Journal of Electronic Imaging

JElectronicImaging.org

Robust subspace clustering via joint weighted Schatten-p norm and L_q norm minimization

Tao Zhang
Zhenmin Tang
Qing Liu

Tao Zhang, Zhenmin Tang, Qing Liu, "Robust subspace clustering via joint weighted Schatten-p norm and L_q norm minimization," *J. Electron. Imaging* **26**(3), 033021 (2017), doi: 10.1117/1.JEI.26.3.033021.

Robust subspace clustering via joint weighted Schatten-p norm and L_q norm minimization

Tao Zhang,* Zhenmin Tang, and Qing Liu
Nanjing University of Science and Technology, School of Computer Science and Engineering, Nanjing, China

Abstract. Low-rank representation (LRR) has been successfully applied to subspace clustering. However, the nuclear norm in the standard LRR is not optimal for approximating the rank function in many real-world applications. Meanwhile, the L_{21} norm in LRR also fails to characterize various noises properly. To address the above issues, we propose an improved LRR method, which achieves low rank property via the new formulation with weighted Schatten-p norm and L_q norm (WSPQ). Specifically, the nuclear norm is generalized to be the Schatten-p norm and different weights are assigned to the singular values, and thus it can approximate the rank function more accurately. In addition, L_q norm is further incorporated into WSPQ to model different noises and improve the robustness. An efficient algorithm based on the inexact augmented Lagrange multiplier method is designed for the formulated problem. Extensive experiments on face clustering and motion segmentation clearly demonstrate the superiority of the proposed WSPQ over several state-of-the-art methods. *© 2017 SPIE and IS&T* [DOI: 10.1117/1.JEI.26.3.033021]

Keywords: low-rank representation; weighted Schatten-p norm; L_q norm; face clustering; motion segmentation.
Paper 170053 received Jan. 23, 2017; accepted for publication May 15, 2017; published online Jun. 6, 2017.

1 Introduction

In recent years, subspace clustering techniques have attracted much attention due to its widespread applications in computer vision and machine learning, such as image representation,[1,2] face clustering,[3,4] and motion segmentation.[5,6] Subspace clustering refers to the task of recovering low-dimensional subspace structures[7] in the high-dimensional data and clustering each group of data into its own subspace simultaneously. Until now, many subspace clustering methods have been proposed in the literature[8] and they can mainly be divided into four categories: iterative methods,[9,10] algebraic methods,[11,12] statistical methods,[13,14] and spectral clustering-based methods.[15,16]

The spectral clustering-based methods can be implemented easily, solved efficiently, and achieve competitive results in real-world applications. The spectral clustering-based methods perform subspace clustering in two stages: they first learn an affinity matrix from the observed data and then apply spectral clustering algorithms[17,18] on the affinity matrix to get the final clustering results. Therefore, how to construct a good affinity matrix is the main challenge for spectral clustering-based methods. Inspired by recent advances in sparse[19,20] recovery theory, sparse subspace clustering (SSC)[21] was developed for subspace clustering. SSC constructs the affinity matrix[22] using the sparse representation of each data point with respect to all the other data points, which is obtained by solving an L_1 norm minimization problem. SSC is more robust to noise and outliers and does not require the number and dimensions of subspaces in advance. However, SSC considers each data point individually, which fails to discover the global structure of data.

Recently, low rank matrix recovery methods[23,24] have attracted great attention in the fields of subspace learning,[25] transfer learning,[26] and multiview learning.[27] Among them, low-rank representation (LRR)[28] is one typical model for solving the subspace clustering problem. By solving a rank minimization problem, LRR seeks the lowest rank representation of the whole data and thus can capture the global structure of data. Nevertheless, rank minimization of a matrix is a challenging NP-hard problem, and it is commonly relaxed to a convex nuclear norm minimization (NNM) problem.[24] Although NNM can find a globally optimal solution of the relaxed problem, this solution may be suboptimal for the original problem.

Many efficient nonconvex approaches have been proposed to solve the rank minimization problem. Fazel et al. developed the Log-det function[29] to replace the rank function, but it has a significantly biased approximation for small singular values. Hu et al. proposed the truncated nuclear norm (TNN),[30] which minimizes the smallest several singular values for the matrix completion problem. TNN requires an accurate estimation of the rank of the data matrix, which is a hard task in practice. In addition, Nie et al. proposed the Schatten-p norm,[31] which is a generalization of the nuclear norm. Theoretically, it has been proven that the Schatten-p norm minimization achieves better recovery for the matrix completion problem while a weaker restricted isometry property is required.[32]

Later, the Schatten-p norm has been widely studied. Nie and Huang[33] proposed a new subspace clustering model based on the Schatten-p norm, which can not only directly learn the group indicator instead of affinity matrix to obtain the low rank structure but also better approximate the low rank constraint. Yu and Schuurmans[34] optimized the unitarily invariant norm, which is a generalization of the Schatten-p norm, and derived an efficient closed-form solution for the subspace clustering problem. Moreover, Xie et al. proposed a

*Address all correspondence to: Tao Zhang, E-mail: 310062204@njust.edu.cn

more flexible model, namely the weighted Schatten-p norm,[35] which assigns weights to different singular values for the Schatten-p norm. The weighted Schatten-p norm is a better approximation to the rank function as it considers the importance of different rank components. Until now, it has been successfully applied to many image processing problems, such as image denoising, background subtraction, and hyperspectral image restoration.[36] The success inspires us to use the weighted Schatten-p norm to approximate the rank function in the subspace clustering problem.

The L_{21} norm in the standard LRR is used to characterize the noise, but it is only suitable for sample-specific corruptions. Later, some methods[37,38] use other norms to describe different types of noise. For example, the L_1 norm can describe random corruptions, and the squared Frobenius norm is suitable for Gaussian noise. Generally speaking, the above norms can be only applied to one specific type of noise, which limits the application for different settings. A generic model, i.e., L_q norm can be used to describe various noises by setting different values for q. For example, the L_q norm reduces to the L_1 norm if $q = 1$ and reduces to the squared Frobenius norm if $q = 2$.

To improve the robustness of LRR for the subspace clustering problem, an LRR model, i.e., weighted Schatten-p norm and L_q norm regularized LRR (WSPQ), is proposed in this paper. Specifically, the weighted Schatten-p norm and L_q norm are introduced to approximate the rank of the coefficient matrix and complex noises, respectively. The proposed WSPQ can better approximate the rank function and is more robust to various noises, such as illumination variations, corruptions, and occlusions. Based on the inexact augmented Lagrange multiplier (ALM)[39] method, we design an efficient optimization algorithm to solve the proposed model. Experiments on several datasets, i.e., extended Yale B, AR face datasets, and Hopkins155 motion segmentation dataset, show that the proposed WSPQ has better clustering performance than the previous methods. Figure 1 shows the proposed WSPQ for subspace clustering.

The remainder of this paper is organized as follows. We give an overview of the related works in Sec. 2. In Sec. 3, we present the proposed WSPQ and the detailed optimization. The experimental results are reported in Sec. 4. Finally, we conclude this paper in Sec. 5.

2 Related Work

In this section, some research background about subspace clustering is introduced, such as the subspace clustering problem, the LRR model, and the inexact ALM method, which is a basic optimization technique for LRR.

2.1 *Subspace Clustering Problem*

Let $\mathbf{X} = \{x_i \in \mathbf{R}^d\}_{i=1}^n$ be a given set of points drawn from an unknown of k linear or affine subspaces $\{S_i\}_{i=1}^k$ of unknown dimensions $d_k = \dim(S_k)$. The task of subspace clustering is to find the number of subspaces k and cluster each data point x_i according to the k subspaces. Figure 2 is the illustration of the subspace clustering problem.

2.2 *LRR Model*

The recently proposed LRR performs subspace clustering excellently. The task of LRR is to find the lowest rank representation among all the linear combinations of the bases in a given dictionary. Let $\mathbf{X} = \{x_i \in \mathbf{R}^d\}_{i=1}^n$ be a given set of n points in the d-dimensional space. LRR aims to solve the following rank minimization problem:

$$\min_{\mathbf{Z}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t. } \mathbf{X} = \mathbf{AZ}, \tag{1}$$

where $\mathbf{A} = [a_1, a_2, \ldots, a_n] \in \mathbf{R}^{d\times n}$ is the dictionary matrix and $\mathbf{Z} = [z_1, z_2, \ldots, z_n] \in \mathbf{R}^{n\times n}$ is the coefficient matrix with each z_i being the representation coefficient of x_i.

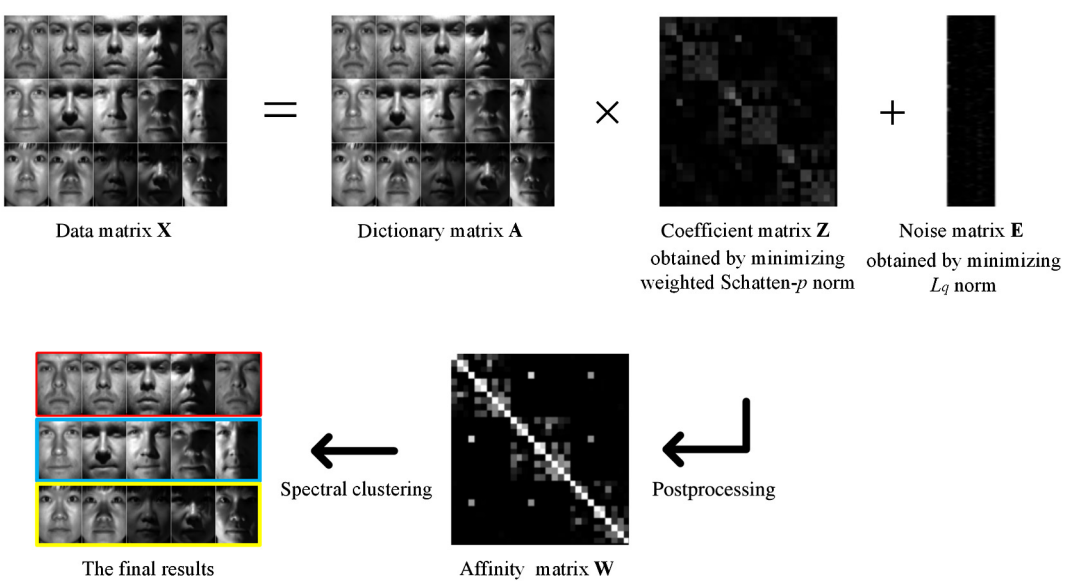


Fig. 1 Illustration of the proposed model. The data matrix **X** contains three subjects of face images, and we set the dictionary matrix $\mathbf{A} = \mathbf{X}$. By solving the weighted Schatten-p norm and L_q norm minimization problem, we obtain the coefficient matrix **Z**. The affinity matrix **W** is acquired by a postprocessing of the coefficient matrix **Z**, and the final results are obtained by employing the classical spectral clustering algorithm.

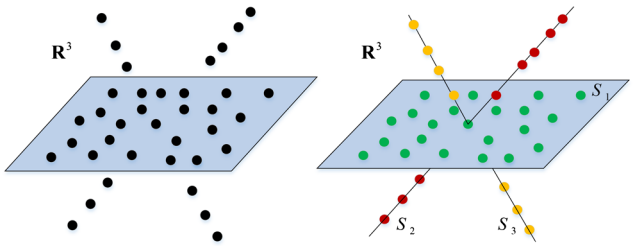


Fig. 2 Illustration of subspace clustering problem. The left is a set of sample points drawn from a union of three subspaces. The right is the clustering results: two lines (one-dimensional subspace) and a plane (two-dimensional subspace).

Since Eq. (1) is NP-hard, LRR uses the nuclear norm to replace the rank function and solves the following convex optimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* \quad \text{s.t. } \mathbf{X} = \mathbf{AZ}, \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, defined as the sum of all singular values of the matrix. Considering that the observed data are usually noisy or corrupted, a more robust objective function for LRR can be expressed as

$$\min_{\mathbf{Z}} (\|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_l) \quad \text{s.t. } \mathbf{X} = \mathbf{AZ} + \mathbf{E}, \tag{3}$$

where $\lambda > 0$ is a balance parameter, $\mathbf{E} \in \mathbf{R}^{d \times n}$ represents the unknown noise or corruption, and $\|\cdot\|_l$ can be L_1 norm $\|\mathbf{E}\|_1 = \sum_{i=1}^{d} \sum_{j=1}^{n} |\mathbf{E}_{ij}|$, L_{21} norm $\|\mathbf{E}\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{d} \mathbf{E}_{ij}^2}$, or squared Frobenius norm $\|\mathbf{E}\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{n} \mathbf{E}_{ij}^2$. The standard LRR uses the L_{21} norm to measure the noise term. Then the inexact ALM method is employed to solve Eq. (3). After obtaining the minimizer $(\mathbf{Z}^*, \mathbf{E}^*)$, the affinity matrix used for subspace clustering is defined as

$$\mathbf{W} = |\mathbf{Z}^*| + |\mathbf{Z}^{*T}|. \tag{4}$$

2.3 *Inexact ALM Method*

The inexact ALM method is widely used for solving the low rank matrix recovery problem. Consider the following constrained optimization problem:

$$\min_{\mathbf{x},\mathbf{y}} [f(\mathbf{x}) + g(\mathbf{y})] \quad \text{s.t. } A(\mathbf{x}) + B(\mathbf{y}) = \mathbf{c}, \tag{5}$$

where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{c}$ are the matrices, f and g are the convex functions, and A and B are the linear mappings. The augmented Lagrangian function of Eq. (5) can be expressed as follows:

$$L(\mathbf{x}, \mathbf{y}, \mathbf{n}) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{n}, A(\mathbf{x}) + B(\mathbf{y}) - \mathbf{c} \rangle + \frac{\mu}{2} \|A(\mathbf{x}) + B(\mathbf{y}) - \mathbf{c}\|_F^2, \tag{6}$$

where $\mathbf{n}$ is a Lagrange multiplier and μ is a penalty parameter. By fixing the other variables, the inexact ALM method minimizes $\mathbf{x}$ and $\mathbf{y}$ alternately, and the updating rules are as follows:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}_k, \mathbf{n}_k) = \arg\min_{\mathbf{x}} \left[f(\mathbf{x}) + \frac{\mu_k}{2} \|A(\mathbf{x}) + B(\mathbf{y}_k) - \mathbf{c} + \mathbf{n}_k/\mu_k\|_F^2 \right], \tag{7}$$

$$\mathbf{y}_{k+1} = \arg\min_{\mathbf{y}} L(\mathbf{x}_{k+1}, \mathbf{y}, \mathbf{n}_k) = \arg\min_{\mathbf{y}} \left[g(\mathbf{y}) + \frac{\mu_k}{2} \|A(\mathbf{x}_{k+1}) + B(\mathbf{y}) - \mathbf{c} + \mathbf{n}_k/\mu_k\|_F^2 \right]. \tag{8}$$

Equations (7) and (8) can be solved by some convex or nonconvex optimization methods. The Lagrange multiplier $\mathbf{n}$ and penalty parameter μ are adaptively updated as follows:

$$\mathbf{n}_{k+1} = \mathbf{n}_k + \mu_k [A(\mathbf{x}_{k+1}) + B(\mathbf{y}_{k+1}) - \mathbf{c}], \tag{9}$$

$$\mu_{k+1} = \min(\mu_{\max}, \rho \mu_k), \tag{10}$$

where $\mu_{\max}$ is an upper bound of μ and ρ is a constant. The variables are iteratively updated by Eqs. (7)–(10), until the convergence conditions are met.

3 Weighted Schatten-*p* Norm and L_q Norm Regularized LRR

In this section, we introduce the proposed WSPQ model. First, we discuss the motivation of this work. Second, we formulate the objective function of WSPQ. Third, we provide an efficient optimization algorithm to solve the objective function. Finally, we present the whole subspace clustering algorithm of WSPQ and discuss the convergence of WSPQ.

3.1 *Motivation*

This work is motivated by the following two aspects: on one hand, recent researches[38,40] have shown that the standard LRR uses the nuclear norm to approximate the rank function, which leads to a suboptimal solution, and cannot capture the global structures of data well. On the other hand, LRR utilizes the L_{21} norm to represent noise. In this case, it only describes the sample-specific corruptions and is not robust to the various types of noise. This paper focuses on improving the robustness of LRR for subspace clustering. We investigate an improved LRR model, which has a better rank approximation and a generic noise representation.

3.2 *Problem Formulation*

For coefficient matrix $\mathbf{Z}$, the rank function $\text{rank}(\mathbf{Z}) = \sum_{i=1}^{n} \sigma_i^0(\sigma_i \neq 0) = \|\sigma\|_0$ is the L_0 norm of the singular values, and the nuclear norm $\|\mathbf{Z}\|_* = \sum_{i=1}^{n} \sigma_i^1 = \|\sigma\|_1$ is the L_1 norm of the singular values. The L_1 norm is not a perfect approximation to the L_0 norm; thus, many nonconvex surrogate functions have been proposed.[29–35] The weighted Schatten-p norm, a better surrogate of the rank function, is defined as

$$\|\mathbf{Z}\|_{w,S_p} = \left(\sum_{i=1}^{n} w_i \sigma_i^p \right)^{\frac{1}{p}} \quad (0 < p \leq 1), \tag{11}$$

where $w = [w_1, w_2, \ldots, w_n]$ is the weight vector. The weighted Schatten-p norm of coefficient matrix $\mathbf{Z}$ to the power p is

$$\|\mathbf{Z}\|_{w,S_p}^p = \sum_{i=1}^{n} w_i \sigma_i^p \qquad (0 < p \leq 1). \tag{12}$$

Specifically, when $p = 1$ and $w_i = 1$, the weighted Schatten-1 norm reduces to the nuclear norm. If we define $0^0 = 0$, when $p = 0$ and $w_i = 1$, the weighted Schatten-0 norm reduces to the rank function. Compared to the nuclear norm, the weighted Schatten-p norm possesses two advantages: first, the Schatten-p norm is a generalization of the nuclear norm, which can approximate the rank function more exactly and guarantee the flexibility of dealing with different real problems; second, the weighting strategy considers the importance of different rank components. It can penalize the smaller singular values and preserve the major low rank structure of data, thus improving rank approximation.

The observed data often contains various noises, such as Gaussian noise and different corruptions. In this work, we utilize the L_q norm to describe the noise, which is defined as

$$\|\mathbf{E}\|_q = \sum_{i=1}^{d}\sum_{j=1}^{n} |\mathbf{E}_{ij}|^q \qquad (0 < q < \infty). \tag{13}$$

The L_1 norm and the squared Frobenius norm are two special cases of the L_q norm. The L_q norm can handle different noises with different values of q and thus is more generic and robust.

In summary, we use the weighted Schatten-p norm and L_q norm to replace the low rank term and noise term of the conventional LRR, respectively. The objective function of the proposed WSPQ is formulated as follows:

$$\min_{\mathbf{Z}} (\|\mathbf{Z}\|_{w,S_p}^p + \lambda \|\mathbf{E}\|_q) \quad \text{s.t. } \mathbf{X} = \mathbf{AZ} + \mathbf{E}. \tag{14}$$

3.3 *Optimization*

Similar to Liu et al.,[28] we present an inexact ALM iterative scheme to solve the optimization problem in Eq. (14). We first introduce an auxiliary variable $\mathbf{J}$ to separate the variable in the objective function and then convert Eq. (14) into the following equivalent problem:

$$\min_{\mathbf{Z}} (\|\mathbf{J}\|_{w,S_p}^p + \lambda \|\mathbf{E}\|_q) \quad \text{s.t. } \mathbf{X} = \mathbf{AZ} + \mathbf{E}, \quad \mathbf{Z} = \mathbf{J}. \tag{15}$$

As we know, the inexact ALM aims to convert the constrained optimization problem into an unconstrained problem; thus, we minimize the following augmented Lagrangian function of Eq. (15):

$$\begin{aligned} L(\mathbf{Z}, \mathbf{J}, \mathbf{E}, \mathbf{Y}_1, \mathbf{Y}_2, \mu) = {} & \|\mathbf{J}\|_{w,S_p}^p + \lambda \|\mathbf{E}\|_q \\ & + \langle \mathbf{Y}_1, \mathbf{X} - \mathbf{AZ} - \mathbf{E} \rangle + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J} \rangle \\ & + \frac{\mu}{2} (\|\mathbf{X} - \mathbf{AZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2), \end{aligned} \tag{16}$$

where $\mathbf{Y}_1 \in \mathbf{R}^{d \times n}$ and $\mathbf{Y}_2 \in \mathbf{R}^{n \times n}$ are Lagrangian multipliers and $\mu > 0$ is a penalty parameter. We update the variables $\mathbf{Z}$, $\mathbf{J}$, and $\mathbf{E}$ alternately while fixing the others. The updating rules are as follows:

$$\begin{aligned} \mathbf{J}_{k+1} &= \arg\min_{\mathbf{J}} L(\mathbf{Z}_k, \mathbf{J}_k, \mathbf{E}_k, \mathbf{Y}_{1,k}, \mathbf{Y}_{2,k}, \mu_k) \\ &= \arg\min_{\mathbf{J}} \left[\frac{1}{\mu_k} \|\mathbf{J}\|_{w,S_p}^p + \frac{1}{2} \|\mathbf{J} - (\mathbf{Z}_k + \mathbf{Y}_{2,k}/\mu_k)\|_F^2 \right], \end{aligned} \tag{17}$$

$$\begin{aligned} \mathbf{Z}_{k+1} &= \arg\min_{\mathbf{Z}} L(\mathbf{Z}_k, \mathbf{J}_{k+1}, \mathbf{E}_k, \mathbf{Y}_{1,k}, \mathbf{Y}_{2,k}, \mu_k) \\ &= \arg\min_{\mathbf{Z}} [\|\mathbf{E}_k - (\mathbf{X} - \mathbf{AZ} + \mathbf{Y}_{1,k}/\mu_k)\|_F^2 \\ &\quad + \|\mathbf{J}_{k+1} - (\mathbf{Z} + \mathbf{Y}_{2,k}/\mu_k)\|_F^2], \end{aligned} \tag{18}$$

$$\begin{aligned} \mathbf{E}_{k+1} &= \arg\min_{\mathbf{E}} L(\mathbf{Z}_{k+1}, \mathbf{J}_{k+1}, \mathbf{E}_k, \mathbf{Y}_{1,k}, \mathbf{Y}_{2,k}, \mu_k) \\ &= \arg\min_{\mathbf{E}} \left[\frac{\lambda}{\mu_k} \|\mathbf{E}\|_q + \frac{1}{2} \|\mathbf{E} - (\mathbf{X} - \mathbf{AZ}_{k+1} + \mathbf{Y}_{1,k}/\mu_k)\|_F^2 \right]. \end{aligned} \tag{19}$$

1. *updating* $\mathbf{J}$: Equation (17) is a weighted Schatten-p norm minimization problem, which can be solved via Lemma 1.

Lemma 1. Let the SVD of $\mathbf{Y} \in \mathbf{R}^{d \times n}$ be $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$ with $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, where $r = \min(d, n)$, and $\sigma_1 > \sigma_2 > \ldots > \sigma_r$. If $0 \leq w_1 \leq w_2 \leq \ldots \leq w_r$, the optimal solution $\mathbf{X}^*$ of the following problem:[36]

$$\min_{\mathbf{X} \in \mathbf{R}^{d \times n}} \left(\frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{X}\|_{w,S_p}^p \right) \tag{20}$$

is $\mathbf{X}^* = \mathbf{U}\Lambda\mathbf{V}^T$ with $\Lambda = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_r)$, where δ_i can be obtained by solving the following problem:

$$\begin{aligned} & \min_{\delta_1, \delta_2, \ldots, \delta_r} \sum_{i=1}^{r} \left[\frac{1}{2} (\delta_i - \sigma_i)^2 + \lambda w_i \delta_i^p \right], \quad \text{s.t. } \delta_i \geq 0 \\ & i = 1, 2, \ldots r. \end{aligned} \tag{21}$$

Equation (21) can be decoupled into r independent subproblems and each subproblem can be solved via the generalized soft-thresholding (GST) algorithm, which is described in Lemma 2.

Lemma 2. Given $y \in \mathbf{R}$ and $\lambda > 0$, the optimal solution x^* of the following problem:[41]

$$\min_x \left[\frac{1}{2} (x - y)^2 + \lambda |x|^p \right] \tag{22}$$

is given by

$$x^* = \begin{cases} 0 & |y| \leq \tau_p^{\mathrm{GST}}(\lambda) \\ \mathrm{sgn}(y) S_p^{\mathrm{GST}}(|y|, \lambda) & |y| > \tau_p^{\mathrm{GST}}(\lambda) \end{cases}, \tag{23}$$

where $\tau_p^{\mathrm{GST}}(\lambda) = [2\lambda(1-p)]^{\frac{1}{2-p}} + \lambda p [2\lambda(1-p)]^{\frac{p-1}{2-p}}$ and $S_p^{\mathrm{GST}}(|y|, \lambda)$ is obtained by solving the problem below:

$$S_p^{\mathrm{GST}}(|y|,\lambda)-|y|+\lambda p[S_p^{\mathrm{GST}}(|y|,\lambda)]^{p-1}=0. \tag{24}$$

In addition, Eq. (20) has the optimal solution only when $0\le w_1\le w_2\le\ldots\le w_r$ holds, that is, w_i must be in nondescending order. The weighting strategy in Eq. (17) aims to protect the major low rank structure of $\mathbf{J}$. Therefore, large singular values should be shrunk less than the small ones, which inspires us to set the weight vector w_i to be inversely proportional to $\sigma_i(\mathbf{J})$, where $\sigma_i(\mathbf{J})$ is the i'th singular value of $\mathbf{J}$. Thus, an intuitive way for setting weights is as follows:

$$w_i=C/\sigma_i(\mathbf{J}), \tag{25}$$

where C is a constant. Considering the effect of the size of the data matrix $\mathbf{X}\in\mathbf{R}^{d\times n}$, we rewrite Eq. (25) as

$$w_i=C\sqrt{dn}/\sigma_i(\mathbf{J}). \tag{26}$$

To avoid dividing by zero, we add a parameter $0<\varepsilon_1\le 0$ to Eq. (26), and thus we have

$$w_i=C\sqrt{dn}/[\sigma_i(\mathbf{J})+\varepsilon_1]. \tag{27}$$

Since $\sigma_i(\mathbf{J})$ is unknown before Eq. (17) is solved, we propose an alternative method to calculate w_i: we first initialize $\sigma_i(\mathbf{J})$ by $\sigma_i(\mathbf{X})$, where $\sigma_i(\mathbf{X})$ is the i'th singular value of $\mathbf{X}$, and then iteratively update w_i and $\sigma_i(\mathbf{J})$ until the optimal solution is obtained. Many experiments have shown that the optimal solution can be obtained with fewer iterations. The iteration number is simply set to 3 in this work. The optimization procedure of Eq. (17) is described in Algorithm 1.

Algorithm 1: Solving Eq. (17).

Input: Data matrix $\mathbf{X}$, parameter p;

Step 1: Compute the SVD: $\mathbf{X}=\mathbf{U}_{\mathbf{X}}\Sigma_{\mathbf{X}}\mathbf{V}_{\mathbf{X}}{}^T$, where $\Sigma_{\mathbf{X}}=\mathrm{diag}[(\delta_0)_1,(\delta_0)_2,\ldots,(\delta_0)_r]$;

Step 2: Compute the SVD: $(\mathbf{Z}_k+\mathbf{Y}_{2,k}/\mu_k)=\mathbf{U}\Sigma\mathbf{V}^T$, where $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_r)$;

Step 3: for $j=0$ to 2 do

for $i=1$ to r do

$(w_j)_i=C\sqrt{dn}/[(\delta_j)_i+\varepsilon_1]$.

Solve the following problem by GST

$(\delta_{j+1})_i=\arg\min_{(\delta_j)_i}\frac{1}{2}[(\delta_j)_i-\sigma_i]^2+\frac{(w_j)_i}{\mu_k}(\delta_j)_i^p$;

end

end

Step 4: Calculate: $\Lambda=\mathrm{diag}[(\delta_{j+1})_1,(\delta_{j+1})_2,\ldots,(\delta_{j+1})_r]$;

Output: Optimal solution $\mathbf{J}_{k+1}=\mathbf{U}\Lambda\mathbf{V}^T$.

2. *updating* $\mathbf{Z}$: Equation (18) has a closed-form solution

$$\mathbf{Z}=(\mathbf{I}+\mathbf{A}^T\mathbf{A})^{-1}[\mathbf{A}^T(\mathbf{X}-\mathbf{E}_k)+\mathbf{J}_{k+1}+(\mathbf{A}^T\mathbf{Y}_{1,k}-\mathbf{Y}_{2,k})/\mu_k]. \tag{28}$$

3. *updating* $\mathbf{E}$: Equation (19) is an L_q norm minimization problem. We shift $|\mathbf{E}_{ij}|^q$ to $(|\mathbf{E}_{ij}|+\varepsilon_2)^q$ with $0<\varepsilon_2\le 1$, such that Eq. (19) can be relaxed as follows

$$\mathbf{E}_{k+1}=\arg\min_{\mathbf{E}}\left[\frac{\lambda}{\mu_k}\sum_{i=1}^{d}\sum_{j=1}^{n}(|\mathbf{E}_{ij}|+\varepsilon_2)^q+\frac{1}{2}\|\mathbf{E}-(\mathbf{X}-\mathbf{A}\mathbf{Z}_{k+1}+\mathbf{Y}_{1,k}/\mu_k)\|_F^2\right]. \tag{29}$$

We linearize the objective function of Eq. (29) using the first degree Taylor expansion with respect to $\mathbf{E}$ at $\mathbf{E}=\mathbf{E}_k$ and then add a proximal term. Thus, the variable $\mathbf{E}$ can be updated by minimizing the following problem:

$$\begin{aligned}\mathbf{E}_{k+1}&=\arg\min_{\mathbf{E}}\left\{\frac{\lambda}{\mu_k}\sum_{i=1}^{d}\sum_{j=1}^{n}[|(\mathbf{E}_{ij})_k|+\varepsilon_2]^q+(\mathbf{M}_{ij})_k[|\mathbf{E}_{ij}|-|(\mathbf{E}_{ij})_k|]+\frac{1}{2}\|\mathbf{E}_k-(\mathbf{X}-\mathbf{A}\mathbf{Z}_{k+1}+\mathbf{Y}_{1,k}/\mu_k)\|_F^2\right.\\&\left.+\langle\mathbf{E}_k-(\mathbf{X}-\mathbf{A}\mathbf{Z}_{k+1}+\mathbf{Y}_{1,k}/\mu_k),\mathbf{E}-\mathbf{E}_k\rangle+\frac{\eta}{2}\|\mathbf{E}-\mathbf{E}_k\|_F^2\right\}\\&=\arg\min_{\mathbf{E}}\left[\frac{\lambda}{\mu_k\eta}\sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{M}_{ij})_k|\mathbf{E}_{ij}|+\frac{1}{2}\left\|\mathbf{E}-\left\{\mathbf{E}_k-\frac{1}{\eta}[\mathbf{E}_k-(\mathbf{X}-\mathbf{A}\mathbf{Z}_{k+1}+\mathbf{Y}_{1,k}/\mu_k)]\right\}\right\|_F^2\right],\end{aligned} \tag{30}$$

where $(\mathbf{M}_{ij})_k=q/[|(\mathbf{E}_{ij})_k|+\varepsilon_2]^{1-q}$ is the weight corresponding to $\mathbf{E}_k$. We can obtain $\mathbf{E}_{k+1}$ by solving Eq. (30), which can be calculated separately, and each $\mathbf{E}_{ij}$ is derived using Lemma 3.

Lemma 3. Given y, $w\in\mathbf{R}$, $w\ge 0$, and $\lambda>0$, the optimal solution of the following problem:[42]

$$S_{\lambda w}(y)=\arg\min_x\left[\lambda w|x|+\frac{1}{2}(x-y)^2\right] \tag{31}$$

is given by

$$S_{\lambda w}(y)=\begin{cases}y-\lambda w & \text{if } y>\lambda w,\\ y+\lambda w & \text{if } y<-\lambda w,\\ 0 & \text{otherwise.}\end{cases} \tag{32}$$

Therefore, $\mathbf{E}_{k+1}$ can be updated by

$$\mathbf{E}_{k+1}=S_{\lambda\mathbf{M}/\mu_k\eta}\left\{\mathbf{E}_k-\frac{1}{\eta}[\mathbf{E}_k-(\mathbf{X}-\mathbf{A}\mathbf{Z}_{k+1}+\mathbf{Y}_{1,k}/\mu_k)]\right\}. \tag{33}$$

In summary, the complete algorithm of solving problem Eq. (14) is outlined in Algorithm 2.

Algorithm 2: Solving problem Eq. (14) by inexact ALM.

Input: Data matrix $\mathbf{X}$, parameters λ, p, q and $\eta > \frac{1}{2}$;

Initialize: $\mathbf{Z}_0 = \mathbf{J}_0 = 0$, $\mathbf{E}_0 = 0$, $\mathbf{Y}_{1,0} = 0$, $\mathbf{Y}_{2,0} = 0$, $\mu_0 = 10^{-6}$, $\mu_{\max} = 10^{10}$, $\rho_0 = 1.1$, $\varepsilon = 10^{-8}$, $k = 0$;

While not converged **do**

Step 1: Fix the others and update $\mathbf{J}_{k+1}$ by Eq. (17);

Step 2: Fix the others and update $\mathbf{Z}_{k+1}$ by Eq. (18);

Step 3: Fix the others and update $\mathbf{E}_{k+1}$ by Eq. (19);

Step 4: Update the multipliers by $\mathbf{Y}_{1,k+1} = \mathbf{Y}_{1,k} + \mu_k(\mathbf{X} - \mathbf{A}\mathbf{Z}_{k+1} - \mathbf{E}_{k+1})$, $\mathbf{Y}_{2,k+1} = \mathbf{Y}_{2,k} + \mu_k(\mathbf{Z}_{k+1} - \mathbf{J}_{k+1})$;

Step 5: Update the parameter μ by $\mu_{k+1} = \min(\mu_{\max}, \rho\mu_k)$;

Step 6: Update k by $k = k + 1$;

Step 7: Check the convergence condition $\|\mathbf{X} - \mathbf{A}\mathbf{Z} - \mathbf{E}\|_\infty < \varepsilon$ and $\|\mathbf{Z} - \mathbf{J}\|_\infty < \varepsilon$;

End while

Output: Optimal solution $(\mathbf{Z}^*, \mathbf{E}^*)$.

3.4 *Constructing an Affinity Matrix*

We consider constructing an affinity matrix $\mathbf{W}$ from the coefficient matrix $\mathbf{Z}^*$. Different from the traditional method using Eq. (4), we utilize the postprocessing technique that considers the angular information.[28] Let the skinny SVD of $\mathbf{Z}^*$ be $\mathbf{Z}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{V}^*)^T$; we define $\mathbf{H} = \mathbf{U}^*(\Sigma^*)^{1/2}$, and then $\mathbf{W}$ is obtained by the following equation:

$$\mathbf{W}_{ij} = [(\mathbf{H}\mathbf{H}^T)_{ij}]^{2\alpha}. \tag{34}$$

The obtained $\mathbf{W}$ is positive and can increase the separation of data points from different subspaces. A large α may break the affinities between data points of same subspace; thus, we simply set $\alpha = 2$ in this work. We apply a spectral clustering algorithm on $\mathbf{W}$ to cluster data into its own subspace. Algorithm 3 summarizes the complete subspace clustering algorithm of WSPQ.

Algorithm 3: The WSPQ algorithm.

Input: Data matrix $\mathbf{X}$;

Step 1: Solve the following problem by Algorithm 2: $\min_{\mathbf{Z}}(\|\mathbf{Z}\|^p_{w,S_p} + \lambda\|\mathbf{E}\|_q)$ s.t. $\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E}$, and obtain the optimal solution $(\mathbf{Z}^*, \mathbf{E}^*)$;

Step 2: Compute the skinny SVD: $\mathbf{Z}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{V}^*)^T$;

Step 3: Calculate: $\mathbf{H} = \mathbf{U}^*(\mathbf{\Sigma}^*)^{1/2}$;

Step 4: Construct the affinity matrix $\mathbf{W}$ by Eq. (34);

Step 5: Apply spectral clustering algorithm[18] on $\mathbf{W}$;

Output: The clustering results.

3.5 *Convergence Analysis*

This section discusses the convergence of Algorithm 2. When the objective function is smooth, the convergence of the exact ALM algorithm has been proven in the literature.[43] However, for the inexact ALM algorithm, the proof of convergence is complicated and difficult. In fact, it has been proven that the inexact ALM is convergent when two variable matrices iterate alternately.[39,44] Since the objective function of the proposed WSPQ is not smooth and the optimization algorithm (Algorithm 2) includes three iterating variable matrices ($\mathbf{J}$, $\mathbf{Z}$, and $\mathbf{E}$), it would be not easy to prove the convergence in theory. Fortunately, according to the theoretical analysis,[45] there are some sufficient conditions (but may not be necessary) to guarantee the convergence of Algorithm 2.

(1) The penalty parameter μ is upper bounded.
(2) The dictionary matrix $\mathbf{A}$ is of full column rank.
(3) The error $\Delta_k = \|(\mathbf{Z}_k, \mathbf{J}_k) - (\mathbf{Z}, \mathbf{J})\|$ is monotonically decreasing, where $(\mathbf{Z}_k, \mathbf{J}_k)$ is the solution of the k'th iteration and $(\mathbf{Z}, \mathbf{J})$ is the real solution by solving $\min L(\mathbf{Z}, \mathbf{J}, \mathbf{E}, \mathbf{Y}_1, \mathbf{Y}_2, \mu)$.

Condition 1 is guaranteed by step 5 in Algorithm 2. Similar to LRR, we replace the dictionary $\mathbf{A}$ by its orthogonal basis, which not only reduces the complexity of the algorithm but also guarantees condition 2. Condition 3 cannot be proven directly, but the convexity of the Lagrange function guarantees that Δ_k is monotonically decreasing to a certain extent.

4 Experiments

In this section, we evaluate the effectiveness of the proposed WSPQ method on two real-world applications of subspace clustering: face clustering and motion segmentation. We conduct face clustering experiments on the extended Yale B and AR dataset and motion segmentation experiments on the Hopkins155 dataset. WSPQ is compared to several related subspace clustering methods, including principal component analysis (PCA), local subspace affinity (LSA),[15] spectral curvature clustering (SCC),[16] least squares regression (LSR),[46] low rank subspace clustering (LRSC),[47] Schatten-p norm regularized LRR (SPM),[40] SSC,[21] and LRR.[28] All the experiments are conducted using MATLAB® 2012a in a laptop with Intel Core i7 4710MQ CPU and 8GRAM.

4.1 *Settings*

For WSPQ, we use the data matrix as the dictionary matrix, namely $\mathbf{A} = \mathbf{X}$. The step factor ρ in the inexact ALM method influences the convergence speed, which is set to 1.1. The convergence threshold ε is set to 10^{-8} for face clustering and 10^{-5} for motion segmentation. We also present the grid search method to select the parameters of WSPQ. Specially, the parameters p, q, and λ are learned by searching the grid from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, $\{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$, and $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$, respectively.

For comparison methods, the source codes are downloaded or provided by the authors, and the algorithm settings

are obtained according to the publications. Specifically, for LSR, we use the method described in Theorem 6.[46] For LRSC, we use the method described in Lemma 1[47] for motion segmentation and an ALM variant for face clustering. For SSC, we use the sparse outlying entries variation for face clustering and the noisy variation for motion segmentation. For LRR, we report the results for both cases of LRR (without postprocessing) and LRRH (with postprocessing). For SPM, we report the results with postprocessing. We carefully tune the parameters of the comparison methods and report their best results.

Two popular evaluation metrics, clustering error and confusion matrix, are used to measure the performance of different subspace clustering methods. The clustering error is defined as follows:

$$\text{error} = \frac{N_{\text{error}}}{N_{\text{total}}}, \tag{35}$$

where N_{error} represents the number of incorrectly assigned samples and N_{total} represents the number of total samples. The smaller the clustering error is, the better the clustering quality is.

Assume that we have already known the number of clusters k, then $\mathbf{C} \in \mathbf{R}^{k\times k}$ is defined as the confusion matrix, where $\mathbf{C}_{ij}$ is obtained by first counting the number of samples that are clustered into cluster j but actually belonged to cluster i and then dividing them by the number of i'th cluster samples. Each column and row of the confusion matrix $\mathbf{C}$ represents the predicted cluster and actual cluster, respectively, and thus can reveal more accurate index and structural information.

4.2 *Face Clustering on the Extended Yale B Dataset*

It has been shown that, under the Lambertian assumption, the set of face images of multiple subjects with varying illumination lies close to a union of 9D subspaces.[48] Thus, face clustering is a typical application of subspace clustering.

In this section, we evaluate the clustering performance of WSPQ on the extended Yale B dataset. The extended Yale B dataset is a benchmark for the face clustering problem, which consists of 2432 human face images of 38 subjects. Each subject contains 64 face images taken under different illuminations, and each image is cropped into 192×148 pixels. To reduce the computational cost of all algorithms, we resize each image to 48×42 pixels in our experiment. Thus, each 2016D vectorized face image is treated as a data point. Some sample images from the extended Yale B dataset are shown in Fig. 3(a).

4.2.1 *Experiment under different illuminations*

We design three experiments on the extended Yale B dataset to demonstrate the effectiveness of the proposed algorithm. In the first experiment, we randomly select K (ranging from 2 to 10) subjects from the 38 subjects in the whole dataset. For each given cluster number K, we conduct 20 test runs on different randomly chosen subjects and compute the clustering results by averaging the clustering errors from the 20 tests.

Table 1 shows the average results of all the algorithms. As we can see, the average clustering error of WSPQ reaches 3.62% and outperforms the other algorithms. Meanwhile, as the number of subjects increases, the clustering errors of WSPQ remain at a low level, while those of other

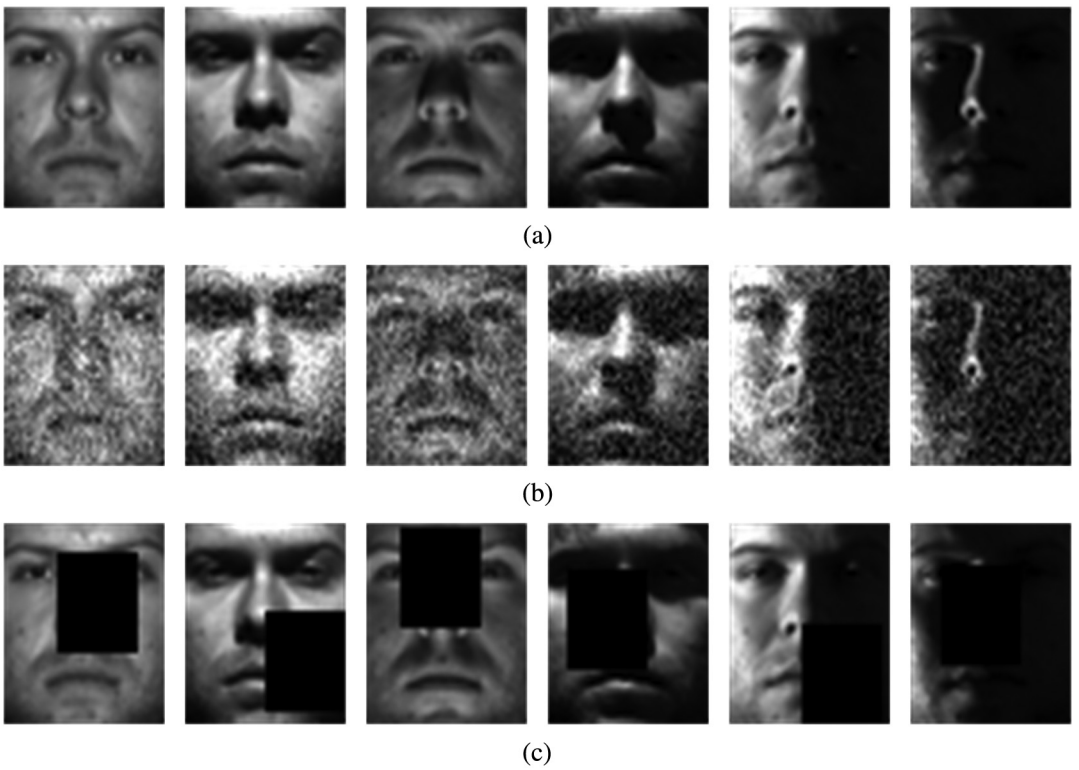


Fig. 3 Some sample images from the extended Yale B dataset. (a) Some original face images under different illuminations; (b) some corrupted face images with Gaussian noise; and (c) some corrupted face images with random block occlusion.

Table 1 Clustering errors (%) of all the algorithms on the extended Yale B dataset.

Subjects	Algorithms									
	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
2	38.83	43.65	18.54	4.02	3.54	4.29	2.14	1.92	1.74	**1.07**
4	61.09	54.96	40.29	11.67	5.41	6.41	4.53	3.57	3.29	**2.53**
6	68.26	58.78	61.22	28.19	13.32	14.82	7.24	5.89	4.80	**2.81**
8	67.91	59.37	65.19	39.97	27.84	22.10	14.72	8.07	6.42	**4.67**
10	71.17	60.12	74.62	42.06	32.43	29.73	15.38	10.74	11.91	**7.02**
Avg	61.45	55.38	51.97	25.18	16.51	15.47	8.80	6.04	5.63	**3.62**

algorithms increase in different degrees. The results show that the proposed WSPQ is effective for different illuminations. Since the sparse representation is also an efficient data representation, SSC obtains relatively low clustering errors. We also observe that SPM performs better than LRRH, which indicates that the Schatten-p norm is more suitable than the nuclear norm for estimating the rank. Moreover, the average clustering error of LRRH is 8.80%, while that of LRR is 15.47%, which shows that the postprocessing technique can improve the clustering performance. Obviously, the clustering errors of LSA and SCC are much larger than other algorithms. This is because they do not explicitly deal with data noise.

4.2.2 *Experiment with Gaussian noise*

In the second experiment, we evaluate the robustness of WSPQ to corrupted face images with Gaussian noise. The first 10 subjects of the extended Yale B dataset are used in the experiment, and we randomly select some face images to be corrupted using Gaussian noise with zero mean and deviation of 0.1. The percentage of corrupted face images is gradually increased from 20% to 100%, in steps of 20%. Figure 3(b) shows some corrupted sample images with Gaussian noise.

Table 2 shows the clustering errors of all the algorithms with Gaussian noise. It can be seen that WSPQ is obviously superior to comparison algorithms among all the cases. As the number of corrupted face image increases, the performance of all algorithms declines in different degrees. In the case of corrupting all face images, the clustering error of WSPQ is 27.34%, which improves by 5.10% over the best result provided by SPM. This demonstrates that WSPQ is more robust to Gaussian noise than the other algorithms. We also observe that SSC performs worse than LRRH, which indicates that the sparse model is sensitive to Gaussian noise.

4.2.3 *Experiment with block occlusion*

In the third experiment, we further verify the robust performance of WSPQ against block occlusion. We corrupt some face images from the first 10 subjects of the extended Yale B dataset with a fixed black block occlusion of size 24×21 pixels, and the position of the blocks is randomly chosen. Following the strategy in Sec. 4.2.2, we experiment on different percentages of corrupted face images varying from 20% to 100%, in steps of 20%. Figure 3(c) shows some corrupted sample images with random block occlusion.

The clustering errors of all the algorithms with random block occlusion are reported in Table 3. As can be seen, WSPQ outperforms all the comparison algorithms among all the cases, which again verifies the robustness of WSPQ. As the corrupted face image increase, the clustering

Table 2 Clustering errors (%) of all the algorithms on the extended Yale B dataset with Gaussian noise.

Corruption (%)	Algorithms									
	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
20	71.88	64.11	71.02	42.22	37.69	34.61	20.81	18.70	31.03	**14.88**
40	73.75	65.08	74.14	42.45	38.44	36.05	22.09	19.13	37.91	**15.42**
60	72.66	67.34	79.91	43.10	39.05	40.17	25.42	24.72	47.04	**17.69**
80	72.50	67.89	81.09	42.11	39.66	42.47	33.92	29.45	47.27	**25.09**
100	71.72	73.36	81.75	44.56	40.42	43.89	34.38	33.91	51.17	**27.34**
Avg	72.50	67.56	77.58	42.89	39.05	39.44	27.32	25.18	42.88	**20.08**

Table 3 Clustering errors (%) of all the algorithms on the extended Yale B dataset with block occlusion.

Occlusion (%)	Algorithms									
	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
20	72.66	70.16	75.55	47.19	46.33	50.02	33.28	30.27	31.88	**21.25**
40	74.75	77.66	77.19	49.38	48.67	51.17	34.38	33.52	49.14	**29.75**
60	75.38	80.71	81.02	66.35	60.78	61.72	54.06	55.81	61.78	**50.41**
80	75.06	83.98	82.11	72.03	73.05	73.05	68.52	67.92	70.47	**66.12**
100	76.75	84.69	83.20	79.92	78.83	85.08	80.23	77.37	77.03	**75.89**
Avg	74.92	79.44	79.81	62.97	61.53	64.21	54.09	52.98	58.08	**48.68**

task becomes more challenging, and all the algorithms obtain very high clustering errors. In particular, when all face images are corrupted, the clustering error of all the algorithms is nearly 80%. The reason is that many useful features in the human face (month, nose, and eyes) may be occluded by the random block. The more corrupted the face images are, the less useful the information that can be obtained is.

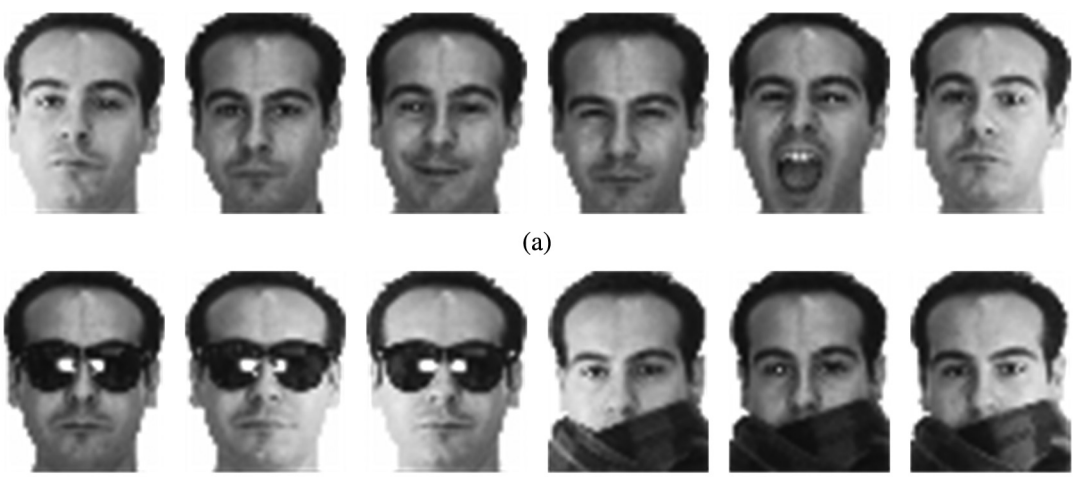
4.3 *Face Clustering on the AR Dataset*

In this section, we further evaluate different subspace clustering algorithms on the AR dataset. This dataset consists of over 4000 human face images of 126 subjects (70 men and 56 women). For each subject, 26 face images were taken in two sessions, separated by two weeks (14 days). These images feature frontal view faces with different facial expressions (neutral, smile, anger, and scream), illumination conditions (left, right, and all side lights on), and occlusions (sunglasses and scarf). In our experiment, we make use of the first 20 subjects and resize each image to 40×50 pixels. Some sample images from the AR dataset are shown in Fig. 4.

4.3.1 *Experiment without occlusion*

In this experiment, we evaluate the performance of WSPQ on the AR dataset without occlusion. For each subject, eight face images with different facial expressions and six face images with varying illuminations are selected for clustering tasks. Similar to the experiments on the extended Yale B dataset, we also randomly select K (ranging from 2 to 10) subjects from the first 20 subjects in the whole dataset. For each given cluster number K, the clustering results are based on 20 test runs on different randomly chosen subjects.

The average results of all the algorithms are shown in Table 4. Obviously, WSPQ obtains the best experimental results among all the cases, verifying the effectiveness of WSPQ on face clustering. The clustering errors of LSR and SSC reach 8.16% and 8.33%, respectively, which are lower than the four low rank algorithms (LRSC, LRR, LRRH, and SPM). The reason is that conventional low rank models fail to accurately approximate the rank in cases of large variations. WSPQ obtains better results due to its superior rank approximation and adaptive noise representation. In addition, LRR and LRRH have similar



(a)

(b)

Fig. 4 Some sample images from the AR dataset. (a) Some face images without occlusion; (b) some face images with real occlusion.

Table 4 Clustering errors (%) of all the algorithms on the AR dataset without occlusion.

	Algorithms									
Subjects	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
2	21.57	37.57	13.15	3.11	4.54	2.94	3.05	2.29	2.21	**1.65**
4	36.93	40.95	25.71	5.79	7.90	4.11	4.92	4.03	3.67	**3.04**
6	40.38	41.17	27.29	9.83	17.61	8.73	9.11	7.89	7.45	**5.33**
8	39.61	63.25	40.07	7.98	39.04	14.72	16.92	15.21	10.43	**7.19**
10	40.43	65.82	42.91	14.08	39.85	23.51	24.18	19.46	17.91	**12.72**
Avg	35.78	49.75	29.83	8.16	21.79	10.81	11.64	9.78	8.33	**5.98**

performances, which illustrates that the postprocessing technique cannot work well on the AR dataset.

4.3.2 *Experiment with real occlusion*

To validate the robustness of WSPQ to different noises, we further test WSPQ on the AR dataset with real occlusion. All 26 face images of each subject are used in this experiment, including 14 images with different expressions and illuminations and 12 images with real occlusion (sunglasses and scarf). Since each subject has few sample images and the data noise is complex, it is challenging for clustering algorithms.

Following the strategy in Sec. 4.3.1, we report the average results of all the algorithms in Table 5. As can be seen, similar to previous experimental results, the proposed WSPQ is superior to the other algorithms. The results demonstrate again that WSPQ works well in face clustering.

Table 5 Clustering errors (%) of all the algorithms on the AR dataset with real occlusion.

	Algorithms									
Subjects	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
2	28.46	40.26	19.74	9.06	14.59	12.32	10.76	8.83	9.53	**7.97**
4	54.62	44.10	35.97	15.64	22.95	21.31	18.47	14.69	17.73	**11.57**
6	53.85	45.94	36.75	25.77	36.62	34.15	29.16	23.10	29.02	**19.74**
8	62.59	69.21	48.40	28.41	51.08	40.23	35.94	33.43	37.17	**25.05**
10	67.77	74.09	52.71	33.83	52.71	42.01	39.37	38.05	41.52	**28.92**
Avg	54.46	54.72	38.71	22.54	35.59	30.01	26.74	23.62	27.01	**18.65**

Fig. 5 Three example frames with the extracted feature from the Hopkins155 dataset.

Table 6 Clustering errors (%) of all the algorithms on the Hopkins155 dataset.

Algorithms	PCA	LSA	SCC	LSR	LRSC	LRR	LRRH	SPM	SSC	WSPQ
					Two motions					
Mean	16.13	3.64	2.15	2.49	3.33	3.23	1.33	1.39	1.53	**1.06**
Std	15.53	7.96	7.31	7.27	8.26	8.25	4.48	4.56	6.42	**4.14**
Max	49.79	40.88	48.83	39.49	40.34	41.16	**33.30**	35.51	47.18	**33.30**
					Three motions					
Mean	23.44	7.13	8.01	5.93	6.97	8.20	2.51	2.47	4.40	**1.97**
Std	17.58	9.47	13.74	9.88	8.83	10.11	5.20	5.04	9.33	**4.87**
Max	58.03	44.98	52.83	38.29	31.12	36.17	26.81	25.49	41.25	**24.93**

4.4 *Motion Segmentation on the Hopkins155 Dataset*

Given a set of two-dimensional point trajectories extracted from a video sequence, motion segmentation refers to the task of clustering the trajectories into groups according to their underlying motions. Under the affine projection model, the trajectories associated with a single rigid motion lie in an affine subspace.[49] Therefore, the trajectories of n rigid motions lie in a union of n affine subspaces, and the motion segmentation problem reduces to a subspace clustering problem.

In this section, we evaluate the proposed WSPQ for the motion segmentation task on the Hopkins155 dataset. This dataset consists of 155 video sequences, where 120 sequences have two motions and 35 sequences have three motions. Each sequence is a separate clustering task, so there are 155 clustering tasks in total. Three example frames with the extracted feature are shown in Fig. 5.

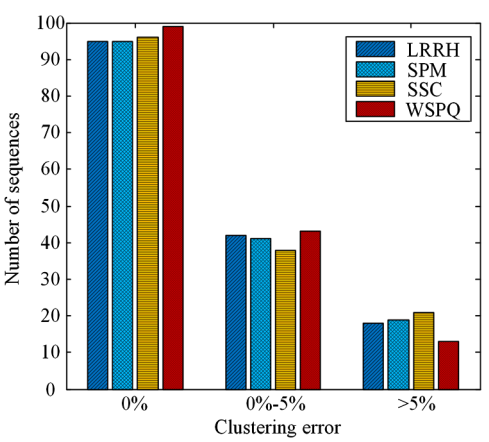


Fig. 6 The distribution of the clustering errors of LRRH, SPM, SSC, and WSPQ on the Hopkins155 dataset.

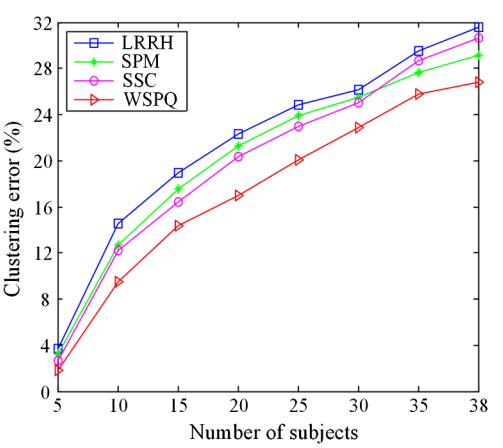


Fig. 8 The clustering performance of LRRH, SPM, SSC, and WSPQ with different numbers of subjects.

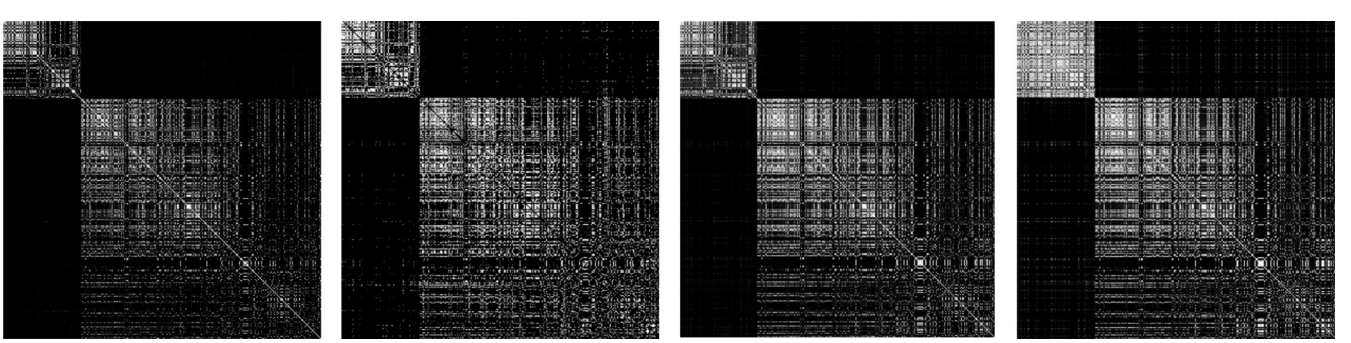


Fig. 7 The affinity matrix of one example sequence. From left to right: WSPQ, SSC, SPM, and LRRH.

We report the mean, standard deviation, and maximum of clustering errors over the whole dataset in Table 6. From the results listed in Table 6, we observe that WSPQ outperforms the other algorithms. Since LRRH, SPM, SSC, and WSPQ achieve competitive results, we further compare the distribution of the clustering errors of them, and the results are shown in Fig. 6. As we can see, WSPQ obtains accurate clustering results for most sequences, and nearly 100 sequences are clustered accurately.

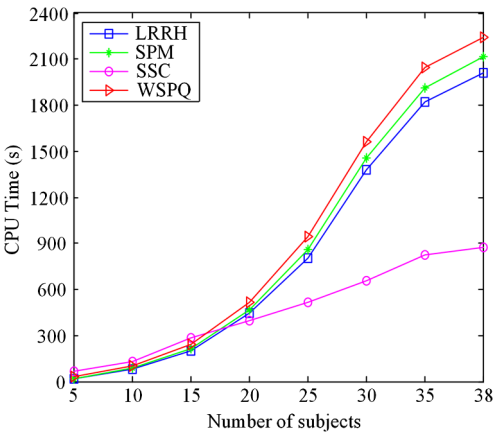


Fig. 9 The CPU time of LRRH, SPM, SSC, and WSPQ with different numbers of subjects.

We next show the affinity matrices of one example sequence, i.e., the 40th sequence obtained by LRRH, SPM, SSC, and WSPQ in Fig. 7. Clearly, there are two diagonal blocks in each affinity matrix, which indicates that the example sequence has two motions. If the number of nonzero entries lying outside the diagonal blocks is small, the spectral clustering algorithm can obtain a better result. The matrix of the proposed WSPQ obviously has bright diagonal blocks and is clearer than others outside the diagonal blocks. This demonstrates the superiority of our algorithm in affinity matrix construction.

4.5 *Large-Scale Subspace Clustering*

In this section, we carry out an in-depth study for large-scale subspace clustering. Instead of using at most 10 subjects to perform clustering experiments in Secs. 4.2 and 4.3, we conduct experiments on the whole extended Yale B dataset with more subjects. We range the number of subjects from 5 to 38 and report the results of four competitive algorithms (LRRH, SPM, SSC, and WSPQ). The clustering performance and CPU time are shown in Figs. 8 and 9, respectively.

From Fig. 8, we can see that the four algorithms all perform worse as the number of subject increases and the proposed WSPQ always achieves the best results no matter how many subjects are used. From Fig. 9, it is observed that WSPQ is more time consuming than LRR and SPM. Although the computational cost of WSPQ is a little higher,

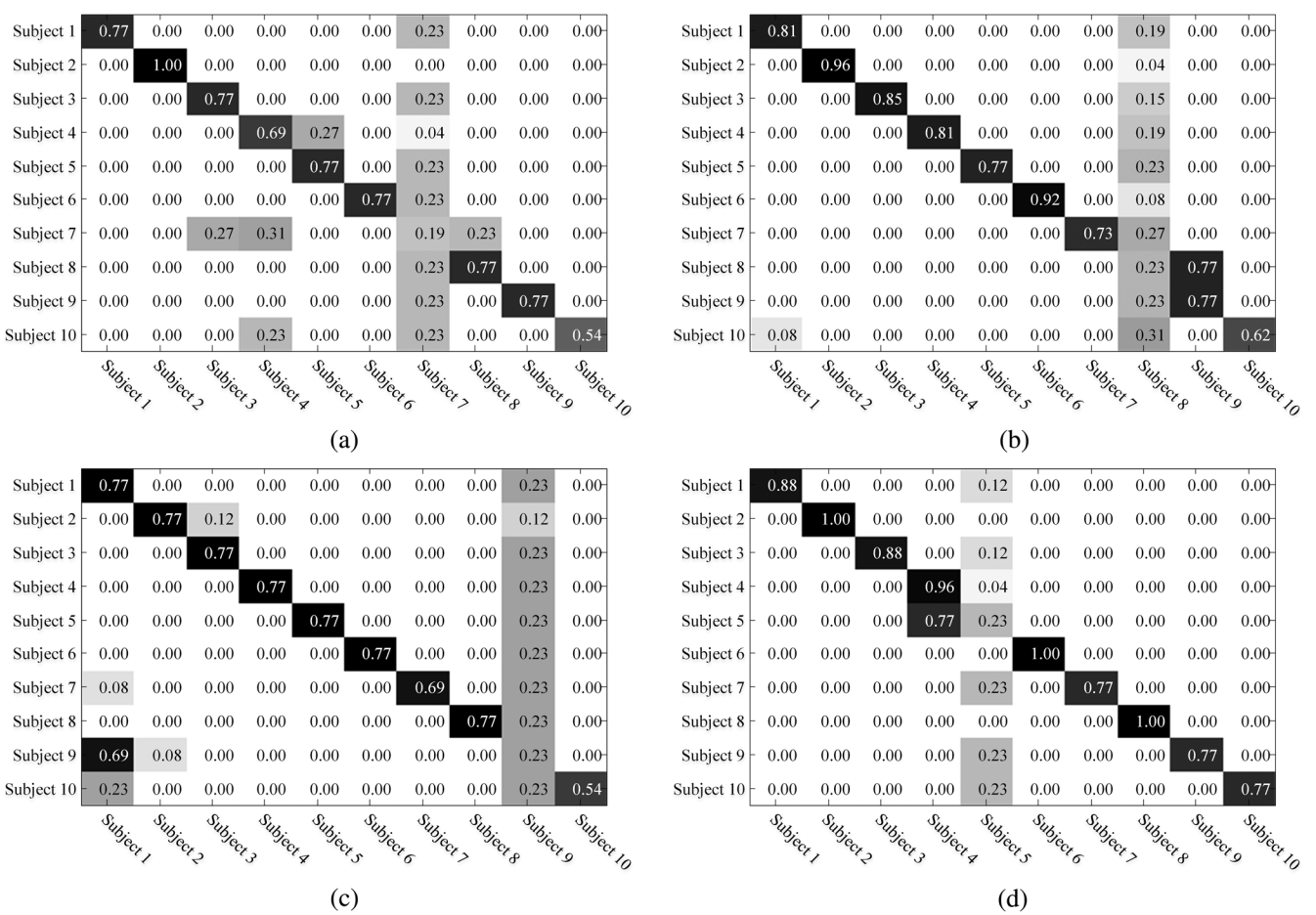


Fig. 10 The confusion matrices of (a) LRRH, (b) SPM, (c) SSC, and (d) WSPQ on the AR dataset.

it can obtain much performance improvement. In addition, the growth rate of SSC is much slower than that of LRR variant algorithms (LRRH, SPM, and WSPQ); thus, it runs fast when more subjects are used.

4.6 *Evaluation by Confusion Matrix*

To examine the performance of different subspace clustering algorithms more carefully, the confusion matrix is presented in this section. The confusion matrix can not only show the clustering accuracy from different angles but also reveal more structural information. We use the first 10 subjects of the AR dataset with real occlusion and report the results of LRRH, SPM, SSC, and WSPQ in this experiment. The confusion matrices of the four algorithms are shown in Fig. 10. It can be seen that the proposed WSPQ again achieves more clustering accuracy and outperforms other algorithms, which demonstrates the effectiveness of both weighted Schatten-p norm and L_q norm.

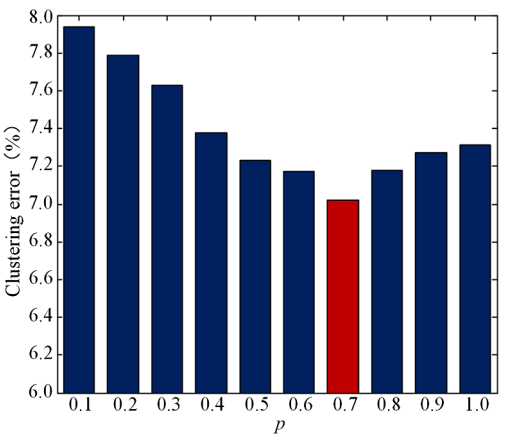


Fig. 11 The clustering performance with the varied parameter p.

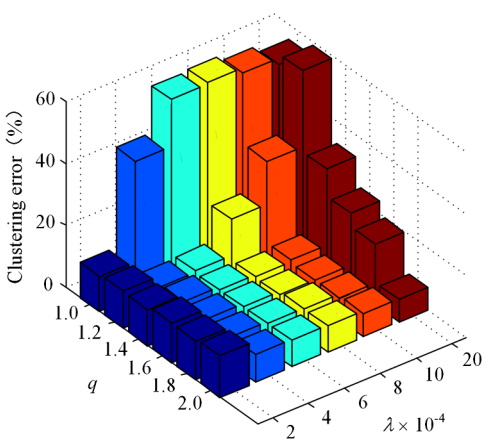


Fig. 12 The clustering performance with the varied parameters q and λ.

4.7 *Parameter Selection*

The proposed WSPQ includes three important parameters (p, q, and λ), where p is used to control the low rank term and q and λ are used to select a suitable noise model and balance the effect of noise, respectively. In this section, we experiment on the extended Yale B dataset (in the case of 10 subjects) to study the effectiveness of the parameters in WSPQ.

First, we range the value of p from 0.1 to 1.0; the experimental results are shown in Fig. 11. It is observed that WSPQ achieves the best results when $p = 0.7$ and the clustering performance is acceptable when p is not set to a small value. In theory, as p decreases, the weighted Schatten-p norm is getting closer to the rank function, so it seems that we can achieve better results with a smaller p. However, the input data contain various noises, which may destroy the low rank structure in different degrees. Taking a small value of p means a strict rank approximation, and it is not suitable when the low rank property is weak. Actually, according to different input data, we can always get an appropriate rank approximation by varying the value of p. In addition, the parameters q and λ both control the noise term. Thus, we do parameter selection by varying q and λ simultaneously, and the experimental results are shown in Fig. 12. As can be seen, WSPQ can achieve promising results when $q \in [1.4, 2.0]$ and $\lambda \in [4 \times 10^{-4}, 10 \times 10^{-4}]$.

4.8 *Convergence Verification*

In this section, we empirically verify the convergence of the proposed WSPQ. We first construct five independent subspaces $\{S_i\}_{i=1}^{5} \subset R^{100}$ and then sample 40 data vectors from each subspace. We randomly select 20% of samples to be corrupted by adding Gaussian noise with zero mean and deviation of $0.2\|\mathbf{X}\|_F$, where $\mathbf{X}$ is the created data matrix. After each iteration, we record the relative error

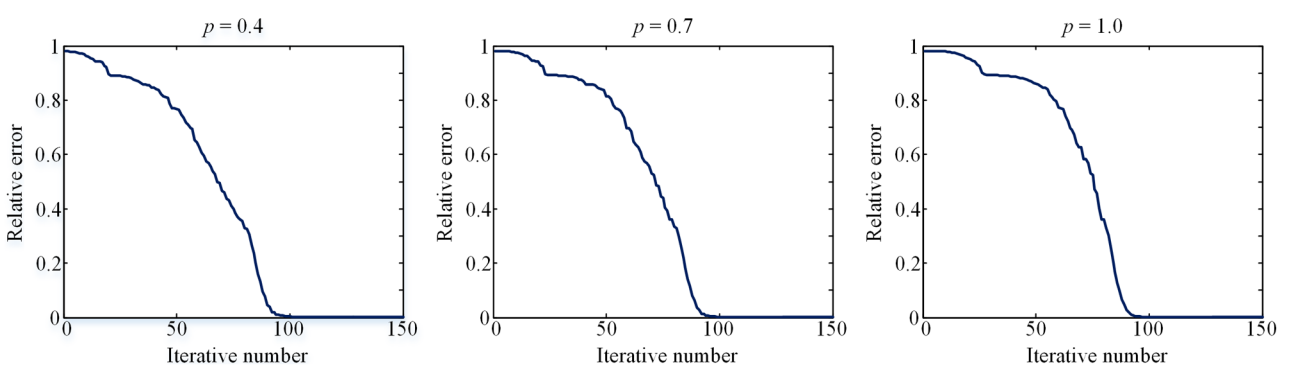


Fig. 13 The convergence curves of WSPQ.

$\frac{\|\mathbf{X}-\mathbf{AZ}-\mathbf{E}\|_F}{\|\mathbf{X}\|_F}$. Figure 13 has shown the convergence curves of WSPQ when $p = 0.4$, 0.7, and 1.0. We can see that WSPQ reaches convergence within 100 iterations.

5 Conclusions

This paper proposes a subspace clustering algorithm, named WSPQ, for face clustering and motion segmentation. Specifically, WSPQ uses the weighted Schatten-p norm and L_q norm to model the low rankness and noise in LRR, respectively. The weighted Schatten-p norm can better approximate the rank function, and the L_q norm can handle complex noise well. Moreover, an efficient algorithm based on the inexact ALM method is developed to solve the optimization problem. Using a postprocessing technique, the affinity matrix is constructed and used for the spectral clustering algorithm. Experimental results on two face datasets and one motion segmentation dataset demonstrate the superiority of the proposed WSPQ over the state-of-the-art algorithms. Our future work will introduce the manifold model into WSPQ, so the global and local structures of data can be captured simultaneously. Another important work is how to select the parameters adaptively. We need to do more theoretical research on it.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61473154) and the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Grant No. 30920140122007).

References

1. Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory* **55**(11), 5302–5316 (2009).
2. W. Hong et al., "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Process.* **15**(12), 3655–3671 (2006).
3. J. Ho et al., "Clustering appearances of objects under varying illumination conditions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 1, pp. 11–18, IEEE (2003).
4. P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1801–1807, IEEE (2011).
5. S. Rao et al., "Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(10), 1832–1845 (2010).
6. S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," in *Proc. of the Int. Conf. on Computer Vision*, pp. 4453–4461, IEEE (2015).
7. X. Shen, Q. Sun, and Y. Yuan, "A unified multiset canonical correlation analysis framework based on graph embedding for multiple feature extraction," *Neurocomputing* **148**, 397–408 (2015).
8. R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.* **28**(2), 52–68 (2011).
9. P. Tseng, "Nearest q-flat to m points," *J. Optim. Theory Appl.* **105**(1), 249–252 (2000).
10. T. Zhang, A. Szlam, and G. Lerman, "Median k-flats for hybrid linear modeling with many outliers," in *Proc. of the IEEE 12th Int. Conf. on Computer Vision Workshops*, pp. 234–241, IEEE (2009).
11. J. P. Costeria and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vision* **29**(3), 159–179 (1998).
12. R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(12), 1945–1959 (2005).
13. M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.* **11**(2), 443–482 (1999).
14. M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* **24**(6), 381–395 (1981).
15. J. Yan and M. Pollefeys, "A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. of the European Conf. on Computer Vision*, pp. 94–106, Springer, New York (2006).
16. G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vision* **81**(3), 317–330 (2009).
17. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 888–905 (2000).
18. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856, MIT, Cambridge, Massachusetts (2002).
19. E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory* **51**(12), 4203–4215 (2005).
20. R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B* **73**, 267–288 (1996).
21. E. Elhamifar and R. Vidal, "Sparse subspace clustering: algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2765–2781 (2013).
22. X. Shen and Q. Sun, "A novel semi-supervised canonical correlation analysis and extensions for multi-view dimensionality reduction," *J. Visual Commun. Image Represent.* **25**(8), 1894–1904 (2014).
23. E. J. Candes et al., "Robust principal component analysis," *J. ACM* **58**(3), 1–37 (2011).
24. E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.* **9**(6), 717–772 (2009).
25. X. Shen et al., "Semi-paired discrete hashing: learning latent hash codes for semi-paired cross-view retrieval," *IEEE Trans. Cybern.* **PP**, 1–14 (2016).
26. Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.* **24**(11), 4322–4334 (2015).
27. Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. of the Int. Conf. on Data Mining*, pp. 110–119, IEEE (2014).
28. G. Liu et al., "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 171–184 (2013).
29. M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proc. of the American Control Conf.*, pp. 2156–2162, IEEE (2003).
30. Y. Hu et al., "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(9), 2117–2130 (2013).
31. F. Nie, H. Huang, and C. H. Ding, "Low-rank matrix recovery via efficient Schatten p-norm minimization," in *AAAI Conf. on Artificial Intelligence*, pp. 655–661 (2012).
32. L. Liu, W. Huang, and D. R. Chen, "Exact minimum rank approximation via Schatten p-norm minimization," *J. Comput. Appl. Math.* **267**, 218–227 (2014).
33. F. Nie and H. Huang, "Subspace clustering via new low-rank model with discrete group structure constraint," in *Proc. of the 25th Int. Joint Conf. on Artificial Intelligence*, pp. 1874–1880 (2016).
34. Y. Yu and D. Schuurmans, "Rank/norm regularization with closed-form solutions: application to subspace clustering," in *Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence* (2011).
35. Y. Xie et al., "Weighted Schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.* **25**(10), 4842–4857 (2016).
36. Y. Xie et al., "Hyperspectral image restoration via iteratively regularized weighted Schatten p-norm minimization," *IEEE Trans. Geosci. Remote Sens.* **54**(8), 4642–4659 (2016).
37. Y. Zheng et al., "Practical low-rank matrix approximation under robust L1-norm," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1410–1417, IEEE (2012).
38. Z. Kang, C. Peng, and Q. Cheng, "Robust subspace clustering via smoothed rank approximation," *IEEE Signal Process. Lett.* **22**(11), 2088–2092 (2015).
39. Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC Technical Report UILU-ENG-09-2215 (2009).
40. X. Zhang et al., "Schatten-q regularizer constrained low rank subspace clustering model," *Neurocomputing* **182**, 36–47 (2016).
41. W. Zuo et al., "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. of the Int. Conf. on Computer Vision*, pp. 217–224 (2013).
42. E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for \ell_1-minimization: methodology and convergence," *SIAM J. Optim.* **19**(3), 1107–1130 (2008).
43. D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York (2014).
44. Y. Zhang, "Recent advances in alternating direction methods: practice and theory," in *IPAM Workshop on Continuous Optimization* (2010).
45. J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.* **55**(1–3), 293–318 (1992).

46. C. Lu et al., "Robust and efficient subspace segmentation via least squares regression," in *Proc. of the European Conf. on Computer Vision*, pp. 347–360, Springer, New York (2012).
47. R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognit. Lett.* **43**, 47–61 (2014).
48. R. Basri and D. W. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(2), 218–233 (2003).
49. C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int. J. Comput. Vision* **9**(2), 137–154 (1992).

Tao Zhang received his BS degree in network engineering from Nanjing University of Science and Technology, Nanjing, China, in 2010. Currently, he is a PhD candidate at Nanjing University of Science and Technology. His current research interests include pattern recognition, computer vision, subspace learning, and low rank and sparse theory.

Zhenmin Tang received his PhD degree from Nanjing University of Science and Technology, Nanjing, China. He now is a professor at Nanjing University of Science and Technology. His major research areas include intelligent systems, pattern recognition, and image processing. He is also the leader of several key programs of the National Nature Science Foundation of China.

Qing Liu received his BS degree in information and computing science from Inner Mongolia University and his MS degree in pattern recognition and intelligent systems from Jiangsu University, China, in 2009 and 2013, respectively. He is currently pursuing his PhD in Nanjing University of Science and Technology. His current research interests include image processing and machine learning.