# Adaptive Weighted Sparse Principal Component Analysis for Robust Unsupervised Feature Selection

Shuangyan Yi, Zhenyu He, *Senior Member, IEEE*, Xiao-Yuan Jing, *Senior Member, IEEE*, Yi Li, Yiu-Ming Cheung, *Fellow, IEEE*, and Feiping Nie

***Abstract*— Current unsupervised feature selection methods cannot well select the effective features from the corrupted data. To this end, we propose a robust unsupervised feature selection method under the robust principal component analysis (PCA) reconstruction criterion, which is named the adaptive weighted sparse PCA (AW-SPCA). In the proposed method, both the regularization term and the reconstruction error term are constrained by the $\ell_{2,1}$-norm: the $\ell_{2,1}$-norm regularization term plays a role in the feature selection, while the $\ell_{2,1}$-norm reconstruction error term plays a role in the robust reconstruction. The proposed method is in a convex formulation, and the selected features by it can be used for robust reconstruction and clustering. Experimental results demonstrate that the proposed method can obtain better reconstruction and clustering performance, especially for the corrupted data.**

***Index Terms*— $\ell_{2,1}$-norm, clustering, feature selection, reconstruction.**

## I. Introduction

IN IMAGE processing [1]–[5], machine learning [6]–[10], and object tracking [11]–[13], data are often formed as the high-dimensional feature vectors. Among these high-dimensional features, they are inevitably correlated, redundant, or noisy, which may depress the performance [14]–[16] in reconstruction and clustering. This shows that not all features are valuable for the learning task. Consequently, it is necessary to reduce those features hindering learning tasks by the technique of feature selection.

According to whether labels are used or not, feature selection methods are grouped into two categories: supervised feature selection methods that use labels in training and unsupervised feature selection methods that do not use labels in training. The supervised feature selection methods focus on selecting those essential discriminative features guided by the labels of the training data [17]–[19], while the unsupervised feature selection methods focus on selecting those representative features without the guidance of labels [20]–[22]. Since the labeled data are often expensive to obtain in practical applications, the unsupervised feature selection methods are more important.

Earlier, the unsupervised feature selection methods aim to independently calculate the score of each feature and select the top ranking features according to the calculated scores [23]. The way of score computation for a single feature neglects the feature correlations [24], [25], which may hinder the acquisition of the optimal feature subset. Therefore, many spectral feature selection methods have been proposed to exploit feature correlations [14], [21], [22], [24]–[28]. In detail, the Laplacian score method [24] is proposed to evaluate the importance of features by considering their local correlations. Furthermore, multi-cluster feature selection (MCFS) [25] is proposed to select those features, such that the multi-cluster structure of data can be best preserved. After the exploitation of feature correlations, some unsupervised feature selection methods with discriminative ability [14], [22], [26]–[28] are proposed by introducing pseudo-labels. For example, unsupervised discriminative feature selection (UDFS) [26] designs a discriminative criterion that preserves local discrimination information of the original high-dimensional data in the low-dimensional subspace to select the discriminative features. Similar to the formulation of UDFS, nonnegative discriminative feature selection (NDFS) [27] uses nonnegative spectral analysis to learn the more ideal clustering pseudo-labels, thereby selecting the discriminative features. It can be generalized that the regularization terms of the spectral feature selection methods are often constrained by the $\ell_1$-norm or $\ell_{2,1}$-norm. In fact, $\ell_{2,1}$-norm has often been used in the reconstruction term for enhancing the robustness to outliers [29]–[32], in the regularization term for selecting the effective features, and in

Manuscript received February 4, 2018; revised June 12, 2018, September 29, 2018, January 26, 2019, and June 12, 2019; accepted July 5, 2019. Date of publication August 28, 2019; date of current version June 2, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61672183 and Grant 61772141, in part by the Shenzhen Research Council under Grant JCYJ20170413104556946, Grant JCYJ20160406161948211, Grant JCYJ20160226201453085, and Grant KJYY20170724152625446, in part by the Natural Science Foundation of Guangdong Province under Grant 2015A030313544, in part by the Guangdong Provincial Natural Science Foundation under Grant 17ZK0422, and in part by the Guangzhou Science and Technology Planning Project under Grant201804010347. *(Corresponding authors: Zhenyu He; Xiao-Yuan Jing.)*

S. Yi is with the Institute of Information Technology, Shenzhen Institute of Information Technology, Shenzhen 518172, China, and also with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: shuangyanshuangfei@163.com).

Z. He and Y. Li are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: zhenyuhe@hit.edu.cn; ly-res@.163.com).

X.-Y. Jing is with the State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China, and also with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: jingxy2000@126.com).

Y.-M. Cheung is with the Department of Computer Science, Hong Kong Baptist University (HKBU), Hong Kong, also with the Institute of Research and Continuing Education, Hong Kong Baptist University (HKBU), Hong Kong, and also with the United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai 519000, China (e-mail: ymc@comp.hkbu.edu.hk).

F. Nie is with the Optical Image Analysis and Learning Center, Northwestern Polytechnical University, Xi'an 710000, China (e-mail: feipingnie@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2928755

both the reconstruction term and the regularization term for multi-class classification problem [28].

Generally, the above-mentioned spectral feature selection methods can effectively select the most useful features by discovering the manifold structure of data. However, the learning of manifold structure depends on a graph Laplacian based on original data construction. When the original data contain a large amount of noise, noisy features may hinder the correct construction of graph Laplacian, and then, the spectral feature selection methods may become unstable or invalid [22], [33]. To this end, robust spectral feature selection methods [14], [22], [33] are proposed to improve the robustness to outliers in feature selection. More specifically, robust unsupervised feature selection (RUFS) method [14] and robust spectral learning for unsupervised feature selection (RSFS) [22] are proposed to consider the data noise in the learning of pseudo-cluster labels. The structured optimal graph feature selection method (SOGFS) [33] is proposed to adaptively learn a robust graph Laplacian. However, these robust spectral feature selection methods are robust to outliers only when the data are corrupted slightly. This is because they do not have the reconstruction term and can only select the effective features from the original data. In fact, when the original data are heavily corrupted, these spectral feature selection methods will select many noisy features inevitably. In addition, all spectral feature selection methods are basically in the same mode, that is, they combine graph embedding and sparse spectral regression to evaluate the effectiveness of features. Hence, these models are usually complex and include more than one parameter.

In order to construct a simple yet effective feature selection method, we propose to select the useful features from a perspective of robust principal component analysis (PCA) reconstruction. More specifically, we first establish the relationship between optimal mean robust PCA (OMRPCA) [30] and feature selection by imitating the self-contained regression type of PCA [34] and obtain the self-contained regression type of OMRPCA. By relaxing for the self-contained regression type of OMRPCA, we obtain its non-convex formulation. In this way, robust PCA methods (e.g., OMRPCA) and the technique of feature selection can be successfully connected. Furthermore, we make a change of variable for the non-convex formulation so that the formulation after variable substitution is convex. For simplicity, the convex formulation is named the adaptive weighted sparse PCA (AW-SPCA). The main contributions of this paper include the followings.

1) We propose an RUFS method based on a robust PCA reconstruction criterion. The proposed method is in a convex formulation and can obtain the global optimal solution.
2) We propose that when the data are corrupted heavily, the effective features should be chosen from the reconstructed data.

The rest of this paper is organized as follows. The preliminaries are introduced in Section II. The proposed method and its theoretical analyses are introduced in Section III and IV, respectively. The experiments are performed in Section V to demonstrate the effectiveness of the proposed method. Finally, a conclusion is given in Section VI.

## II. Preliminaries

In this section, the concept of principal component is first expressed, and then, the essential relationship between PCA, the regression type of PCA, and the self-contained regression type of PCA is explained.

### A. Principal Component Analysis

For a matrix $\boldsymbol{A} \in \mathbb{R}^{l\times k}$, we denote the $(i, j)$th element of $\boldsymbol{A}$ by $a_{ij}$ and the $j$th column of $\boldsymbol{A}$ by $\boldsymbol{a}_j$. The $\ell_{2,1}$-norm and Frobenius-norm of $\boldsymbol{A}$ in this paper are defined as $\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{k} \|\boldsymbol{a}_j\|_2$ and $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{l}\sum_{j=1}^{k} a_{ij}^2$. Given the data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$, where each data point $\boldsymbol{x}_i \in \mathbb{R}^m$ represents a vectorized image and $n$ is the number of data points. Assume that matrix $\boldsymbol{X}$ has been centralized, PCA aims to seek a transformation matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] \in \mathbb{R}^{m\times d}$ with $d \ll m$, such that its reconstruction error is minimized as follows:

$$\min_{\boldsymbol{U}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X}\|_F^2, \quad \text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}. \tag{1}$$

After the optimal transformation matrix $\boldsymbol{U}^*$ is learned, the transformed data are denoted by $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{U}^* \in \mathbb{R}^{n\times d}$, and each column of $\boldsymbol{Z}$ (i.e., $\boldsymbol{z}_j \in \mathbb{R}^n$, $j \in \{1, 2, \ldots, d\}$) is a principal component. From $\boldsymbol{z}_j = \boldsymbol{X}^T\boldsymbol{u}_j^*$, it can be seen that each principal component is a linear combination of all the $m$ original features, whose combination coefficient $\boldsymbol{u}_j^* \in \mathbb{R}^m$ corresponds to the $j$th column of $\boldsymbol{U}^*$.

### B. Self-Contained Regression Type of Principal Component Analysis

Suppose given the principal components of PCA, i.e., $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d]$. Here, $\boldsymbol{z}_j \in \mathbb{R}^n$, $j \in \{1, 2, \ldots, d\}$. The regression type [34] of each principal component $\boldsymbol{z}_j$ can be expressed as

$$\min_{\boldsymbol{c}_j} \|\boldsymbol{z}_j - \boldsymbol{X}^T\boldsymbol{c}_j\|_2^2 \tag{2}$$

where the column vector $\boldsymbol{c}_j \in \mathbb{R}^m$ is the vector of regression coefficients of the principal component $\boldsymbol{z}_j$.

The regression type of all $d$ principal components are calculated as follows:

$$\min_{\boldsymbol{C}} \textstyle\sum_{j=1}^{d} \|\boldsymbol{z}_j - \boldsymbol{X}^T\boldsymbol{c}_j\|_2^2. \tag{3}$$

Write $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_d] \in \mathbb{R}^{m\times d}$. Problem (3) is, therefore, equivalent to the following formulation:

$$\min_{\boldsymbol{C}} \|\boldsymbol{Z} - \boldsymbol{X}^T\boldsymbol{C}\|_F^2. \tag{4}$$

In SPCA [34], each column of $\boldsymbol{C}$ is referred as a loading corresponding to a principal component. Let the singular value decomposition (SVD) of $\boldsymbol{X}$ be $\boldsymbol{X} = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{B}^T$. Then, $\boldsymbol{B}\boldsymbol{\Sigma}^T$ are the principal components [34]. Since $\boldsymbol{U}^*$ is column-orthogonal, we have $\boldsymbol{X}^T\boldsymbol{U}^* = \boldsymbol{B}\boldsymbol{\Sigma}^T$. Therefore, $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{\Sigma}^T = \boldsymbol{X}^T\boldsymbol{U}^*$. Substituting $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{U}^*$ into problem (4), we have

$$\min_{\boldsymbol{C}} \|\boldsymbol{X}^T\boldsymbol{U}^* - \boldsymbol{X}^T\boldsymbol{C}\|_F^2. \tag{5}$$

Since $\boldsymbol{U}^*$ is fixed and column-orthogonal, there must exist a column-orthogonal matrix $\boldsymbol{U}_\perp^*$, such that $[\boldsymbol{U}^*, \boldsymbol{U}_\perp^*]$ is an $m \times m$ orthogonal matrix. At this moment, $\boldsymbol{U}^{*T}\boldsymbol{U}_\perp^* = \boldsymbol{O}$, and so we have

$$\|\boldsymbol{X}^T\boldsymbol{U}^* - \boldsymbol{X}^T\boldsymbol{C}\|_F^2 = \|\boldsymbol{X} - \boldsymbol{U}^*\boldsymbol{C}^T\boldsymbol{X}\|_F^2. \tag{6}$$

Therefore, the following self-contained regression type is produced when $\boldsymbol{U}^*$ is not fixed:

$$\min_{\boldsymbol{C},\boldsymbol{U}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{C}^T\boldsymbol{X}\|_F^2, \quad \text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \tag{7}$$

where $\boldsymbol{C}$ is the regression coefficient matrix, and the subspace spanned by the columns of $\boldsymbol{C}^*$ is the same as that spanned subspace by the columns of $\boldsymbol{U}^*$ (see [34, Th. 3]). That is, $\boldsymbol{X}^T\boldsymbol{C}^*$ approximates to principal components $\boldsymbol{X}^T\boldsymbol{U}^*$. In problem (7), $\boldsymbol{U}$ is often called the auxiliary transformation matrix, and the regression coefficient matrix $\boldsymbol{C}$ is often called the transformation matrix. Generally, $\boldsymbol{C}$ in problem (7) has the better explanatory significance than $\boldsymbol{U}$ in problem (1) because any sparse norm can be added to $\boldsymbol{C}$.

### C. Optimal Mean Robust Principal Component Analysis

In the above-mentioned PCA, we assume that the data matrix $\boldsymbol{X}$ has been centralized, that is, the data mean is zero. In fact, the mean of data is usually not zero. By denoting the mean vector by a variable $\boldsymbol{b}$, OMRPCA is proposed to optimize the following $\ell_{2,1}$ minimization problem:

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{b}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\|_{2,1} \\ &\text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \end{aligned} \tag{8}$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ is a transformation matrix, $\boldsymbol{b} \in \mathbb{R}^m$ is a mean vector, and $\boldsymbol{1} \in \mathbb{R}^n$ is a column vector with all its elements being one.

Using some mathematical techniques [18], problem (8) can be converted to the following formulation:

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{b}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}} - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}}\|_F^2 \\ &\text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \end{aligned} \tag{9}$$

where $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix, whose $j$th diagonal element is $(1/2\|[(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)]_j\|_2)$. Note that $[(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)]_j$ means the $j$th column of $(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)$. In essence, $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ induced by $\|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\|_{2,1}$ can be viewed as the weight matrix of the data samples, and thus, we say that OMRPCA is an adaptive weighted PCA with an automatic scheme of optimal mean removal.

Note that the optimal transformation matrix $\boldsymbol{U}^*$ and optimal mean $\boldsymbol{b}^*$ are learned on the original data points that are not centralized. For any one original data point $\boldsymbol{x}$, the reconstructed data by OMRPCA are $\boldsymbol{U}^*\boldsymbol{U}^{*T}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{b}$.

## III. Adaptive Weighted Sparse Principal Component Analysis

In this section, the non-convex and convex formulas of the proposed method are first deduced, and then, the optimization algorithm and discussion are given.

### A. Non-Convex Formulation

Inspired by the self-contained regression type of PCA, we argue that $\min_{\boldsymbol{Q},\boldsymbol{U},\boldsymbol{b}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}} - \boldsymbol{U}\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}}\|_F^2$ is the self-contained regression type of OMRPCA (i.e., $\min_{\boldsymbol{U},\boldsymbol{b}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}} - \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\sqrt{\boldsymbol{D}}\|_F^2$) when $\boldsymbol{D}$ is fixed and $\boldsymbol{U}$ is a column-orthogonal matrix. Once a regression type is produced, an arbitrary sparse regularization term can be added. Therefore, we can add the sparse regularization term $\|\boldsymbol{Q}\|_{2,1}$ to penalize each column of $\boldsymbol{Q}$, i.e., all $d$ regression coefficients as a whole, and obtain $m$ penalty values. These obtained penalty values can govern the use or removal of features. Based on the self-contained regression type of OMRPCA, we propose the relaxed sparse self-contained regression type of OMRPCA as follows:

$$\begin{aligned} &\min_{\boldsymbol{Q},\boldsymbol{U},\boldsymbol{b}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{U}\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\|_{2,1} + \lambda\|\boldsymbol{Q}\|_{2,1} \\ &\text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \end{aligned} \tag{10}$$

where each column of matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ represents a vectorized image, $\boldsymbol{Q} \in \mathbb{R}^{d\times m}$ is the transformation matrix, $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ is the auxiliary transformation matrix, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{1} \in \mathbb{R}^n$ have been defined in problem (8), and the parameter $\lambda \geq 0$ plays an important role in balancing the loss term and the regularization term.

From problem (10), it can be seen that $\boldsymbol{Q} \in \mathbb{R}^{d\times m}$ is first used to transform the centralized data $(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)$ to the low-dimensional data $\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)$, and then, $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ is used to transform the low-dimensional data to the original data.

In fact, when $\boldsymbol{Q} = \boldsymbol{U}^T$, problem (10) becomes the sparse self-contained regression type of OMRPCA. However, in problem (10), $\boldsymbol{Q}$ does not necessarily equal $\boldsymbol{U}^T$, and therefore, we say that problem (10) is a relaxation of the sparse self-contained regression type of OMRPCA.

### B. Convex Formulation

The problem (10) is non-convex and cannot get the global optimal solution. Since $\boldsymbol{U}$ is column-orthogonal, by the definition of $\ell_{2,1}$-norm, we have $\|\boldsymbol{Q}\|_{2,1} = \|\boldsymbol{U}\boldsymbol{Q}\|_{2,1}$. Replacing $\boldsymbol{U}\boldsymbol{Q}$ in problem (10) with $\boldsymbol{A}$ yields

$$\min_{\boldsymbol{b},\boldsymbol{A}} \|(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T) - \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^T)\|_{2,1} + \lambda\|\boldsymbol{A}\|_{2,1}. \tag{11}$$

However, it is still not convex. Furthermore, we have the following equivalent formulation:

$$\min_{\boldsymbol{b},\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{X} - (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b}\boldsymbol{1}^T\|_{2,1} + \lambda\|\boldsymbol{A}\|_{2,1}. \tag{12}$$

For any one original data point $\boldsymbol{x}$, its reconstruction is $\boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b}$. Replacing $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b}$ with $\boldsymbol{v}$, the following convex surrogate is produced:

$$\min_{\boldsymbol{v},\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{X} - \boldsymbol{v}\boldsymbol{1}^T\|_{2,1} + \lambda\|\boldsymbol{A}\|_{2,1}. \tag{13}$$

Problem (13) is convex and can get its global optimal solution. For any original vectorized image $\boldsymbol{x}$, the vectorized image is reconstructed as $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}$.

According to the properties of $\ell_{2,1}$-norm [18], problem (13) can be rewritten as

$$\min_{\boldsymbol{v},\boldsymbol{A}} \|(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T)\sqrt{\boldsymbol{W}_1}\|_F^2+\lambda\|\boldsymbol{A}\sqrt{\boldsymbol{W}_2}\|_F^2 \quad (14)$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{n\times n}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{m\times m}$ are the two diagonal matrices, whose $j$th diagonal elements are expressed as $(1/2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2)$ and $(1/2\|\boldsymbol{a}_j\|_2)$, respectively. Note that $[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j$, $j \in \{1,2,\ldots,n\}$, means the $j$th column of matrix $\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T$, and $\boldsymbol{a}_j$, $j \in \{1,2,\ldots,m\}$, means the $j$th column of matrix $\boldsymbol{A}$. When $\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2 = 0$, we let $\boldsymbol{W}_1^{jj} = (1/2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2+\zeta)$, where $\zeta$ is a very small constant. Similarly, when $\|\boldsymbol{a}_j\|_2 = 0$, we let $\boldsymbol{W}_2^{jj} = (1/2\|\boldsymbol{a}_j\|_2+\zeta)$. In this way, the smaller $\boldsymbol{W}_1^{jj}$ is, the higher possibility to be outliers the $j$th sample has. Here, $\sqrt{\boldsymbol{W}_1}$ gives the weights of the data samples. The clean samples are weighted more heavily, while the samples that are outliers are weighted less heavily. This leads to the robustness of our method to outliers. Moreover, the regularization term $\boldsymbol{A}\sqrt{\boldsymbol{W}_2}$ can guide the selection of features. When a suitable parameter $\lambda$ is adjusted, our method can select the representative features.

It is worth noting that when $\lambda = 0$, the convex objective function in problem (13) has the trivial solution $\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{v} = \boldsymbol{0}$, which can be avoided by setting $\lambda \neq 0$. Therefore, in problem (13), the parameter $\lambda$ is set as $\lambda > 0$.

### C. Optimal Solution

The problem (14) is solved by using the iterative re-weighting method, which includes the following two steps.

*Step 1:* Given $\boldsymbol{A}$, the optimization problem (14) becomes the computation of $\boldsymbol{v}$

$$\min_{\boldsymbol{v}} \left\|(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T)\sqrt{\boldsymbol{W}_1}\right\|_F^2. \quad (15)$$

Taking the derivative of (15) with respect to $\boldsymbol{v}$ to be zero, we get $\boldsymbol{v} = ((\boldsymbol{X}\boldsymbol{W}_1-\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}_1)\boldsymbol{1}/\boldsymbol{1}^T\boldsymbol{W}_1\boldsymbol{1})$.

*Step 2:* Given $\boldsymbol{v}$, the optimization problem (14) becomes the computation of $\boldsymbol{A}$

$$\min_{\boldsymbol{A}} \|(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T)\sqrt{\boldsymbol{W}_1}\|_F^2+\lambda\|\boldsymbol{A}\sqrt{\boldsymbol{W}_2}\|_F^2. \quad (16)$$

Taking the derivative of (16) with respect to $\boldsymbol{A}$ to be zero, we get $\boldsymbol{A} = (\boldsymbol{X}\boldsymbol{W}_1-\boldsymbol{v}\boldsymbol{1}^T\boldsymbol{W}_1)\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^T+\lambda\boldsymbol{W}_2)^{-1}$.

Iterating the above-mentioned two steps will obtain the global optimal solution. See Algorithm 1 for more details.

### D. Discussion

Here, we discuss the relationship between problems (10) and (13). First, we propose problem (10), but its objective function is not convex. In order to construct a convex objective function, we produce problem (13) by replacing $\boldsymbol{U}\boldsymbol{Q}$ of problem (10) with $\boldsymbol{A}$ and $(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{b}$ of problem (12) with $\boldsymbol{v}$.

Before giving the relationship between problems (10) and (13), we first make the following definitions. Suppose $(\boldsymbol{Q}^0,\boldsymbol{U}^0,\boldsymbol{b}^0)$ and $(\boldsymbol{A}^*,\boldsymbol{v}^*)$ are the optimal solutions to problems (10) and (13), respectively, we have

**Algorithm 1** Optimization of Problem

**Input**: Data matrix $\boldsymbol{X}$ and parameter $\lambda$;

1: Initialize $\boldsymbol{W}_1 = \boldsymbol{I}$, $\boldsymbol{W}_2 = \boldsymbol{I}$ and $\boldsymbol{v} = \boldsymbol{0}$;

2: **while** not converge **do**

2.1: Compute $\boldsymbol{A} = (\boldsymbol{X}\boldsymbol{W}_1-\boldsymbol{v}\boldsymbol{1}^T\boldsymbol{W}_1)\boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^T+\lambda\boldsymbol{W}_2)^{-1}$;

2.2: Compute $\boldsymbol{v} = \frac{(\boldsymbol{X}\boldsymbol{W}_1-\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}_1)\boldsymbol{1}}{\boldsymbol{1}^T\boldsymbol{W}_1\boldsymbol{1}}$;

2.3: Compute $\boldsymbol{W}_1 = \begin{bmatrix} \frac{1}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_1\|_2} & & \\ & & \\ & & \ddots \end{bmatrix}$;

2.4: Compute $\boldsymbol{W}_2 = \begin{bmatrix} \frac{1}{2\|\boldsymbol{a}_1\|_2} & & \\ & \frac{1}{2\|\boldsymbol{a}_2\|_2} & \\ & & \ddots \end{bmatrix}$;

**end while**

**Output**: Optimal transformation matrix $\boldsymbol{A}^*$ and mean vector $\boldsymbol{v}^*$.

$f^0 = \|(\boldsymbol{X}-\boldsymbol{b}^0\boldsymbol{1}^T)-\boldsymbol{U}^0\boldsymbol{Q}^0(\boldsymbol{X}-\boldsymbol{b}^0\boldsymbol{1}^T)\|_{2,1}+\lambda\|\boldsymbol{Q}^0\|_{2,1}$ and $f^* = \|\boldsymbol{X}-\boldsymbol{A}^*\boldsymbol{X}-\boldsymbol{v}^*\boldsymbol{1}^T\|_{2,1}+\lambda\|\boldsymbol{A}^*\|_{2,1}$, respectively.

If $\boldsymbol{A} = \boldsymbol{U}^0\boldsymbol{Q}^0$ and $\boldsymbol{v} = (\boldsymbol{I}-\boldsymbol{U}^0\boldsymbol{Q}^0)\boldsymbol{b}^0$, then $(\boldsymbol{A},\boldsymbol{v})$ is a feasible solution to problem (13). Because of the convexity of the objective function in problem (13), the objective function value $f^0$ obtained by a feasible solution $(\boldsymbol{A},\boldsymbol{v})$ is obviously no less than the objective function value $f^*$ obtained by the optimal solution $(\boldsymbol{A}^*,\boldsymbol{v}^*)$, i.e., $f^* \leq f^0$. Therefore, problem (13) can always obtain the better optimal solution than problem (10) in theory.

In fact, we also compared the off-line results of problems (10) and (13) and found that $f^* \leq f^0$ was always correct on the used data sets. As a consequence, the convex formulation [see problem (13)] performs slightly better than the non-convex formulation [see problem (10)], and in most cases, the gap is about 1%. Considering the importance of convex formulation, we only explore problem (13) in the following.

## IV. Theoretical Analyses of Problem (13)

In this section, the theoretical analyses of Algorithm 1, including a convergence analysis and a computational complexity analysis, are introduced.

### A. Convergence Analysis

Before proving the convergence Algorithm 1, Lemma 1 [35] is first introduced as follows.

*Lemma 1*: For any nonzero vectors $\boldsymbol{U},\boldsymbol{q} \in \mathbb{R}^c$

$$\|\boldsymbol{U}\|_2-\frac{\|\boldsymbol{U}\|_2^2}{2\|\boldsymbol{q}\|_2} \leq \|\boldsymbol{q}\|_2-\frac{\|\boldsymbol{q}\|_2^2}{2\|\boldsymbol{q}\|_2}. \quad (17)$$

Based on Lemma 1, we propose the following Theorem 1.

*Theorem1*: Algorithm 1 will monotonically decrease the value of the objective function of the optimization problem (13) in each iteration and converges to the global optimal solution.

*Proof:* In each iteration, the updated $\boldsymbol{v}$ and $\boldsymbol{A}$ values are denoted by $\widetilde{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{A}}$. Since the updated values $\widetilde{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{A}}$ are the

optimal solution of problem (13), according to the definition of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, we have

$$\operatorname{tr}\left(\sum_{j=1}^{n}\frac{\|[\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}\right)+\lambda\operatorname{tr}\left(\sum_{j=1}^{m}\frac{\|\widetilde{\boldsymbol{a}}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}\right)$$
$$\leq \operatorname{tr}\left(\sum_{j=1}^{n}\frac{\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}\right)+\lambda\operatorname{tr}\left(\sum_{j=1}^{m}\frac{\|\boldsymbol{a}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}\right). \tag{18}$$

On the one hand, according to Lemma 1, we have

$$\|[\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T]_j\|_2-\frac{\|[\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}$$
$$\leq \|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2-\frac{\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}. \tag{19}$$

Using matrix calculus for (19), we have the following formulation:

$$\sum_{j=1}^{n}\|[\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T]_j\|_2-\sum_{j=1}^{n}\frac{\|[\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}$$
$$\leq \sum_{j=1}^{n}\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2-\sum_{j=1}^{n}\frac{\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2^2}{2\|[\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T]_j\|_2}. \tag{20}$$

On the other hand, according to Lemma 1, we have

$$\|\widetilde{\boldsymbol{a}}_j\|_2-\frac{\|\widetilde{\boldsymbol{a}}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}\leq\|\boldsymbol{a}_j\|_2-\frac{\|\boldsymbol{a}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}. \tag{21}$$

Using matrix calculus for (21), we have the following formulation:

$$\lambda\left(\sum_{j=1}^{m}\left(\|\widetilde{\boldsymbol{a}}_j\|_2-\frac{\|\widetilde{\boldsymbol{a}}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}\right)\right)\leq\lambda\left(\sum_{j=1}^{m}\left(\|\boldsymbol{a}_j\|_2-\frac{\|\boldsymbol{a}_j\|_2^2}{2\|\boldsymbol{a}_j\|_2}\right)\right). \tag{22}$$

By summing for (18), (20), and (22), we have

$$\|\boldsymbol{X}-\widetilde{\boldsymbol{A}}\boldsymbol{X}-\widetilde{\boldsymbol{v}}\boldsymbol{1}^T\|_{2,1}+\lambda\|\widetilde{\boldsymbol{A}}\|_{2,1}$$
$$\leq\|\boldsymbol{X}-\boldsymbol{A}\boldsymbol{X}-\boldsymbol{v}\boldsymbol{1}^T\|_{2,1}+\lambda\|\boldsymbol{A}\|_{2,1}. \tag{23}$$

Since the objective function of problem (13) has an obvious lower bound 0, Algorithm 1 converges to the global optimal solution.

### B. Computational Complexity Analysis

In each iteration, the computational complexity of Algorithm 1 mainly focuses on two steps: the first one is the computational complexity of $\boldsymbol{v}$ with $O(mn^2)$, and the other one is the computational complexity of $\boldsymbol{A}$ with $O(m^3)$ at most. Therefore, the computational complexity of one iteration will be up to $O(m^3)$. If Algorithm 1 needs $t$ iterations, the total computational complexity is $O(tm^3)$.

TABLE I
DATA SET DESCRIPTION

| Datasets | Sample | Feature | Class |
|---|---|---|---|
| ORL | 400 | 1024 | 40 |
| COIL20 | 1440 | 1024 | 20 |
| COIL100 | 7200 | 1024 | 100 |
| USPS | 9298 | 256 | 10 |
| UMIST | 2576 | 575 | 20 |
| Isolet1 | 1560 | 617 | 26 |
| LUNG | 203 | 3312 | 5 |
| Binary Alphadigs | 1404 | 320 | 36 |
| Leukemia1 | 72 | 5327 | 3 |
| Tumors9 | 60 | 5726 | 9 |

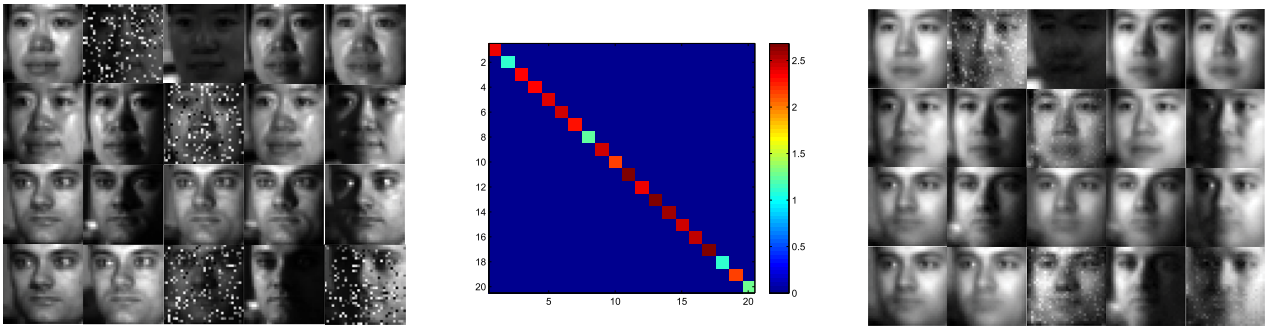
## V. EXPERIMENTS

The proposed method can select the effective feature for robust reconstruction and clustering, and thus, the experiments are implemented in the following two groups.

1) *Experiments of Robust Reconstruction:* In this group, the following three data sets are used for reconstruction.
   a) *Corrupted PIE Data Set:* The PIE face data set [36] contains 68 classes, each class contains 24 face images, and each image is with $32\times 32$ pixels. Based on the PIE data set, the corrupted PIE data set is made in the following way. First, ten samples per class from the PIE data set are randomly selected. Of these 680 images selected from the PIE face data set, 20% of them are corrupted by pepper and salt noise, and the corruption proportion is 20% of an image.
   b) *Yale10 Data Set:* This data set [37] contains 10 classes, each class contains 64 images, and each image is with $42\times 48$ pixels. More than half of the data images are corrupted by "shadows" and noise, and so the corruptions in the Yale10 data set are heavy.
   c) *Background–Foreground Separation Data Set:* This data set [38] is a collection of 502 images captured by a static camera over one day, where the size of each image is $120\times 160$ pixels. Therefore, this data set has a static background with changes in the illumination. Moreover, 40% of this data set contain people in various locations, and the people can be regarded as noise.
2) *Experiments of Robust Clustering:* In this group, ten data sets, i.e., ORL, COIL20, COIL100, USPS, UMIST, Isolet1, LUNG, Binary Alphadigits, Leukemia1, and Tumors9, are used, and the details are given in Table I.

### A. Experiments of Robust Reconstruction

In order to demonstrate that AW-SPCA can perform robust reconstruction, we take the first 20 images of the corrupted PIE data set as an example to elaborate the ability of robust reconstruction of AW-SPCA (see Fig. 1). More specifically, Fig. 1(a) shows the used 20 images, in which the 2nd, 8th, 18th, and 20th images are corrupted by pepper and salt noise, Fig. 1(b) shows the corresponding weights of these 20 images, and Fig. 1(c) shows the reconstruction results of these 20 images. Fig. 1 demonstrates the robustness of the proposed method



(a) Origin images. (b) Adaptive weight matrix $\sqrt{W_1}$. (c) Reconstruction images.

Fig. 1. Illustration of robust reconstruction of AW-SPCA. (a) First 20 images from the corrupted PIE data set. (b) Diagonal elements of $W_1$ which means the weights of the 20 samples. (c) Reconstruction results of original images.

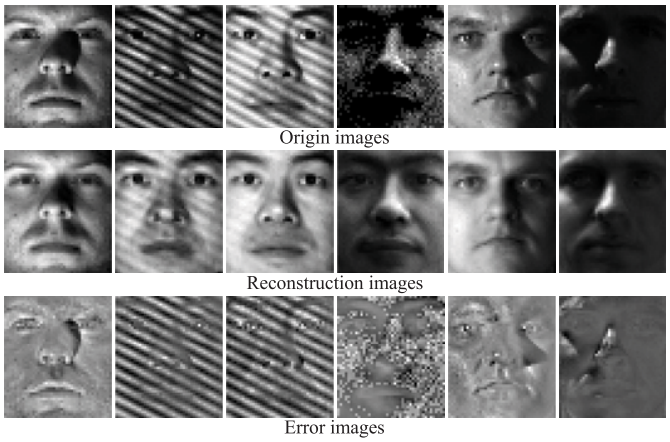



Fig. 2. Some images reconstructed by AW-SPCA with $\lambda = 0.5$ on the Yale10 data set.

to outliers, whose robust principal is the adaptive weighting ability. That is, if one data sample is corrupted, its weight (reflected by a corresponding diagonal element of $\sqrt{W_1}$) is, adaptively, assigned for a small value; otherwise, its weight is, adaptively, assigned for a large value. This can soften the impact of outliers on reconstruction. Moreover, the proposed method is also performed on the Yale10 data set, and the experimental results (see Fig. 2) demonstrate the robustness of the proposed method to the varying illumination, shadows, and noise.

In addition, the proposed method is performed on the background–foreground separation data set [38]. The background–foreground separation data set can be seen as a data set with noise. That is, the background is fixed, while the foregrounds are dynamic that can be regarded as noise. Some experimental results of the proposed method are shown in Fig. 3, where the top row shows some original images (i.e., the background is corrupted by foregrounds), the middle row shows the recovered background images, and the bottom row shows the noise images. Fig. 3 shows that the proposed method can well separate the noise (i.e., people) from the corrupted background.

### B. Experiments of Robust Clustering

In order to demonstrate that the features selected by AW-SPCA can obtain robust clustering performance, this paper adopts two common clustering evaluation metrics,

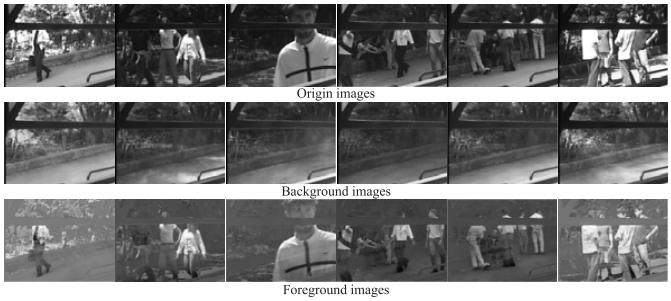



Fig. 3. Recovery results of AW-SPCA on the background–foreground separation data set.

namely accuracy (ACC) and normalized mutual information (NMI). The proposed AW-SPCA method is compared with a baseline method and four spectral feature selection methods, i.e., UDFS [26], RUFS [14], RSFS [22], and SOGFS [33]. Unlike the spectral feature selection methods, the baseline method uses all the features of a data set to perform clustering. In this paper, all the methods search for the best optimal parameters within the search range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. After the effective features are selected by different methods, $k$-means is adopted to perform clustering according to these selected features. To reduce the dependence of $k$-means on initialization values, we count the clustering results of 20 times, and the average clustering result along with the standard deviation is reported.

*1) Clustering Experiments on Ten Data Sets Without Noise:* For the first eight data sets in Table I, the number of selected features (that is, #features) is initially set to 50 with an incremental interval of 50. Fig. 4 intuitively shows the variation of the clustering results (ACC) of different methods with the variation of #features. As shown in Fig. 4, with the limited features, the proposed method and RSFS are superior to baseline, UDFS, RUFS, and SOGFS in most cases. Furthermore, Tables II and III report the clustering results of different methods on ten data sets without noise, where the average rank of each method is in the last row. Note that the average rank is the average of ranking scores, where ranking scores are obtained by ranking the ACC of a method on all data sets. From Tables II and III, it can be seen that the performance of most feature selection methods is better than that of the baseline method, and the performance of the proposed method

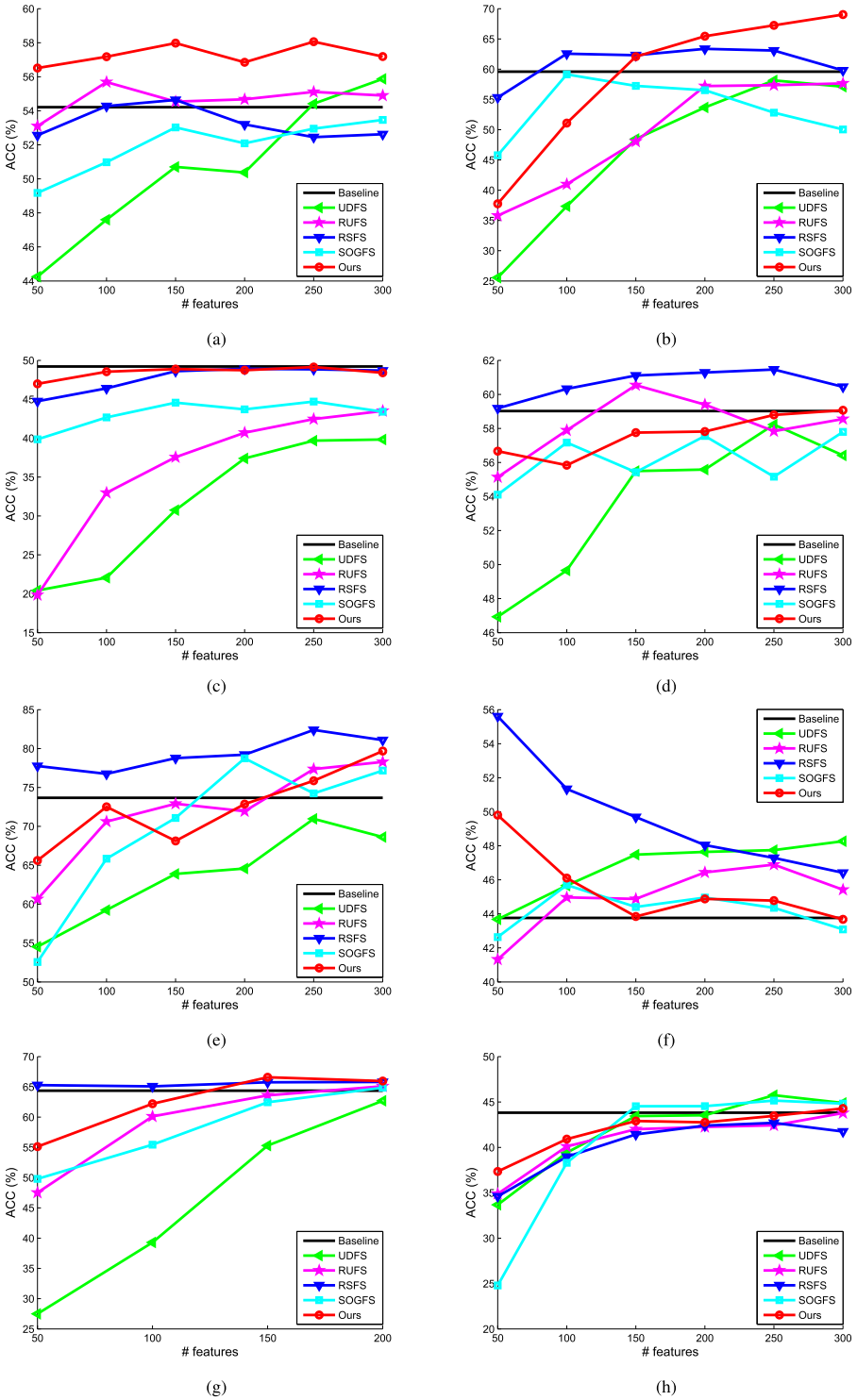



Fig. 4. Illustration of the variation of clustering results (ACC) with the number variation of selected features on eight data sets. (a) ORL. (b) Isolate 1. (c) COIL 100. (d) COIL 20. (e) LUNGML. (f) UMIST. (g) USPS. (h) Binary Alphadigs.

is close to that of the spectral feature selection methods. Although the proposed method obtains the best clustering result from the statistical significant perspective, it does not always show a superior result on the data sets without noise. To this end, we further implement all the methods on the data sets with noise.

*2) Robust Clustering Experiments on ORL Data Set With Noise:* Here, the robust clustering ability of the proposed method is verified on the corrupted ORL data set with different corruption ratios. More specifically, we randomly choose 20% images from the total 400 images and make the selected images to be randomly corrupted by pepper and salt noise, with corruptions of 10% and 20%, respectively. For each method, 20 tests are conducted on the corrupted ORL data set, and the average clustering result is shown in Fig. 5.

Since the feature selection method is proposed from the perspective of robust reconstruction, it can not only select effective features from the original data but also select effective features from the reconstructed data. Here, we label Ours1 and Ours2 as the way to select the effective features

TABLE II

CLUSTERING RESULTS (ACC%±STD) OF SIX FEATURE SELECTION METHODS ON TEN DATA SETS, WHERE THE BOLD FONTS MARK THE BEST RESULTS, THE UNDERLINED FONTS MARK THE SECOND BEST RESULTS, AND THE PARENTHESES CORRESPOND TO THE NUMBER OF SELECTED FEATURES

| | Baseline | UDFS[26] | RUFS[14] | RSFS[22] | SOGFS[33] | Ours |
|---|---|---|---|---|---|---|
| ORL | 54.21±3.48 | 55.88±3.45(300F) | 55.11±2.15(250F) | 54.64±3.27(150F) | 53.99±2.03(200F) | **58.82±3.12**(150F) |
| COIL20 | 59.02±4.90 | 58.23±4.34(250F) | 60.55±3.90(150F) | **61.46±4.56**(250F) | 57.80±5.39(300F) | 59.07±4.02(300F) |
| COIL100 | **49.27±1.97** | 39.82±1.32(300F) | 43.49±1.99(100F) | 48.69±2.23(300F) | 44.70±1.53(250F) | 49.14±2.19(250F) |
| USPS | 64.37±3.60 | 62.71±3.48(200F) | 65.10±1.63(200F) | 65.09±4.99(100F) | 64.87±2.46(200F) | **66.59±3.35**(150F) |
| UMIST | 43.76±2.40 | 48.27±2.94(300F) | 46.90±2.83(250F) | **55.61±3.60**(50F) | 45.70±3.09(100F) | 49.80±2.97(50F) |
| Isolet | 59.59±3.77 | 57.12±4.44(300F) | 57.65±3.85(300F) | 63.36±3.38(200F) | 59.13±5.62(100F) | **69.04±3.89**(300F) |
| LUNG | 73.67±9.02 | 70.96±7.05(250F) | 78.28±3.92(300F) | **82.39±5.14**(250F) | 78.74±11.52(200F) | 79.93±5.90(300F) |
| Binary Alphadigs | 43.82±2.30 | **45.75±2.17**(250F) | 43.78±1.93(300F) | 42.70±1.75(250F) | 45.16±2.09(250F) | 44.29±1.63(300F) |
| Leukemia1 | **55.21±9.09** | 49.72±4.12(300F) | 51.74±4.25(300F) | 51.25±6.64(300F) | 49.44±4.47(300F) | 54.51±7.86(300F) |
| Tumors9 | 38.50±3.28 | 37.50±3.03(800F) | 37.42±2.83(300F) | 37.50±3.31(500F) | 34.08±3.35(1000F) | **41.58±4.21**(500F) |
| average rank | 3.600 | 4.300 | 3.800 | 2.900 | 4.600 | **1.800** |

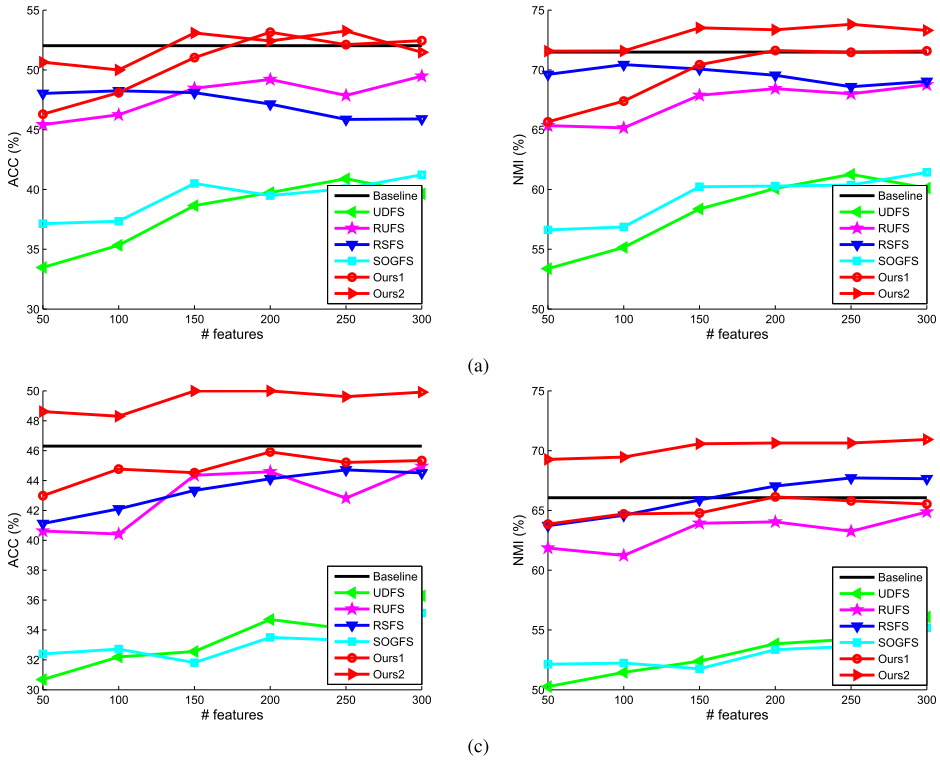



Fig. 5. Clustering results (ACC and NMI) on the ORL data set with different corruption ratios. (a) ACC on the corrupted data set with a corruption ratio of 10%. (b) NMI on the corrupted data set with a corruption ratio of 10%. (c) ACC on the corrupted data set with a corruption ratio of 20%. (d) NMI on the corrupted data set with a corruption ratio of 20%.

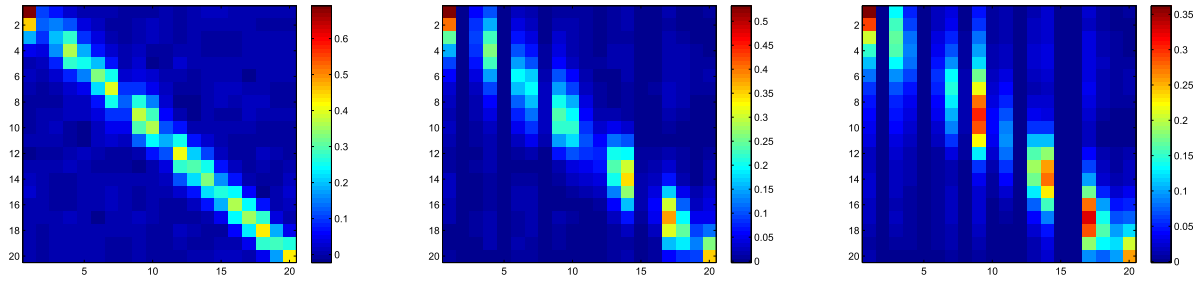



Fig. 6. Feature selection of the proposed method with different $\lambda$ values on the PIE data set. (a) A on $\lambda = 0:1$. (b) A on $\lambda = 0:3$. (c) A on $\lambda = 0:5$.

from the original data and the reconstructed data, respectively. Fig. 5 shows the clustering results (ACC and NMI) of different methods on the corrupted ORL data sets, where Fig. 5(a) and (b) shows the clustering results (ACC and NMI) on the ORL data set with a corruption ratio of 10%, while Fig. 5(c) and (d) shows the clustering results (ACC and NMI) on the ORL data set with a corruption ratio of 20%. From Fig. 5, we observe the following three phenomena. First,

TABLE III

CLUSTERING RESULTS (NMI%±STD) OF SIX FEATURE SELECTION METHODS ON TEN DATA SETS, WHERE THE BOLD FONTS MARK THE BEST RESULTS, THE UNDERLINED FONTS MARK THE SECOND BEST RESULTS, AND THE PARENTHESES CORRESPOND TO THE NUMBER OF SELECTED FEATURES

| | Baseline | UDFS[26] | RUFS[14] | RSFS[22] | SOGFS[33] | Ours |
|---|---|---|---|---|---|---|
| ORL | 74.93±1.99 | 75.10±1.60(300F) | 75.40±1.07(250F) | 75.77±1.49(150F) | 74.73±1.17(200F) | **76.76±1.44**(150F) |
| COIL20 | 74.53±2.18 | 71.95±1.92(250F) | 74.98±1.44(150F) | **76.59±1.85**(250F) | 72.15±2.44(300F) | 75.29±1.49(300F) |
| COIL100 | 75.78±0.67 | 68.11±0.64(300F) | 71.86±0.74(100F) | 76.57±0.58(300F) | 71.21±0.70(250F) | **76.77±0.50**(250F) |
| USPS | 60.50±1.42 | 58.97±1.68(200F) | 60.68±0.70(200F) | **63.96±2.41**(100F) | 60.24±0.77(200F) | 62.37±2.02(150F) |
| UMIST | 64.95±1.59 | 63.60±1.65(300F) | 66.96±0.66(250F) | **73.20±1.70**(50F) | 65.41±1.43(100F) | 69.94±1.79(50F) |
| Isolet | 75.25±1.74 | 73.77±2.01(300F) | 73.40±1.68(300F) | 78.70±1.15(200F) | 72.55±1.57(100F) | **80.78±1.40**(300F) |
| LUNG | 56.48±3.96 | 50.70±4.36(250F) | 55.23±2.98(300F) | **66.69±3.71**(250F) | 59.32±7.67(200F) | 56.79±3.47(300F) |
| Binary Alphadigs | 59.04±1.05 | **60.39±1.17**(250F) | 59.13±0.93(300F) | 59.20±1.06(250F) | 60.37±0.86(250F) | 59.08±0.85(300F) |
| Leukemia1 | **21.17±13.78** | 11.21±6.04(300F) | 16.90±11.83(300F) | 15.78±10.88(300F) | 9.51±5.23(300F) | 18.37±14.23(300F) |
| Tumors9 | 39.74±3.44 | 34.92±3.59(800F) | 36.00±3.62(300F) | 35.99±4.37(500F) | 31.04±2.57(1000F) | **41.41±3.86**(500F) |
| average rank | 3.700 | 4.900 | 3.600 | 2.100 | 4.700 | **2.000** |

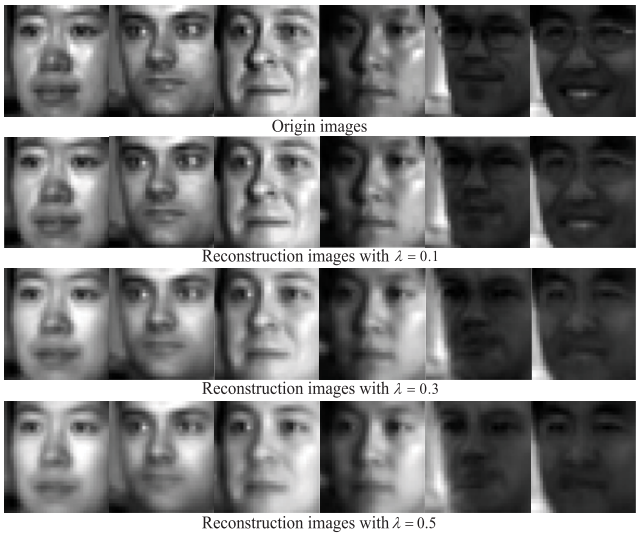



Fig. 7. Some reconstruction images by AW-SPCA with different $\lambda$ values.

both two robust spectral feature selection methods, i.e., RUFS and RSFS, outperform UDFS and SOGFS. This is because the design of UDFS does not consider the effect of noise in the learning of pseudo-cluster labels. Although SOGFS considers the effect of noise, it does not obtain robust performance. This may be because the adaptive similarity matrix learned by SOGFS is still affected when the data corruption degree is large. Second, our method, including Ours1 and Ours2, especially Ours2, obtains the best clustering result than the spectral feature selection methods. This is because the spectral feature selection methods need to construct a graph Laplacian that is easily affected by outliers and thus interferes with the result of feature selection, while our method selects those important features guided by the robust reconstruction and thus it can soften the impact of noise. Third, with an increase of corruption ratio, i.e., the corruption ratio increases from 10% to 20%, Ours2 performs far better than Ours1. This is because Ours1 will inevitably select many noisy features from the original data as the corruption ratio increases. At this point, it will be seriously affected by outliers, and the clustering performance will become worse. This indicates that when the corruption ratio of data becomes large, it is very important to design a feature selection method that can select effective features from the reconstructed data.

In fact, a few of the existing spectral feature selection methods have the direct reconstruction term to select the effective features from its reconstructed data. Besides, the spectral feature selection methods often include many model parameters, while our method only includes a model parameter. Therefore, our method is simple but effective.

### C. Parameter Settings

For a fair experimental comparison, the grid search method is used in each group of experimental methods, and the optimal parameter values are obtained from the same parameter search range.

In the first group of reconstruction experiment, there is a model parameter, i.e., $\lambda$. The best optimal parameter is searched from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, and then, this search range is narrowed. The best optimal parameter with the minimal reconstruction error can be got. For the proposed method, different values of $\lambda$ correspond to different degrees of feature selection. As shown in Fig. 6, the proposed method selects fewer features for reconstruction with the increasing value of $\lambda$, while when the $\lambda$ value increases, the quality of the reconstructed images may decrease because of the loss of lots of information (see Fig. 7). Therefore, a balance between feature selection and reconstruction can be achieved by adjusting $\lambda$.

In the second group of clustering experiment, there is a model parameter (i.e., $\lambda$) and a feature parameter (#features). In order to evaluate the influence of these two parameters on the experiment, different combinations of the model parameter set and the feature number set are carried out for each data set. The best optimal parameters (i.e., $\lambda$ and #features) with the best clustering results can be got. Fig. 8 shows the average clustering results of the proposed method with different $\lambda$ and #features values. As can be seen, the proposed method is less sensitive to the choice of $\lambda$ within wide ranges and more sensitive to #features.

### D. Convergence Curves

The proposed method can obtain the global optimal solution, whose theoretical analyses have been given in Section IV-A. Here, the convergence curves on three data sets are taken as examples to verify the convergence of the proposed method, which can be seen from Fig. 9.

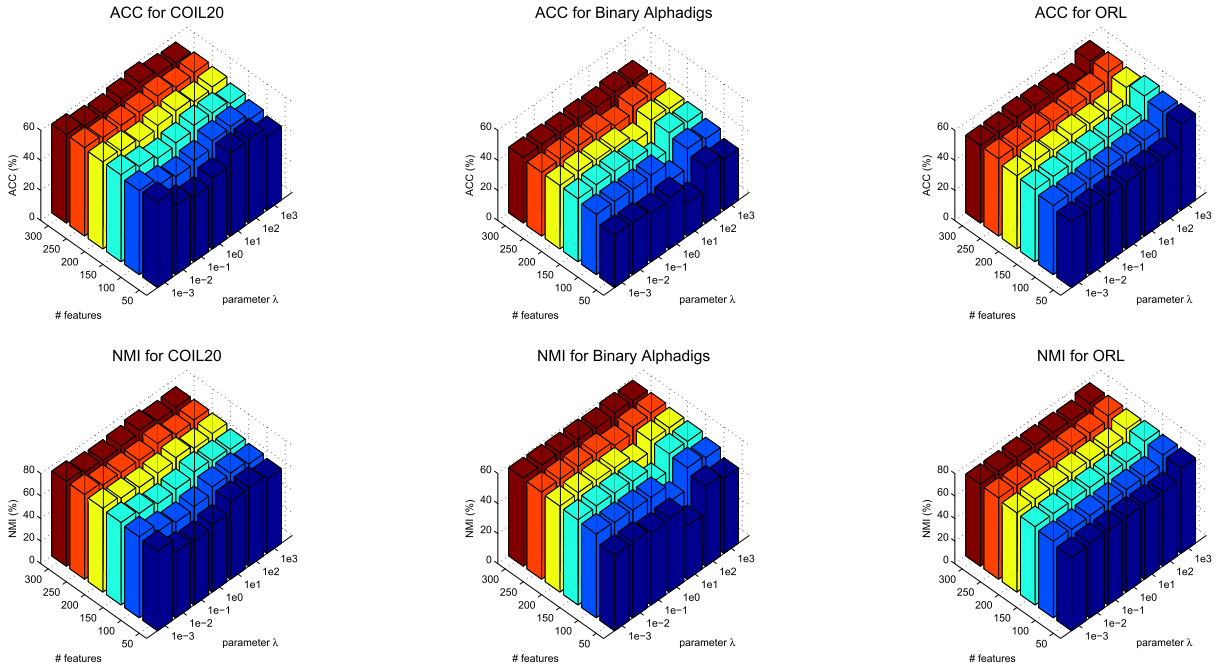
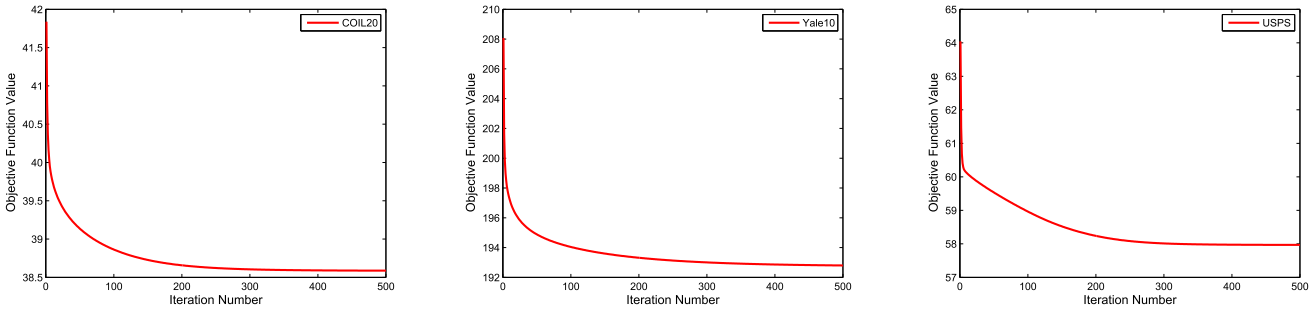



Fig. 8. Clustering results (ACC and NMI) of AW-SPCA with different $\lambda$ and #features values on different data sets.




(a) (b) (c)

Fig. 9. Convergence curves of AW-SPCA on three data sets. (a) COIL20. (b) Yale10. (c) USPS.

## VI. Conclusion

The proposed method is in a convex formulation and can obtain the global optimal solution. It is simple but effective. More specifically, only by using the robust PCA criteria to select effective features, this proposed method can obtain clustering results similar to that of the spectral feature selection methods on the noiseless data sets and is far better than that of the spectral feature selection methods on the noisy data sets. Therefore, the proposed method is a significant feature selection method, especially when the data set is heavily corrupted by noise.

## References

[1] Q. Peng *et al.*, "A hybrid of local and global saliencies for detecting image salient region and appearance," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 1, pp. 86–97, Jan. 2017.

[2] W. Ou *et al.*, "Robust face recognition via occlusion dictionary learning," *Pattern Recognit.*, vol. 47, no. 4, pp. 1559–1572, Apr. 2014.

[3] Q. Ge *et al.*, "Structure-based low-rank model with graph nuclear norm regularization for noise removal," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3098–3112, Jul. 2017.

[4] X.-Y. Jing *et al.*, "Multi-spectral low-rank structured dictionary learning for face recognition," *Pattern Recognit.*, vol. 59, pp. 14–25, Nov. 2016.

[5] X. You *et al.*, "Robust nonnegative patch alignment for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2760–2774, Nov. 2015.

[6] W.-S. Chen *et al.*, "Supervised kernel nonnegative matrix factorization for face recognition," *Neurocomputing*, vol. 205, pp. 165–181, Sep. 2016.

[7] J. Qian *et al.*, "Accurate tilt sensing with linear model," *IEEE Sensors J.*, vol. 11, no. 10, pp. 2301–2309, Oct. 2011.

[8] W.-S. Chen *et al.*, "Two-step single parameter regularization Fisher discriminant method for face recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 20, no. 2, pp. 189–207, 2006.

[9] W. Ou *et al.*, "Multi-view non-negative matrix factorization by patch alignment framework with view consistency," *Neurocomputing*, vol. 204, pp. 116–124, Sep. 2016.

[10] J. Wen, Y. Xu, and H. Liu, "Incomplete multiview spectral clustering with adaptive graph learning," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2018.28847.

[11] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[12] S. Zhang *et al.*, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[13] X. Sun *et al.*, "Non-rigid object contour tracking via a novel supervised level set model," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3386–3399, Nov. 2015.

[14] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1621–1627.

[15] Z. Lai *et al.*, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 723–735, Apr. 2016.

[16] Z. Lai *et al.*, "Multilinear sparse principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1942–1950, Oct. 2014.
[17] F. Nie *et al.*, "Trace ratio criterion for feature selection," in *Proc. 23rd Amer. Assoc. Artif. Intell.*, 2008, pp. 671–676.
[18] F. Nie *et al.*, "Efficient and robust feature selection via joint $\ell_2$, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[19] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proc. 24th Amer. Assoc. Artif. Intell.*, 2010, pp. 673–678.
[20] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[21] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.
[22] L. Shi, L. Du, and Y.-D. Shen, "Robust spectral learning for unsupervised feature selection," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 977–982.
[23] J. R. King and D. A. Jackson, "Variable selection in large environmental data sets using principal components analysis," *Environmetrics*, vol. 10, no. 1, pp. 67–77, 1999.
[24] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 507–514.
[25] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[26] Y. Yang *et al.*, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.
[27] Z. Li *et al.*, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. Amer. Assoc. Artif. Intell.*, 2012, pp. 1026–1032.
[28] Y. Shi *et al.*, "Feature selection with $\ell_{2,1--2}$ regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4967–4982, Oct. 2018.
[29] C. Ding *et al.*, "$R_1$-PCA: Rotational invariant $\ell_1$-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.
[30] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1062–1070.
[31] X. Shi *et al.*, "Robust principal component analysis via optimal mean by joint $\ell_{2,1}$ and Schatten p-norms minimization," *Neurocomputing*, vol. 283, pp. 205–213, Mar. 2018.
[32] M. Luo *et al.*, "Avoiding optimal mean $\ell_{2,1}$ norm maximization-based robust PCA for reconstruction," *Natural Comput.*, vol. 29, pp. 1124–1150, Apr. 2017.
[33] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. 13th Amer. Assoc. Artif. Intell.*, 2016, pp. 1302–1308.
[34] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, Jun. 2006.
[35] Z. Zheng, "Sparse locality preserving embedding," in *Proc. 2nd Int. Congr. Image Signal Process.*, Oct. 2009, pp. 1–5.
[36] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 46–51.
[37] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 663–670.
[38] F. De la Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, Aug. 2003.

**Shuangyan Yi** received the M.S. and Ph.D. degrees in mathematics and computer science and technology from the Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China. She is currently pursuing the Ph.D. degree with the Institute of Information Technology, Shenzhen Institute of Information Technology, Shenzhen, China, and also with the Shenzhen Key Laboratory of Information Science and Technology, Shenzhen Engineering Laboratory of IS&DRM, Department of Electronics Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen.

Her current research interests include pattern recognition, machine learning, and deep neural network compression.

**Zhenyu He** received the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007.

From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong. He is currently a Full Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His current research interests include machine learning, computer vision, image processing, and pattern recognition.

**Xiao-Yuan Jing** received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, Nanjing, China, in 1998.

He was a Professor with the Department of Computer, Shenzhen Research Student School, Harbin Institute of Technology, Harbin, China, in 2005. He is currently a Professor with the School of Computer Science, Wuhan University, Wuhan, China. He has published over 100 papers in TIP, TIFS, TSE, TCB, TCSVT, TMM, TR, TSMC-B, CVPR, AAAI, IJCAI, ICSE, and PR. His current research interests include pattern recognition, machine learning, image processing, artificial intelligence, and software engineering.

**Yi Li** received the B.E. and M.E. degrees in electronic and information engineering and computer science and technology from the Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China.

His current research interests include machine learning, computer vision, and image processing.

**Yiu-Ming Cheung** received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include artificial intelligence, visual computing, pattern recognition, and optimization.

Dr. Cheung is a fellow of the IET, BCS, and IETI. He is the Founding Chairman of the Computational Intelligence Chapter, IEEE Hong Kong Section. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, *Knowledge and Information Systems*, and the *International Journal of Pattern Recognition and Artificial Intelligence*, among others.

**Feiping Nie** received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. He has published over 100 papers in top journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS/*Techniques in Neurosurgery & Neurology*, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and *Bioinformatics*, and in top conferences, including ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 6500 times. His current research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Dr. Nie is currently a PC member of several prestigious journals and conferences in the related fields. He is also serving as an associate editor for several prestigious journals and conferences in the related fields.