52	22.05	18.90
25	35.59	29.82	31.02	27.82	33.60	24.54	22.54	23.53	19.04
30	33.46	32.51	32.08	27.94	31.56	25.61	23.15	24.43	20.83
35	36.78	32.68	32.91	29.49	35.20	24.19	23.82	23.25	22.01
40	39.24	31.46	33.49	33.54	37.01	25.16	24.56	24.28	23.75
Avg	33.33	30.21	30.14	27.25	31.11	22.93	21.47	21.98	19.39

Table 3

Comparison of all algorithms on the COIL-20 Dataset.

Subjects	LRR	SSC	KSSC	LRSC	SSC-OMP	WSPQ	SCHQ	LRKSC	LAKRSC
2	9.83	4.01	2.51	12.67	10.65	1.49	1.56	1.36	0.18
4	12.91	7.51	3.65	16.33	14.52	2.84	3.48	2.56	1.84
6	14.89	10.41	9.32	24.98	20.14	5.56	8.32	6.83	3.01
8	24.65	21.79	12.06	26.30	24.05	8.57	12.56	9.38	5.82
10	25.06	21.98	15.76	29.91	27.19	12.42	13.25	12.80	7.40
12	28.72	25.96	17.62	32.19	30.25	14.41	16.01	13.73	9.67
14	34.52	31.82	16.68	35.84	32.85	15.54	16.70	13.97	13.72
16	34.41	30.05	20.54	34.65	32.95	15.63	17.85	15.50	15.98
18	35.02	31.84	21.94	37.74	34.59	17.71	19.06	16.51	16.73
Avg	25.16	20.45	12.90	28.03	26.25	10.83	12.60	10.25	8.26

Table 4

Comparison of all algorithms on the Binary Alphadigits Dataset.

Subjects	LRR	SSC	KSSC	LRSC	SSC-OMP	WSPQ	SCHQ	LRKSC	LAKRSC
2	8.81	6.95	6.53	16.82	7.85	6.65	6.01	6.09	4.68
3	15.39	14.72	13.96	26.71	13.32	13.35	13.05	12.67	10.60
5	25.52	25.37	23.65	38.87	22.62	22.84	21.95	21.71	19.59
8	31.56	31.20	32.07	48.91	31.26	30.56	30.69	30.15	28.94
10	43.84	33.26	34.84	51.76	34.52	33.29	32.28	32.63	31.78
Avg	25.13	22.30	22.21	36.61	21.91	21.34	20.80	20.65	19.12

shows some corrupted samples with random block occlusion. It is shown in Fig. 5(b) that the clustering error of each algorithm for block occlusion data. We can see that in all cases, LAKRSC is significantly better than other algorithms, which further demonstrates the robustness of our model. We can observe that all algorithms get high clustering errors as the area of the black occlusion block increases. This is because the more occlusion pixels, the more the facial image is destroyed and the less information is available.

- **ORL Dataset:**

For this dataset, we use Algorithm 2 and randomly choose N subjects from 40 subjects, $N = 10, \dots, 40$. The results of this experiment are shown in Table 2. Obviously, when $N = 20, 25, 30, 35$ and 40, the clustering results of KSSC is lower than SSC. This may be because each image under different subjects is too small to span the entire space. While, our model can better handle this “very-few-samples” by adaptively solving low rank feature maps. Furthermore, as the number of subject increases, the accuracy of all methods decreases, but our model is superior to other methods in all cases. In particular, SSC, KSSC and LRKSC perform better than LRSC and LRR. It is not difficult to find that LRR and LRSC are two traditional low rank algorithms. Therefore, the above phenomenon may be because each facial image in this dataset has detailed changes, the traditional low-rank model cannot accurately approximate the rank of the matrix. WSPQ obtains better clustering results because weighted Schatten p -norm can approximate the actual rank more accurately and the q -norm can deal with noise in data better.

- **COIL-20 Dataset:**

The Algorithm 2 is applied on this dataset with $N = 2, \dots, 18$. As shown in Table 3, the performance of LRKSC and KSSC is always better than SSC, which indicates that the COIL-20 dataset has strong nonlinearity. The better clustering performance of WSPQ shows that the weighted Schatten p -norm can effectively approximate the actual rank. Our model processes nonlinear structural data by learning non-convex low-rank approximation and adaptive kernel mapping, thus achieving the better clustering performance. When $N = 16, 18$, the clustering errors of LAKRSC are lower than 0.48% and 0.22% than LRKSC, respectively. This may be because our model uses correntropy to model noise in data. Correntropy is closely related to the M-estimator [7], which results our model has a potential relationship with

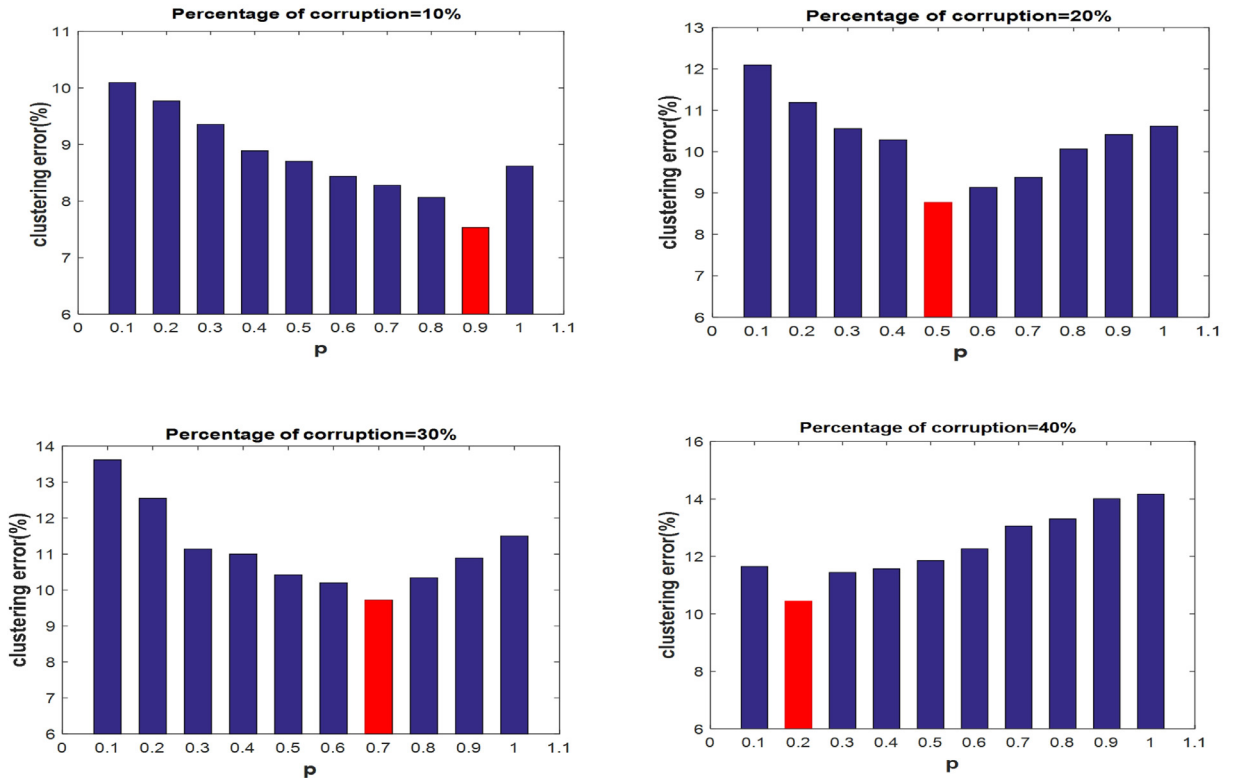


Fig. 6. Clustering results of extended Yale B dataset ($N = 10$) with different p under different percentage corruptions.

M-estimator. However, LAKRSC achieves the best clustering performance on the COIL-20 dataset and its average error is 8.26% which is 1.99% lower than the average error provided by LRKSC.

• Binary Alphanumeric Dataset:

On the Binary Alphanumeric dataset, we use Algorithm 1 and select different N to verify the performance of our model, $N = 2, \dots, 10$. The clustering results are shown in Table 4. Obviously, LAKRSC achieves the best clustering performance on this dataset. Specifically, comparing the second best result provided by LRKSC, LAKRSC achieves an improvement of approximately 1.53%. This result demonstrates that the non-convex low-rank approximation and adaptive kernel is more suitable for modeling complex nonlinear structural data than other low-rank kernels.

4.4. Parameter sensitivity and convergence analysis

LAKRSC have four important parameters, i.e., p , λ_1 , λ_2 , λ_3 . p is used to control low-rank terms, and λ_1 , λ_2 , λ_3 can balance the effects of the regular term and data item. The first 10 subjects of extended Yale B dataset are used to analyze the effectiveness of different p settings on clustering performance. To verify that different percentages of corruptions have different p settings, we add noise to extended Yale B dataset and experiment with p from 0.1 to 1 (interval 0.1) at four different noise levels (percentages=10%, 20%, 30%, 40%).

As shown in Fig. 6, the optimal values of p corresponding to the four noise levels are 0.9, 0.5, 0.7 and 0.2, respectively. Thus it may be known, as the noise level changes, the optimal value of parameter p changes. In theory, in order to achieve good results, weighted Schatten p -norm takes a small value of p , which is closer to the rank function. However, various noise contained in the dataset will destroy the low rank structure to varying degrees such that the low rank property of the dataset is weak, which makes the weight Schatten p -norm unsuitable for taking small value of p . In general, different datasets are suitable for different values of p to obtain an appropriate rank approximation. Next, we use the COIL-20 dataset ($N = 2$) to analyze the effect of parameters λ_i ($i = 1, 2, 3$) on the common balance noise. Fig. 7 shows the experimental results of $\lambda_1 \in [0.8 \times 10^3, 1.1 \times 10^3]$, $\lambda_2 \in [0.5 \times 10^{-2}, 3 \times 10^{-2}]$, $\lambda_3 \in [0.5 \times 10^5, 3 \times 10^5]$. When $\lambda_1 = 0.8 \times 10^3$, $\lambda_2 = 1 \times 10^{-2}$, and $\lambda_3 = 3 \times 10^5$, LAKRSC gets the best clustering performance on the COIL-20 dataset ($N = 2$).

For the convergence analysis, we use the original residual and objective value data of the two datasets of extended Yale B, COIL-20 to make graphs (see Fig. 8) with the number of iterations. Obviously, LAKRSC achieves a rapid reduction of the original residual to close to 10^{-6} in 20 iterations, which indicates that LAKRSC can achieve good clustering effect quickly and efficiently.

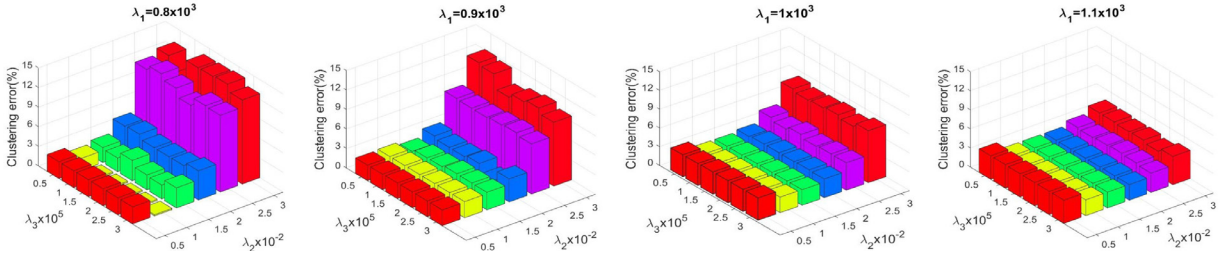


Fig. 7. The clustering performance of COIL-20 dataset ($N = 2$) with the varied parameters λ_i ($i = 1, 2, 3$).

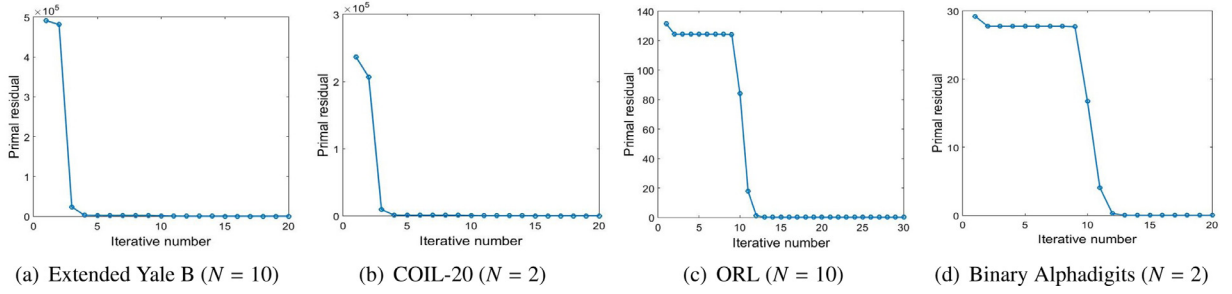


Fig. 8. Convergence curves of LAKRSC on four datasets.

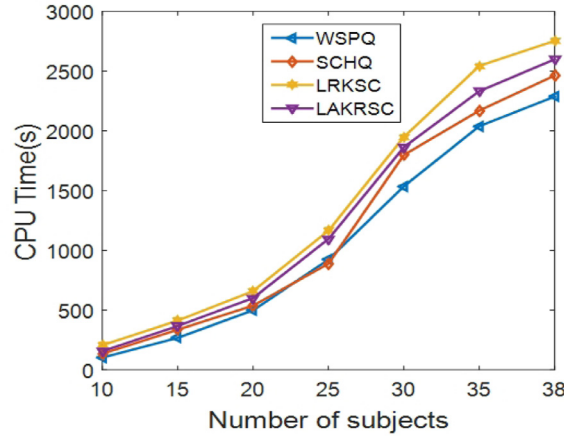


Fig. 9. Running time on extended Yale B dataset.

4.5. Computational time comparison

We measure the running time of four most advanced algorithms in this paper on extended Yale B dataset ($N = 10$). All experiments are performed using a MATLAB 2015b in a desktop computer with an Inter(R) Core(TM) i7-4790 CPU and 8G RAM. In the experiments, the convergence conditions of the compared algorithms are consistent.

Fig. 9 reports the results of running time. The results show that LAKRSC is more effective than LRKSC. Although WSPQ and SCHQ have advantages in computing time, they don't use kernel constraints to process nonlinear structural data. Our model uses the non-convex low-rank approximation and adaptive kernel to map nonlinear structural data to linear space. It takes some time to solve Z in the mapping process. But our model achieves the best clustering performance in all comparison models.

5. Conclusion

A new robust subspace clustering method (LAKRSC) based on non-convex low-rank approximation and adaptive kernel is proposed in this paper. Compared with the existing advanced kernel based methods, LAKRSC takes into account the importance of each rank components by applying weighting strategy. The weighted Schatten p -norm can approximate the actual rank better. Therefore, the data in the feature space obtained by LAKRSC is closer to low-rank and self-expression,

which indicates that our model can better deal with nonlinear structural data. Second, correntropy as a loss function can deal with corruptions damage better and improve robustness of LAKRSC. Then, an algorithm (HQ&ADMM) is designed to solve LAKRSC. Finally, we verify the good clustering performance of LAKRSC by experimenting on four datasets. In addition, we introduce the selection method of parameters in the model and verify the convergence of the model. In the future work, we need to do more theoretical research on how the model adaptively selects parameters. In addition, how to use the proposed model on multi-view data is also a future research content.

Declaration of Competing Interest

None.

Acknowledgment

This work was supported by the [National Natural Science Foundation of China](#) (Grant no. 61772272), the National Nuclear Energy Development Project of China (Grant no. 18zg6103), the [Department of Science and Technology of Sichuan Province](#) (Grant nos. 2019YJ0309, 18YYJC1688), the Scientific Research Project of Education Department of Sichuan Province (Grant no. 18ZB0611) and the Project of Science and Technology of Jiangsu Province (Grant no. BE2017031).

The authors would like to thank Ji Pan for providing the code for LRKSC. Besides, Xue Xuqian and Zhang Xiaoqian equally contributed to this article. In addition, some anonymous commentators have also helped us with our work. We are grateful to them here.

Appendix

We generalize the [Lemma 1](#) as

$$\min_y \left(wy^p + \frac{1}{2}(\delta - y^2)^2 \right) \quad (34)$$

Let the $\tau_p^{GTS}(w_i)$ denotes the critical value of δ and its corresponding local minimum y^* . We solve the following system of nonlinear equations to determine the correct threshold

$$(y^*)^p + \frac{1}{2}(\tau_p^{GTS}(\delta) - (y^*)^2)^2 = \frac{1}{2}(\tau_p^{GTS}(\delta))^2 \quad (35)$$

$$p(y^*)^{p-1} + 2y^*(\tau_p^{GTS}(\delta) - (y^*)^2) = 0 \quad (36)$$

Based on (36), we substitute in (35) with $(y^*)^2 + \frac{wp(y^*)^{p-2}}{2}$, and the new equation is as follows

$$(y^*)^p + \frac{1}{2} \left((y^*)^2 + \frac{wp(y^*)^{p-2}}{2} - (y^*)^2 \right)^2 = \frac{1}{2} \left((y^*)^2 + \frac{wp(y^*)^{p-2}}{2} \right)^2 \quad (37)$$

Thus the only solution of y^* in the range $(0, +\infty)$ can be obtained as

$$y^* = (2 - wp)^{\frac{1}{4-p}} \quad (38)$$

and the thresholding value $\tau_p^{GTS}(\delta)$ is

$$\tau_p^{GTS}(w) = (2 - wp)^{\frac{2}{4-p}} + \frac{wp(2 - wp)^{\frac{p-2}{4-p}}}{2} \quad (39)$$

For any $\delta \in (\tau_p^{GTS}(w), +\infty)$, $wy^p + \frac{1}{2}(\delta - y^2)^2$ has one unique local minimum $S_p^{GTS}(\delta, w)$ in the range of $y \in (0, +\infty)$ which satisfies

$$p(S_p^{GTS}(\delta, w))^{p-1} + 2S_p^{GTS}(\delta, w)(\tau_p^{GTS}(\delta) - (S_p^{GTS}(\delta, w))^2) = 0 \quad (40)$$

Suppose matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ have the same order, according to **Theorem**, we can obtain

$$\begin{aligned} \|\mathbf{S}\|_{\mathbf{w}, S_p}^p + \frac{\mu}{2} \|\mathbf{Q} - \mathbf{S}^T \mathbf{S}\|_F^2 &= tr(\mathbf{w}\mathbf{\Gamma}^p) + \frac{\mu}{2} \left(tr(\mathbf{\Sigma}^T \mathbf{\Sigma}) + tr((\mathbf{\Gamma}^2)^T \mathbf{\Gamma}^2) - 2tr(\mathbf{Q}^T \mathbf{S}^T \mathbf{S}) \right) \\ &\geq tr(\mathbf{w}\mathbf{\Gamma}^p) + \frac{\mu}{2} \left(tr(\mathbf{\Sigma}^T \mathbf{\Sigma}) + tr((\mathbf{\Gamma}^2)^T \mathbf{\Gamma}^2) - 2tr(\mathbf{\Sigma}^T \mathbf{\Gamma}^T \mathbf{\Gamma}) \right) \\ &= tr(\mathbf{w}\mathbf{\Gamma}^p) + \frac{\mu}{2} \|\mathbf{\Sigma} - \mathbf{\Gamma}^2\|_F^2 \end{aligned} \quad (41)$$

According to the **Theorem**, the condition for the equation to occur is $\mathbf{S} = \mathbf{Q}$. Hence, (22) can be equivalent to

$$\min_{\delta_1, \delta_2, \dots, \delta_r} \sum_{i=1}^r (w_i \delta_i^p + \frac{\mu}{2} (\gamma_i - \delta_i^2)^2)$$

$$\text{s.t. } \delta_i \geq 0, \text{ and } \delta_i \geq \delta_j, \text{ for } i \leq j. \quad (42)$$

From Xie et al. [34] known, if the weights satisfy $0 \leq w_1 \leq w_2 \leq \dots \leq w_r$, problem (42) can be solved by r independent subproblems

$$\min_{\delta_i \geq 0} \left(w_i \delta_i^p + \frac{\mu}{2} (\gamma_i - \delta_i^2)^2 \right) \quad (43)$$

There exists a δ_i^* such that

$$\frac{\mu}{2} \left(\delta_i^4 - 2\gamma_i \delta_i^2 + \gamma_i^2 + \frac{2}{\mu} w_i \delta_i^p \right) \geq \frac{\mu}{2} (\gamma_i - (\delta_i^*)^2)^2 + w_i (\delta_i^*)^p, \quad \forall \delta_i \geq 0 \quad (44)$$

References

- [1] F.R. Chung, F.C. Graham, Spectral Graph Theory, Number 92, American Mathematical Soc., 1997.
- [2] T. Deng, D. Ye, R. Ma, H. Fujita, L. Xiong, Low-rank local tangent space embedding for subspace clustering, Inf. Sci. 508 (2020) 1–21.
- [3] Z. Deng, K.-S. Choi, Y. Jiang, J. Wang, S. Wang, A survey on soft subspace clustering, Inf. Sci. 348 (2016) 84–106.
- [4] S. Ding, H. Jia, M. Du, Y. Xue, A semi-supervised approximate spectral clustering algorithm based on HMRF model, Inf. Sci. 429 (2018) 215–228.
- [5] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
- [6] S. Gu, L. Zhang, W. Zuo, X. Feng, Weighted nuclear norm minimization with application to image denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2862–2869.
- [7] R. He, Y. Zhang, Z. Sun, Q. Yin, Robust subspace clustering with complex noise, IEEE Trans. Image Process. 24 (11) (2015) 4001–4013.
- [8] R. He, W.-S. Zheng, T. Tan, Z. Sun, Half-quadratic-based iterative minimization for robust sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2013) 261–275.
- [9] C. Hong, J. Yu, J. Zhang, X. Jin, K.-H. Lee, Multi-modal face pose estimation with multi-task manifold deep learning, IEEE Trans. Ind. Inf. (2018).
- [10] P. Ji, I. Reid, R. Garg, H. Li, M. Salzmann, Low-rank kernel subspace clustering, CoRR, arXiv:1707.04974 (2017).
- [11] M. Keuper, S. Tang, B. Andres, T. Brox, B. Schiele, Motion segmentation & multiple object tracking by correlation co-clustering, IEEE Trans. Pattern Anal. Mach. Intell. (2018).
- [12] K.-C. Lee, J. Ho, D.J. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. (5) (2005) 684–698.
- [13] G. Li, T.K. Pong, Global convergence of splitting methods for nonconvex composite optimization, SIAM J. Optim. 25 (4) (2015) 2434–2460.
- [14] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Advances in Neural Information Processing Systems, 2011, pp. 612–620.
- [15] 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.
- [16] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: ICML, vol. 1, 2010, p. 8.
- [17] L. Liu, W. Huang, D.-R. Chen, Exact minimum rank approximation via Schatten p -norm minimization, J. Comput. Appl. Math. 267 (2014) 218–227.
- [18] W. Liu, P.P. Pokharel, J.C. Principe, Correntropy: properties and applications in non-Gaussian signal processing, IEEE Trans. Signal Process. 55 (11) (2007) 5286–5298.
- [19] X. Liu, S. Zhou, Y. Wang, M. Li, Y. Dou, E. Zhu, J. Yin, Optimal neighborhood kernel clustering with multiple kernels, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- [20] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, Z. Lin, Correntropy induced L2 graph for robust subspace clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1801–1808.
- [21] 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: European Conference on Computer Vision, Springer, 2012, pp. 347–360.
- [22] L. Mirsky, A trace inequality of John von Neumann, Monatsh. Math. 79 (4) (1975) 303–306.
- [23] S.A. Nene, S.K. Nayar, H. Murase, et al., Columbia object image library (coil-20)(1996).
- [24] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: Advances in Neural Information Processing Systems, 2002, pp. 849–856.
- [25] M. Nikolova, M.K. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, SIAM J. Sci. Comput. 27 (3) (2005) 937–966.
- [26] T.-H. Oh, H. Kim, Y.-W. Tai, J.-C. Bazin, I. So Kweon, Partial sum minimization of singular values in RPCA for low-level vision, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 145–152.
- [27] V.M. Patel, H. Van Nguyen, R. Vidal, Latent space sparse subspace clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 225–232.
- [28] V.M. Patel, R. Vidal, Kernel sparse subspace clustering, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 2849–2853.
- [29] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, IEEE, 1994, pp. 138–142.
- [30] X. Shi, Z. Guo, F. Xing, J. Cai, L. Yang, Self-learning for face clustering, Pattern Recognit. 79 (2018) 279–289.
- [31] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), Pattern Recognit. Lett. 43 (2014) 47–61.
- [32] G. Xia, H. Sun, L. Feng, G. Zhang, Y. Liu, Human motion segmentation via robust kernel sparse subspace clustering, IEEE Trans. Image Process. 27 (1) (2017) 135–150.
- [33] S. Xiao, M. Tan, D. Xu, Z.Y. Dong, Robust kernel low-rank representation, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2015) 2268–2281.
- [34] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang, L. Zhang, Weighted Schatten p -norm minimization for image denoising and background subtraction, IEEE Trans. Image Process. 25 (10) (2016) 4842–4857.
- [35] J. Xue, Y. Zhao, W. Liao, J.C.-W. Chan, Nonconvex tensor rank minimization and its applications to tensor recovery, Inf. Sci. 503 (2019) 109–128.
- [36] C. Yang, Z. Ren, Q. Sun, M. Wu, M. Yin, Y. Sun, Joint correntropy metric weighting and block diagonal regularizer for robust multiple kernel subspace clustering, Inf. Sci. 500 (2019) 48–66.
- [37] M. Yin, Y. Guo, J. Gao, Z. He, S. Xie, Kernel sparse subspace clustering on symmetric positive definite manifolds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5157–5164.
- [38] C. You, D. Robinson, R. Vidal, Scalable sparse subspace clustering by orthogonal matching pursuit, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3918–3927.
- [39] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
- [40] J. Yu, M. Tan, H. Zhang, D. Tao, Y. Rui, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
- [41] D. Zhang, Y. Hu, J. Ye, X. Li, X. He, Matrix completion by truncated nuclear norm regularization, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2192–2199.
- [42] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Trans. Image Process. 27 (5) (2018) 2420–2432.

- [43] T. Zhang, Z. Tang, Q. Liu, Robust subspace clustering via joint weighted Schatten-p norm and L_q norm minimization, *J. Electron. Imaging* 26 (3) (2017) 33021.
- [44] X. Zhang, B. Chen, H. Sun, Z. Liu, Z. Ren, Y. Li, Robust low-rank kernel subspace clustering based on the Schatten p-norm and correntropy, *IEEE Trans. Knowl. Data Eng.* (2019).
- [45] X. Zhang, H. Sun, Z. Liu, Z. Ren, Q. Cui, Y. Li, Robust low-rank kernel multi-view subspace clustering based on the Schatten p-norm and correntropy, *Inf. Sci.* 477 (2019) 430–447.
- [46] R. Zhu, M. Dong, J.-H. Xue, Learning distance to subspace for the nearest subspace methods in high-dimensional data classification, *Inf. Sci.* 481 (2019) 69–80.