Pattern Recognition 93 (2019) 164–178

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Robust Jointly Sparse Regression with Generalized Orthogonal Learning for Image Feature Selection

Dongmei Mo[a,b], Zhihui Lai[a,c,*]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[b] Institute of Textiles and Clothing, the Hong Kong Polytechnic University, Hong Kong
[c] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China

ARTICLE INFO

Article history:
Received 2 March 2018
Revised 24 June 2018
Accepted 9 April 2019
Available online 18 April 2019

Keywords:
Dimensionality reduction
Local structure
Joint sparsity
Orthogonality
Orthogonal matching pursuit

ABSTRACT

Ridge regression (RR) and its variants are fundamental methods for multivariable data analysis, which have been widely used to deal with different problems in pattern recognition or classification. However, these methods have their common drawback. That is, the number of the learned projections is limited by the number of class. Moreover, most of these methods do not consider the local structure of the data, which makes them less competitive in the case when data are lying on a lower dimensional manifold. Therefore, in this paper, we propose a robust jointly sparse regression method to integrate the locality geometric structure with generalized orthogonality constraint and joint sparsity into a regression modal to address these problems. The optimization model can be solved by an alternatively iterative algorithm using orthogonal matching pursuit (OMP) and singular value decomposition. Experimental results on face and non-face image database demonstrate the superiority of the proposed method. The matlab code can be found at http://www.scholat.com/laizhihui.

1. Introduction

Since the data in image/video processing, bioinformatics and web data mining are often high dimensional, the computational or memory cost can be very high. Therefore, it is very necessary to have powerful tools to deal with those massive data sets. Feature selection or extraction is considered one of the most effective tools to select or compress the important information into a reduced low-dimensional space [1–3]. Thus, many algorithms have been developed to deal with this problem [4,5]. The most widely used multivariable analysis methods for dimensionality reduction are principal component analysis (PCA) [6], linear discriminant analysis (LDA) [7], ridge regression (RR) and their variations.

However, in many practical applications like face recognition, the data is usually sampled from a nonlinear low-dimensional manifold of the high dimensional ambient space and both of PCA, LDA and RR are not suitable in these cases. Thus, many subspace learning algorithms based on manifold learning are proposed [8–12]. Motivated by the manifold learning methods, RR was extended to have local preserving ability [13–15].

Although all the subspace learning methods mentioned above have their suitable application cases, they still have a major disadvantage. That is, since their learned projections are linear combinations of all the original features, it lacks the interpretation of the results. For example, RR uses L_2-norm on the regularization term and lacks sparsity property. However, many regression methods using L_1-norm on the regularization term can obtain sparse projections, and thus they have attracted much attention in the field of machine learning and pattern recognition [16]. The most representative sparse regression methods are the sparse RR [17], and Elastic Net [18]. Motivated by the sparse RR and the Elastic Net, many subspace learning methods were extended to sparse cases in regression forms [19], including sparse PCA (SPCA) [6], sparse LDA (SLDA) [20], sparse locality preserving embedding (SLPE) [21] and sparse locality preserving projections (SpLPP) [22]. All these methods learn sparse projections by incorporating L_1 norm regularized regression in the process of learning projections. One of the main disadvantages is that they are usually time-consuming because the L_1-norm based methods conduct feature selection on high dimensional image vectors. In addition, the learn projections are not joint sparse. That is, the L_1-norm based sparse learning cannot obtain the joint sparsity which is considered much more effective for feature selection and classification in computer vision or biometric.

Motivated by the property of $L_{2,1}$-norm as regularization for jointly sparse learning, the regression methods are further developed to be the jointly sparse regression [23–25]. Nie et al. proposed efficient and robust feature selection (RFS) [26] via $L_{2,1}$-norms minimization regression. This method uses $L_{2,1}$-norm on both of loss function and the regularization term on the regression

* Corresponding author.
E-mail address: lai_zhi_hui@163.com (Z. Lai).

https://doi.org/10.1016/j.patcog.2019.04.011
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

model so as to not only enhance the robustness to outliers but also guarantee the joint sparse projections for effective feature selection. Other related $L_{2,1}$-norm based regression methods [27–31] were also proposed for jointly sparse subspace learning.

No matter what kinds of variation of the above methods, the basic model is still the form of ridge regression just using different norms as the measurement on the main part or on the regularization term. Therefore, in this paper, we focus on the basic model to develop a generalized ridge regression method to solve the potential drawbacks of the previous methods. First, the number of the learned projections is limited by the number of class (i.e. small-class problem), which means that they cannot obtain enough projections for more effective feature selection. Second, the correlation of the learned sparse projections direction is not taken into consideration. That is, since the projection directions are not mutually orthogonal, the effectiveness of each projection direction is not guaranteed. Third, the robustness as well as flexibility of the previous $L_{2,1}$-norm based methods are unknown since there is no specific technic incorporated into their objective functions to release this problem. Therefore, to release the above problems, we have done some research and the previous work was published as a conference paper [32]. However, the previous work still did not consider the orthogonality of the projection direction. Also, the local geometric structure of the data is ignored. In this paper, we further extend the proposed method in the conference paper [32] into a more general form. That is, one more constraint characterizing the manifold structure of the data is appended to the model in [32]. We call the proposed method Robust Jointly Sparse Regression (RJSR), which aims to solve the problems mentioned above so as to not only improve the performance of feature selection and extraction but also enhance the robustness.

The main contributions of this paper have three-folds:

1) The optimal projections are mutually orthogonal by adding generalized orthogonal constraint and the optimal solution is iteratively learned via Orthogonal Matching Pursuit (OMP). The OMP is able to help the model to obtain more discriminative information for effective feature selection.
2) The robustness is enhanced by not only utilizing $L_{2,1}$-norm instead of L_2-norm on the loss function to reduce the sensitivity to outliers, but also incorporating an elastic factor on the regression model to avoid the overfitting which usually arises in regression-based methods.
3) The proposed method can break through the small-class problem, which exists in the regression-based methods or the LDA-based methods, so as to obtain more projections to improve the performance of pattern recognition or classification. In addition, the convergence of the proposed algorithm is also proved.

The rest of this paper is organized as follows. The related works are presented in Section 2 while Section 3 gives the objective function and the optimal solution of the proposed method. Section 4 presents the theoretical analysis including the convergence of the proposed algorithm and its computational complexity. The experiment is analyzed in Section 5 and the conclusion of the paper is drawn in Section 6.

2. Related works

In this section, we first present the notations used in this paper and then briefly review some related works of the proposed method.

2.1. Notations

Matrices are written as bold italic uppercase letters, i.e. **A,B,X,Y**, etc., the vectors are represented as bold italic lowercase, i.e. **x,y,h**, etc. while scalars are denoted as italic lowercase or uppercase letters, i.e. *i, j, n, K*, α, β, etc. Let the data matrix denoted as $\mathbf{X}\in R^{n\times d}$ where n is the number of the samples and d denotes the feature dimension of each sample, i.e. each row of $\mathbf{X}$ is a sample vector. The label matrix is denoted as $\mathbf{Y}\in R^{n\times c}$ with $\mathbf{Y}_{ij}=1$ while $\mathbf{x}_1$ belongs to the jth class; $\mathbf{Y}_{ij}=0$, otherwise, where c is the number of the classes.

2.2. The LPP and its property

Locality Preserving Projections (LPP) [8] obtains the linear projection to optimally preserve the neighborhood structure of the data set. Supposed $\mathbf{a}$ is one of the projection that projects the data set $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ in R^n to be $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$, namely, $\mathbf{y}=\mathbf{a}^T\mathbf{x}_i$ [33]. Minimizing the following objective function provides the optimal solution of LPP:

$$\begin{aligned}\frac{1}{2}\sum_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\tilde{\mathbf{W}}_{ij} &= \frac{1}{2}\sum_{ij}\left(\mathbf{a}^T\boldsymbol{x}_i-\mathbf{a}^T\boldsymbol{x}_j\right)^2\tilde{\mathbf{W}}_{ij}\\ &= \mathbf{a}^T\mathbf{X}(\tilde{\mathbf{D}}-\tilde{\mathbf{W}})\mathbf{X}^T\mathbf{a}=\mathbf{a}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{a} \qquad (1)\end{aligned}$$

where $\tilde{\mathbf{W}}$ is a symmetric matrix and it elements are defined as follow:

$$\tilde{\mathbf{W}}_{ij}=\begin{cases}\exp(-||\mathbf{x}_i-\mathbf{x}_j||^2/t), & ||\mathbf{x}_i-\mathbf{x}_j||^2 \prec \varepsilon\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

where parameter $\varepsilon\in R$ denotes ε-neighborhoods, $\mathbf{x}_i$ is the K neighborhood to $\mathbf{x}_j$, $\mathbf{x}_j$ is the K neighborhood to $\mathbf{x}_i$ and $\tilde{\mathbf{D}}$ is a diagonal matrix and its elements are row (or column) sum of $\tilde{\mathbf{W}}$, namely, $\tilde{\mathbf{D}}_{ii}=\sum_i\tilde{\mathbf{W}}_{ij}$. $\tilde{\mathbf{L}}=\tilde{\mathbf{D}}-\tilde{\mathbf{W}}$ is the Laplacian matrix.

The optimal $\mathbf{a}$ is given by the minimum eigenvalue solution of the generalized eigenvalue problem [34]:

$$\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{a}=\lambda\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T\mathbf{a} \qquad (3)$$

Thus, the optimal projection matrix for LPP is $\mathbf{A}=[\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^l]$, where the vectors $\mathbf{a}^i(i=0,1,\ldots,l)$ are the eigenvectors corresponding to the smaller eigenvalues.

2.3. The $L_{2,1}$-norm and its property

Recently, the $L_{2,1}$-norm is widely used not only on loss function to reduce the sensitivity to outliers, but also on regularization term to obtain the joint sparsity for feature selection and feature extraction [35].

The $L_{2,1}$-norm of a matrix $\mathbf{M}\in R^{n\times m}$ is defined as

$$\|\mathbf{M}\|_{2,1}=\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}\mathbf{m}_{ij}^2}=\sum_{i=1}^{n}\left\|\mathbf{m}^i\right\|_2 \qquad (4)$$

$L_{2,1}$-norm was first introduced in [36] as the rotational invariant L_1 -norm, it is also fit for multi-task learning [37] and tensor factorization [38]. By using the $L_{2,1}$-norm on both loss function and regularization term, the $L_{2,1}$-norm based methods can easily obtain the discriminative vectors for feature selection by setting the elements in some rows of the projection matrix become zero. Since there is no squared operation in the $L_{2,1}$-norm based methods, they are more robust and less sensitive to outliers than those $L_{2,1}$-norm based methods. The difference between $L_{2,1}$-norm and L_1-norm is that using $L_{2,1}$-norm penalty on the regularization term can obtain jointly sparse projections to improve the performance of feature extraction and selection.

2.4. Orthogonal matching pursuit (OMP)

Orthogonal Matching Pursuit (OMP) is known as the canonical greedy algorithm for sparse learning [39]. It is believed to be valuable to explore alternative approaches like OMP which is not based on optimization to handle the signal recovery problem and compressive sensing [40]. Let $\mathbf{\Phi} \in R^{M \times N}$ denotes a matrix, $\tilde{\mathbf{y}}$ is a vector in R^M. The goal of OMP is to find a coefficient vector $\hat{\mathbf{x}} \in R^N$ with N (N≪M) nonzero terms so as to minimize $\|\tilde{\mathbf{y}} - \mathbf{\Phi}\hat{\mathbf{x}}\|_2$. OMP is usually used to obtain sparse representations for signal $\tilde{\mathbf{y}}$ in signal processing where $\tilde{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{x}}$ denotes an overcomplete dictionary for the signal space [41]. In compressing sensing, $\tilde{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{x}}$ represents compressive measurements of a sparse or nearly sparse signal $\hat{\mathbf{x}}$ to be discovered [42–43].

The major advantage of OMP is its low computational cost and easy implementation. The entire algorithm of OMP is shown in Algorithm 1. In spite of the simplicity, OMP is empirically competitive in approximation performance [44]. The key difference between MP and OMP is that with the finite dictionaries of size M, OMP converges in no more than M iterations while MP does not possess this property [41].

Algorithm 1 Orthogonal matching pursuit.

```
Input: $\mathbf{\Phi}$, $\tilde{\mathbf{y}}$, convergent criterion $\varepsilon$
initialize: $\mathbf{r}^0 = \tilde{\mathbf{y}}$, $\mathbf{x}^0 = \mathbf{0}$, $\mathbf{\Lambda}^0 = \phi$, $\ell = 0$
while not converged do
match: $\mathbf{h}^\ell = \mathbf{\Phi}^T \mathbf{r}^\ell$
identify:
$\mathbf{\Lambda}^{\ell+1} = \mathbf{\Lambda}^\ell \cup \{\arg max_j |\mathbf{h}^\ell(j)|\}$(if multiple maxima exist, choose only one)
update: $\mathbf{x}^{\ell+1} = \arg min_{\mathbf{z}:\mathrm{supp}(\mathbf{z}) \subseteq \mathbf{\Lambda}^{\ell+1}} \|\tilde{\mathbf{y}} - \mathbf{\Phi z}\|_2$, $\mathbf{r}^{\ell+1} = \tilde{\mathbf{y}} - \mathbf{\Phi x}^{\ell+1}$, $\ell = \ell + 1$
end while
output: $\hat{\mathbf{x}} = \mathbf{x}^\ell = \arg min_{\mathbf{z}:\mathrm{supp}(\mathbf{z}) \subseteq \mathbf{\Lambda}^\ell} \|\tilde{\mathbf{y}} - \mathbf{\Phi} z\|_2$
```

3. Robust jointly sparse regression

In this section, we first present the motivations and the novel definitions of the paper, and then introduce our previous work presented in [32]. Based on the groundwork, we finally propose our objective function of the model and its corresponding optimal solution. Some comparison and discussion versus other relevant methods are also made to demonstrate the novelty and the advantages of the proposed method.

3.1. The motivations and the novel definitions

Recently, many ridge regression based methods are proposed to deal with different problems, such as dimensionality reduction, feature selection and sparse learning. Among those methods, a method called robust feature selection (RFS) [45] is widely used in many cases. RFS utilizes the joint $L_{2,\,1}$-norm on both loss function and regularization term to guarantee the robustness to outliers and simultaneously obtain joint sparse protection for feature selection. The idea of the method is based on the property of $L_{2,\,1}$-norm. The RFS can be described as follows:

$$\min_{\mathbf{P}} \|\mathbf{XP} - \mathbf{Y}\|_{2,1} + \lambda \|\mathbf{P}\|_{2,1} \tag{5}$$

where $\mathbf{P} \in R^{d \times c}$ is the projection matrix and λ is the parameter of the modal.

Although RFS can improve the performance of pattern recognition and classification in some degree, it still has some drawbacks. First, it does not take the small-class problem into consideration. That is, the number of the learned projection is still limited by the number of the class, which makes it cannot obtain enough projections for further feature selection in some cases. Second, the loss function in RFS is just a modification of the classical ridge regression, which has the potential risk of overfitting. In this context, the performance of feature selection or classification will be affected. Therefore, we need to define a novel form for robust jointly sparse regression. Motivated by the property of the elastic factor in [46,47] and the decomposition of matrix, we propose the following model in [32]:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{h}} & \left\|\mathbf{XBA}^T + \mathbf{1h}^T - \mathbf{Y}\right\|_{2,1} + \alpha \|\mathbf{B}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{A}^T\mathbf{A} = \mathbf{I} \end{aligned} \tag{6}$$

where $\mathbf{B} \in R^{d \times k}$, $\mathbf{A} \in R^{c \times k}$ is the projection matrix and the auxiliary matrix, respectively. k is the number of objective dimension. $\mathbf{1} \in R^n$ is a constant vector with all elements equaling to 1 and $\mathbf{h} \in R^c$ is an elastic factor incorporated to the loss function to avoid the problem of overfitting so as to enhance the robustness of the method, α is the parameter to balance the two terms.

The common advantage of (6) and RFS is that they all use $L_{2,\,1}$-norm instead of L_2-norm as the basic measurement on the loss function, which provides less sensitivity to outliers. Also, by using the $L_{2,\,1}$-norm penalty on the regularization term, they can guarantee the joint sparsity of the projection matrix. That is, by making the elements in some rows of the learned projection matrix be 0, the important features corresponding to the nonzero elements can be highlighted and simultaneously the unusual information can be filtered out. The difference between (6) and RFS is that although they both can obtain the joint sparsity to improve the performance of feature selection and classification, model (6) has the potential to release the small-class problem and avoid the overfitting problem in ridge regression-based methods. Since the size of the projection matrix $\mathbf{B}$ is $d \times k$, with which we can obtain more projections by setting k as any number larger than the number of class. In this way, we can release the small-class problem and guarantee enough projections to construct a more discriminative subspace for effective feature selection or classification. Moreover, the elastic factor on the loss function is able to provide supplement for the fitting between $\mathbf{XBA}^T$ and $\mathbf{Y}$ such that the potential overfitting problem can be avoided.

Although (6) enjoys many merits which makes it capable to improve the performance of pattern recognition or classification in a degree, it still needs to improve because it neither considers the neighborhood relationship of the original data nor guarantees the orthogonality of the projection directions. As shown in [8,10,48], the information of manifold structure is of great importance in the case when data is not embedded in linear subspace. Also, the orthogonal projection is proved to be more discriminative in most cases. It can also release the correlation of the learned projections so as to construct a more effective and discriminative subspace for feature selection. Therefore, based on the regard of orthogonality of the projection direction, the locality preserving property and the sparsity, a more effective and compact model for feature selection is needed.

In this paper, we propose a novel method to release the problems in (6) and further improve the performance of recognition or classification tasks in face images or other images by integrating the locality of the data as a constraint on the generalized regression model.

3.2. The objective function of RJSR

By combining the joint sparsity, robustness and locality preserving property with the generalized orthogonality constraint on the projection matrix, we have the following objection function of the proposed RJSR:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{h}} & \left\|\mathbf{XBA}^T + \mathbf{1h}^T - \mathbf{Y}\right\|_{2,1} + \alpha \|\mathbf{B}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{A}^T\mathbf{A} = \mathbf{I}, \\ & \mathbf{B}^T\mathbf{X}^T\mathbf{WXB} = \mathbf{I} \end{aligned} \tag{7}$$

where $\mathbf{W} \in R^{n \times n}$ is the affinity graph as in LPP. The elements in $\mathbf{W}$ is defined as

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ is among } K \text{ nearest neighbors of } \mathbf{x}_j \\ & \text{or } \mathbf{x}_j \text{ is among } K \text{ nearest neighbors of } \mathbf{x}_i \\ 0 & \text{otherwise;} \end{cases}$$

By adding the constraint $\mathbf{B}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{B} = \mathbf{I}$ to the objective function, (7) is quite different from (6). (7) not only considers the orthogonality of the projection matrix $\mathbf{B}$, but also preserves the neighborhood relationship of the original data so that it can search a more discriminative subspace for effective feature selection or extraction. Differing from the previous regression methods, the most significant property of the proposed model is that it uses the locality preserving term as a constraint instead of a regularized term.

3.3. The optimal solution of RJSR

For the orthogonality constraint $\mathbf{B}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{B} = \mathbf{I}$, let $\bar{\mathbf{B}} = \sqrt{\mathbf{W}\mathbf{X}}\mathbf{B}$, we have $\bar{\mathbf{B}}^T\bar{\mathbf{B}} = \mathbf{I}$, $\mathbf{B} = \mathbf{C}\bar{\mathbf{B}}$, where $\mathbf{C} = (\sqrt{\mathbf{W}\mathbf{X}})^{-1}$. Therefore, (7) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{A},\bar{\mathbf{B}},\mathbf{h}} \left\|\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\right\|_{2,1} + \alpha\left\|\mathbf{C}\bar{\mathbf{B}}\right\|_{2,1} \\ &\text{s.t.} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}, \\ &\quad \bar{\mathbf{B}}^T\bar{\mathbf{B}} = \mathbf{I} \end{aligned} \tag{8}$$

From the definition of $L_{2,1}$-norm on the projection matrix $\bar{\mathbf{B}}$, we have the following diagonal matrix $\bar{\mathbf{D}}$ with elements on the diagonal defined as [26]:

$$\bar{\mathbf{D}}_{ii} = \frac{1}{2\left\|(\mathbf{C}\bar{\mathbf{B}})^i\right\|_2} \tag{9}$$

where $(\mathbf{C}\bar{\mathbf{B}})^i$ represents the ith row of matrix $\mathbf{C}\bar{\mathbf{B}}$.

Similarly, for $\|\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\|_{2,1}$, the corresponding diagonal matrix $\mathbf{D}$ can be defined as

$$\mathbf{D}_{ii} = \frac{1}{2\left\|(\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^i\right\|_2} \tag{10}$$

where $(\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^i$ represents the ith row of matrix $\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}$.

Therefore, the first part in (8) can be rewritten as

$$\begin{aligned} &\left\|\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\right\|_{2,1} = tr((\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^T \\ &\mathbf{D}(\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})) \end{aligned} \tag{11}$$

and the second part in (8) is written as

$$\alpha\left\|\mathbf{C}\bar{\mathbf{B}}\right\|_{2,1} = tr(\bar{\mathbf{B}}^T\mathbf{C}^T\bar{\mathbf{D}}\mathbf{C}\bar{\mathbf{B}}) \tag{12}$$

From (11) and (12), (8) can be rewritten as

$$\begin{aligned} &\left\|\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\right\|_{2,1} + \alpha\left\|\mathbf{C}\bar{\mathbf{B}}\right\|_{2,1} \\ &= tr((\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^T\mathbf{D}(\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})) \\ &\quad + \alpha tr(\bar{\mathbf{B}}^T\mathbf{C}^T\bar{\mathbf{D}}\mathbf{C}\bar{\mathbf{B}}) \end{aligned} \tag{13}$$

That is,

$$\begin{aligned} &\left\|\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\right\|_{2,1} + \alpha\left\|\mathbf{C}\bar{\mathbf{B}}\right\|_{2,1} \\ &= tr(\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{C}\bar{\mathbf{B}} + 2\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T\mathbf{D}\mathbf{1}\mathbf{h}^T\mathbf{A} - 2\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T\mathbf{D}\mathbf{Y}\mathbf{A} \\ &\quad + \mathbf{h}^T\mathbf{h}\mathbf{1}^T\mathbf{D}\mathbf{1} - 2\mathbf{h}^T\mathbf{Y}^T\mathbf{D}\mathbf{1} + \mathbf{Y}^T\mathbf{D}\mathbf{Y} + \alpha\bar{\mathbf{B}}^T\mathbf{C}^T\bar{\mathbf{D}}\mathbf{C}\bar{\mathbf{B}}) \end{aligned} \tag{14}$$

Since there are three variables (i.e. **A, B** and **h**) in the objective function in (7) and they cannot be obtained directly, we adopt an iterative approach to find the optimal solution. The detail of the approach is described as below.

h step: Set the derivatives of the objective function in (14) with respect to **h** equaling to zero, we have

$$\mathbf{h} = \frac{1}{s}(\mathbf{Y}^T - \mathbf{A}\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T)\mathbf{D}\mathbf{1} \tag{15}$$

where $s = \mathbf{1}^T\mathbf{D}\mathbf{1}$.

A step: Suppose the variable **h** and $\bar{\mathbf{B}}$ are fixed, the optimal solution of (14) can be obtained by minimizing the following problem:

$$\begin{aligned} &\min_{\mathbf{A}} tr\left[\mathbf{A}^T(\mathbf{h}\mathbf{1}^T - \mathbf{Y}^T)\mathbf{D}\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\right] \\ &\quad \text{s.t.} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I} \end{aligned} \tag{16}$$

Theorem 1. [6] *Given an orthogonal matrix* $\mathbf{G} \in R^{c \times k}$ *and a matrix* $\mathbf{Q} \in R^{c \times k}$ *with rankk. Consider the following optimization problem*

$$\begin{aligned} &\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} tr(\mathbf{G}^T\mathbf{Q}) \\ &\quad s.t. \quad \mathbf{G}^T\mathbf{G} = \mathbf{I}_k \end{aligned} \tag{17}$$

Suppose singular value decomposition (SVD) of $\mathbf{Q}$ *is* $\mathbf{Q} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^T$, *then* $\hat{\mathbf{G}} = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T$.

From Theorem 1, we can easily know that for given **h** and $\bar{\mathbf{B}}$ in (14), suppose the SVD of $(\mathbf{h}\mathbf{1}^T - \mathbf{Y}^T)\mathbf{D}\mathbf{X}\mathbf{C}\bar{\mathbf{B}}$ is

$$(\mathbf{h}\mathbf{1}^T - \mathbf{Y}^T)\mathbf{D}\mathbf{X}\mathbf{C}\bar{\mathbf{B}} = \breve{\mathbf{U}}\breve{\mathbf{D}}\breve{\mathbf{V}}^T \tag{18}$$

We have

$$\mathbf{A} = \breve{\mathbf{U}}\breve{\mathbf{V}}^T \tag{19}$$

B step: Since $\mathbf{B} = \mathbf{C}\bar{\mathbf{B}}$, we need to first obtain the optimal value of $\bar{\mathbf{B}}$, and then compute the optimal value of variable **B**.

Theorem 2. *Suppose* **S** *is any symmetric matrix, and the singular value decomposition of* **S** *is* $\overset{\frown}{\mathbf{U}}\overset{\frown}{\mathbf{D}}\overset{\frown}{\mathbf{U}}^T$, *then the following optimization problem*

$$\hat{\mathbf{M}} = \arg\min_{\mathbf{M}} tr(\mathbf{M}^T\mathbf{X}^T\mathbf{S}\mathbf{X}\mathbf{M} - 2\mathbf{M}^T\mathbf{X}^T\mathbf{Y}) \tag{20}$$

is equal to a quadratic form as

$$\hat{\mathbf{M}} = \arg\min_{\mathbf{M}} \left\| \left(\left(\overset{\frown}{\mathbf{D}}^{1/2} \overset{\frown}{\mathbf{U}}^T \right)^T \right)^{-1} \mathbf{Y} - \left(\overset{\frown}{\mathbf{D}}^{1/2} \overset{\frown}{\mathbf{U}}^T \right) \mathbf{X}\mathbf{M} \right\|_2^2 \tag{21}$$

Proof. The proof is in the Appendix.

From (14), by eliminating the terms without variable $\bar{\mathbf{B}}$, we have

$$tr\left[\bar{\mathbf{B}}^T\mathbf{C}^T(\mathbf{X}^T\mathbf{D}\mathbf{X} + \alpha\bar{\mathbf{D}})\mathbf{C}\bar{\mathbf{B}} - 2\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T\mathbf{D}(\mathbf{Y} - \mathbf{1}\mathbf{h}^T)\mathbf{A}\right] \tag{22}$$

That is,

$$tr(\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{H}\mathbf{C}\bar{\mathbf{B}} - 2\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{Z}) \tag{23}$$

where $\mathbf{H} = \mathbf{X}^T\mathbf{D}\mathbf{X} + \alpha\bar{\mathbf{D}}$, $\mathbf{Z} = \mathbf{X}^T\mathbf{D}(\mathbf{Y} - \mathbf{1}\mathbf{h}^T)\mathbf{A}$. Note that **H** is a symmetric matrix, from Theorem 2, we have

$$\begin{aligned} &\min_{\bar{\mathbf{B}}} tr(\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{H}\mathbf{C}\bar{\mathbf{B}} - 2\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{Z}) \\ &\quad \Leftrightarrow \min_{\bar{\mathbf{B}}} \left\| ((\hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T)^T)^{-1}\mathbf{Z} - (\hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T)\mathbf{C}\bar{\mathbf{B}} \right\|_2^2 \end{aligned} \tag{24}$$

where the SVD of **H** *is* $\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{U}}^T$.

Suppose $\mathbf{Y}^* = ((\hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T)^T)^{-1}\mathbf{Z}$, $\mathbf{X}^* = \hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T\mathbf{C}$, (24) can be rewritten as

$$\left\| ((\hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T)^T)^{-1}\mathbf{Z} - (\hat{\mathbf{D}}^{1/2}\hat{\mathbf{U}}^T)\mathbf{C}\bar{\mathbf{B}} \right\|_2^2 = \left\|\mathbf{Y}^* - \mathbf{X}^*\bar{\mathbf{B}}\right\|_2^2 \tag{25}$$

Therefore, the optimal value of variable $\bar{\mathbf{B}}$ can be obtained by solving the following optimization problem

$$\begin{aligned} &\hat{\bar{\mathbf{B}}} = \arg\min_{\bar{\mathbf{B}}} \left\|\mathbf{Y}^* - \mathbf{X}^*\bar{\mathbf{B}}\right\|_2^2 \\ &\quad s.t. \quad \bar{\mathbf{B}}^T\bar{\mathbf{B}} = I \end{aligned} \tag{26}$$

The optimal solution of (26) can be computed using Orthogonal Matching Pursuit (OMP) algorithm. Thus, we can obtain the optimal value of $\bar{\mathbf{B}}$ by calling Algorithm 1 iteratively.

For easy understanding, we conclude the procedures of finding the optimal solution of the objective function (7) or (8) in Algorithm 2.

Algorithm 2 Robust jointly sparse regression.

```
Input: The sample matrix $\mathbf{X} \in R^{n \times d}$, the label matrix $\mathbf{Y} \in R^{n \times c}$, the affinity graph $\mathbf{W} \in R^{n \times n}$, the objective dimension $k (k \le d)$, the parameter $\alpha$, the maximum iteration step $T$.
Output: Low-dimensional and orthogonal discriminative subspace $\mathbf{B} \in R^{d \times k} (k = 1, 2, \ldots, d)$.
Initialize $\boldsymbol{A} \in R^{c \times k}$, $\boldsymbol{B} \in R^{d \times d}$, $\bar{\mathbf{B}} \in R^{n \times k}$, $\mathbf{D} \in R^{n \times n}$, $\bar{\mathbf{D}} \in R^{d \times d}$, $\boldsymbol{h} \in R^{c \times 1}$ randomly, set $\mathbf{1} \in R^{n \times 1}$ with all elements equaling to 1, set $\mathbf{C} = (\sqrt{\mathbf{W}}\mathbf{X})^{-1}$.
for $i = 1, 2, \ldots, T$ do
    update $\mathbf{D}_{ii}$ using $\mathbf{D}_{ii} = \frac{1}{2\|(\mathbf{X}\mathbf{C}\bar{\mathbf{B}}\mathbf{A}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^i\|_2}$;
    update $\bar{\mathbf{D}}_{ii}$ using $\bar{\mathbf{D}}_{ii} = \frac{1}{2\|(\mathbf{C}\bar{\mathbf{B}})^i\|_2}$;
    update $\mathbf{h}$ using $\mathbf{h} = \frac{1}{s}(\mathbf{Y}^T - \mathbf{A}\bar{\mathbf{B}}^T\mathbf{C}^T\mathbf{X}^T)\hat{\mathbf{D}}\mathbf{1}$;
    compute $\mathbf{A}$ using $\mathbf{A} = \underset{\smile}{\mathbf{U}}\underset{\smile}{\mathbf{V}}^T$ in (19);
    compute $\bar{\mathbf{B}}$ using OMP algorithm of (26);
end
Return: $\mathbf{B} = \mathbf{C}\bar{\mathbf{B}}$
```

3.4. Comparison and discussion

LPP is able to preserve neighborhood information of the original data and transfer it into a new subspace, in which the proximity is remained. Neighborhood preserving embedding (NPE) [10] aims to preserve the local neighborhood structure of the data points by using local squares approximations to obtain the affinity weight matrix. But they all ignore the orthogonality of the projection direction.

As indicated in [49], adding orthogonality penalty on the projection directions can improve the ability of preserving the intrinsic geometric structure of the original data. Based on this regard, some orthogonal locality preserving methods are proposed. Orthogonal LPP (OLPP) [50] and orthogonal NPE (ONPE) [51] incorporate the orthogonality penalty to the projection directions to obtain orthogonal basis functions which is more powerful and discriminative than that of the LPP or NPE. In addition, the sparse preserving projection is considered to be quite powerful in sparse subspace learning. By minimizing the L_1-norm regularization-related objective function, the sparsity preserving projections (SPP) [52] and sparse locality preserving embedding (SpLPE) [21] are able to preserve the sparse reconstructive relationship of the original data. Base on the previous work, we can conclude that the information of local geometric structure, the orthogonality of the projection directions as well as the sparsity are consider quite valuable for effective feature selection.

Although LPP and NPE preserve the neighborhood relationship of the data, they do not have sparsity or orthogonality, not to mention that their robustness is not guaranteed. For OLPP and ONPE, even though they obtain the orthogonality of the projection direction, they still ignore the sparsity for effective interpretation of the results. SPP and SpLPE are the algorithms in terms of locality preserving with sparsity. However, since they utilize L_1-norm as the basic measurement on the regularization term, they cannot obtain the joint sparsity for effective feature selection. Differ from these methods, RJSR can not only guarantee the locality preserving ability and joint sparsity for effective feature selection, but also reduce the sensitivity to outliers by using $L_{2,1}$-norm instead of L_2-norm as the basic measurement on the loss function. What is more, the elastic factor incorporated in the loss function is able to release the potential overfitting problem such that the robustness of the model is further enhanced.

Compared with the jointly sparse learning methods, such as RFS, UDFS, USSL, FSSL, L21FLDA [53], the proposed RJSR is able to obtain more discriminative projection direction by adding the orthogonality constraint on the projection. Moreover, RJSR is capable to release the drawback in regression-based or LDA-based methods that the number of the learned projections is limited by the number of class, such that it can obtain more projections to improve the performance of pattern recognition or classification.

In conclusion, by integrating the locality preserving with joint sparsity as well as generalized orthogonality and considering the potential risk of overfitting and the small-class problem, RJSR is different from the existing locality preserving methods, the relevant jointly sparse learning methods and the regression-based methods.

4. Theoretical analysis

In this section, theoretical analysis including the convergence of the proposed algorithm and the corresponding computational complexity is presented.

4.1. The convergence

The following Lemmas are presented to help to verify the convergence of the proposed algorithm.

Lemma 1. *[26] Given any two nonzero constants* a *and* b*, it holds*

$$\sqrt{a} - \frac{a}{2\sqrt{b}} \le \sqrt{b} - \frac{b}{2\sqrt{b}} \tag{27}$$

Lemma 2. *[26] Suppose* $\mathbf{U} \in R$ *is any nonzero matrix, the following inequality holds*

$$\sum_i ||\mathbf{u}_t^i||_2 - \sum_i \frac{||\mathbf{u}_t^i||_2^2}{2||\mathbf{u}_{t-1}^i||_2} \le \sum_i ||\mathbf{u}_{t-1}^i||_2 - \sum_i \frac{||\mathbf{u}_{t-1}^i||_2^2}{2||\mathbf{u}_{t-1}^i||_2} \tag{28}$$

where $\mathbf{u}_t^i$ *and* $\mathbf{u}_{t-1}^i$ *represent the i-th row of matrix* U_t *and* U_{t-1}*, respectively.*

According to Lemmas 1 and 2, we obtain the following theorem.

Theorem 3. *Suppose all parameters on the objective function of (13) are given except* $\mathbf{A}, \bar{\mathbf{B}}, \mathbf{h}, \mathbf{D}, \bar{\mathbf{D}}$*, the proposed Algorithm 2 will monotonically decrease the objective function value during each iteration and finally provide a local optimal solution for the objective function.*

Proof. The proof is in the Appendix.

4.2. Computational complexity analysis

Suppose the dimension of training samples is d and the number of iteration times is T. There are five variables (i.e. $\mathbf{A}, \bar{\mathbf{B}}, \mathbf{h}, \mathbf{D}, \bar{\mathbf{D}}$) need to update iteratively using the proposed Algorithm 2. Computing $\mathbf{h}$ in (15) needs $O(4d^2)$ while it takes $O(d^3)$ to obtain $\mathbf{A}$ from SVD of $(\mathbf{h}\mathbf{1}^T - \mathbf{Y}^T)\mathbf{D}\mathbf{X}\mathbf{C}\bar{\mathbf{B}}$. From (28), the main cost of OMP is $O(kNd)$ [40], where k is the number of sparse approximation in $\bar{\mathbf{B}}$ and N is the number of samples. Computing $\mathbf{D}$ in (10) and $\bar{\mathbf{D}}$ in (9) all need $O(dk)$ and $O(Nc)$. To sum up, the major computational complexity of the proposed algorithm is up to $O(T(d^3 + 4d^2 + kNd + dk + Nc))$.

5. Experiments

In this section, the COIL100 dataset (http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php) is first used to evaluate the performance of the proposed RJSR when images are with rotational variations. Then experiments on other five databases are conducted to evaluate the performance of the proposed method on face databases (i.e. Yale [54], ORL [55] and AR [56] Dataset) and non-face databases (i.e. hyperspectral images from the University of Pavia Data Set (PaviaU) [57], Binary alpha dataset). Plus, the robustness and the flexibility of the proposed method are also evaluated under the case when face images are corrupted by block or noise.

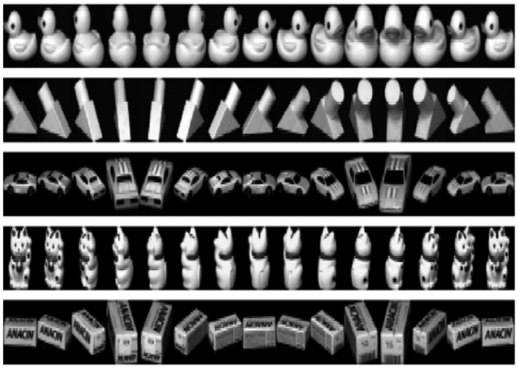


Fig. 1. Sample images on COIL100 database.

Other comparative methods including the locality preserving learning methods LPP [8], OLPP [50] and the fast and orthogonal LPP (FOLPP) [58], the $L_{2,1}$-norm relevant methods like the joint embedding learning and sparse regression (JELSR) [59], the L_1-norm based methods including approximate orthogonal sparse embedding (SLE) [60] and outlier-resisting graph embedding (LPP-L1) [61], are also conducted on all experiments.

5.1. Experiments on manifold learning

In this subsection, the proposed RJSR is applied to verify the performance on keeping the local structure of the data. The COIL100 contains 7200 images from 100 classes and each class has 72 images with rotational variations. In our experiment, all images are converted into gray images with 32 × 32 pixels. Fig. 1 shows the sample images on this database.

On COIL100 dataset, we first learn the projection space by using the proposed RJSR. Then all the images are used to map onto the space. Fig. 2 shows the results that the images mapped onto the two-dimensional plane described by two coordinates of the proposed RJSR. Some representative images are shown closed to the data points in the figures. As can be seen, the viewing point of the images changes smoothly, which indicates that RJSR is able to preserve the local structure of the data.

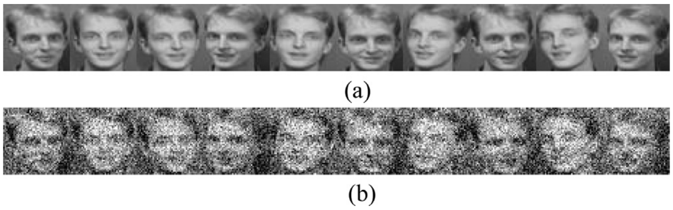


Fig. 3. Sample images on ORL database. (a) Original facial images. (b) Facial images corrupted by gaussian noise.

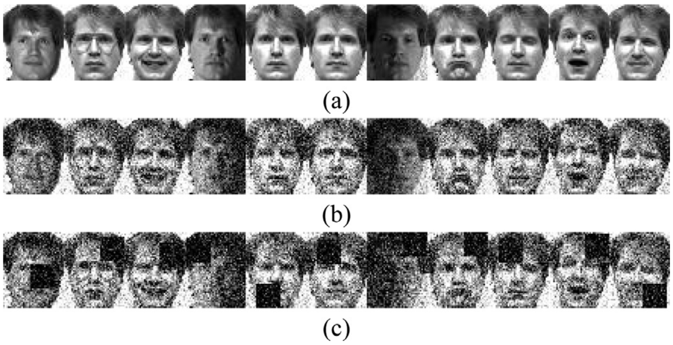


Fig. 4. Sample images on Yale database. (a) Original facial images. (b) Facial images corrupted by gaussian noise. (c) Facial images corrupted by random block with size 15 × 15.

5.2. Experiments on face database

The ORL database [55] has total 400 images from 40 individuals. The images are with variation in pose, facial expression and detail. Fig. 3 (a) shows the sample image of one individual on ORL database.

The Yale dataset [54] contains 165 images from 15 individuals. There are facial expression and lighting conditions various in those images. Fig. 4 (a) presents the sample image of one person on Yale database.

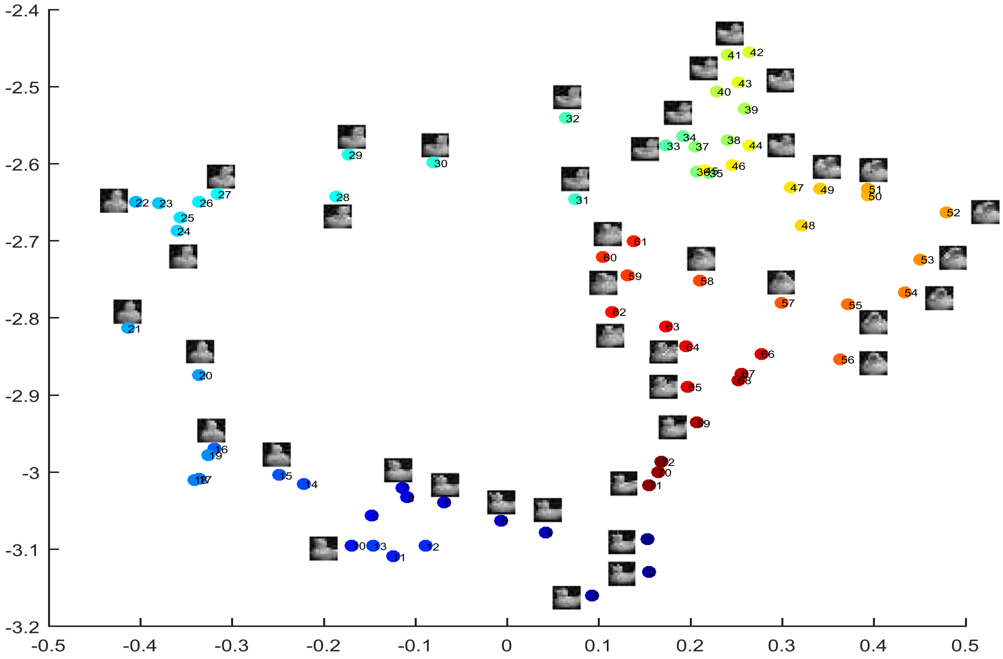


Fig. 2. A two-dimensional representation of the images on COIL100 using the proposed RJSR. Some representative images are shown closed to the corresponding data points.

Fig. 5. Sample images on AR database.

The AR dataset [56] is consist of 2400 images from 120 individuals (65 men and 55 women). The images vary as follows: neutral expression, smiling, angry, screaming, left light on, right light on, all sides light on, wearing sun glasses, wearing sun glasses and left light on, wearing sun glasses and right light on. Fig. 5 shows the sample images on this dataset.

(1) Sensitivity analysis of parameter settings

In this section, experiments are conducted to analyze the influence of the parameters to the proposed RJSR. Since α is used to control and balance the behavior of the proposed algorithm, it needs to be analyzed on all databases. The proposed modal also takes the local structure of the data points into consideration by constructing the affinity graph which is related to the *K*nearest neighbors. Based on this regard, the experiments on all databases were conducted 10 times independently to explore the optimal value of α and *K*.

We explore the best value of α from the area of $[10^{-9}, 10^{-8}, \ldots, 10^{9}]$ while *K* is set as a constant. Similarly, the variation of *K* from 1 to 30 was explored while α was fixed. The performance versus the variation of parameter α and *K* are shown in Fig. 6. From Fig. 6. (a), we can know that the variation of α would affect the performance of RJSR in a degree. The optimal value of α on ORL, Yale, AR, Binary database is $[10^{6}, 10^{7}]$, $[10^{5}, 10^{6}, \| 10^{9}]$, $[10^{6}, 10^{7}, \| 10^{9}]$, $[10^{1}, 10^{2}, 10^{8}]$, respectively. For PaviaU database, it is obvious that the performance of RJSR is not sensitive to the value of α. Although we can set $\alpha = [10^{-9}, 10^{-8}, \ldots, 10^{9}]$ to conduct the experiment on PaviaU database, we select a sub-area of $[10^{5}, 10^{6}, \| 10^{9}]$ for simplicity. Fig. 6. (b) indicates that the optimal value of *K* on ORL, Yale, Binary and PaivaU database is 4, 4, 5, 5, respectively, while it can be set as any integer from 1 to 30 on AR database.

(2) Experiment settings

The images on all databases are cropped and aligned automatically. In each experiment, the images are divided into training set and testing set. The training set is formed by randomly selecting *l* images from each class while the rest of images are used as testing. For fair comparison, the experiments on all databases are conducted 10 times so as to obtain average recognition rate to evaluate the performance of the proposed method and the comparative methods. Since the dimension of the images is usually very high, it tends to be time-consuming to perform feature selection directly. Base on this regard, we divide our experiments into three steps. First, we use PCA as the pre-processing to conduct dimensionality reduction. Second, the training set is used by the feature selection or dimensionality reduction algorithms to learn a low dimensional embedding space. Finally, the testing set is mapped onto the learned subspace and the nearest neighbor classifier is utilized for classification. For easy understanding, the average recognition rate versus the variation of dimension is plotted in figures while the maximum average recognition rate and the corresponding dimension as well as the standard deviations are shown in tables. For the comparative methods which have input parameters, we follow the setting of the parameters as introduced in the original paper.

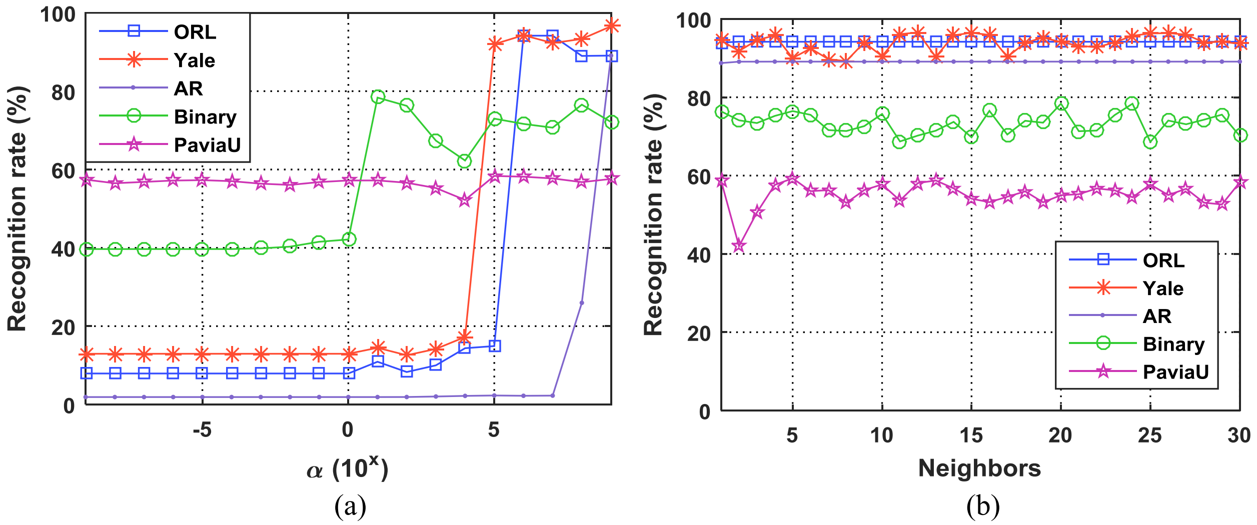
In this experiment, $l (l = 4, 5, 6)$ images of each individual are randomly selected to form the training set and the rest is used for testing. The performance of the proposed method and the comparative methods on Yale, ORL, AR database are shown in Figs. 7 and 8 (a) and Tables 1–3. The experimental results indicate that the proposed method can obtain the best performance.

5.3. Experiments on non-face databases

In this experiment, we evaluate the performance of the proposed method in terms of hyperspectral images and digits and letters images. Base on this regard, the University of Pavia Data Set (PaviaU) and the Binary alpha dataset are used in our experiments.

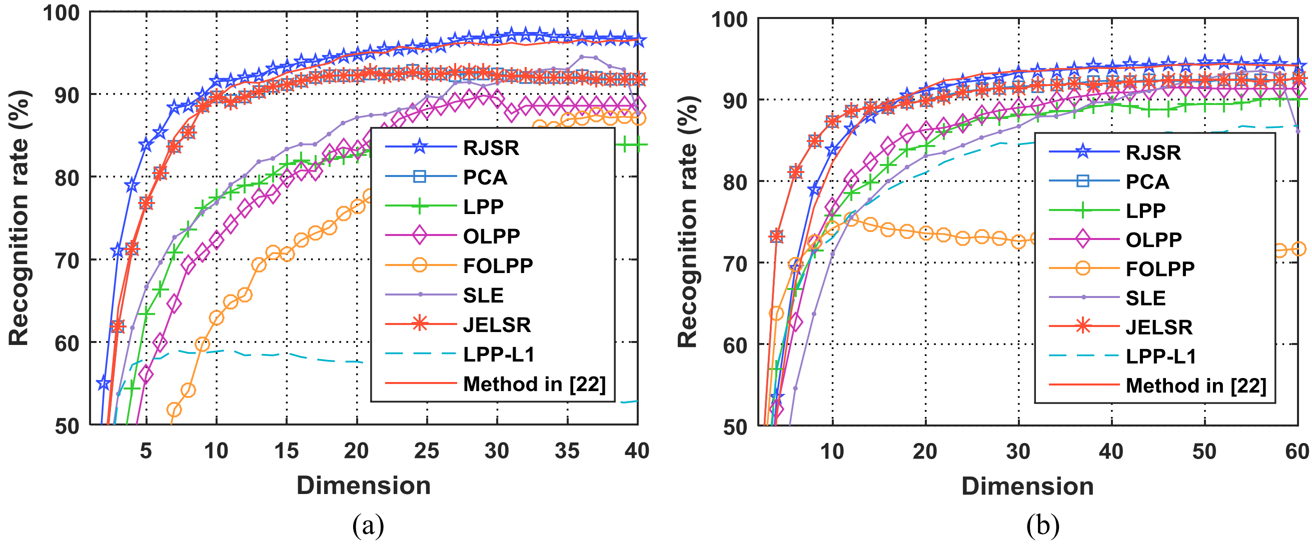
The Binary alpha dataset is comprised of digits from “0″ to “9″ and letters from “A” to “Z” and every class has 39 images with 46×46 pixels. The sample image on this dataset is presented in Fig. 9. (http://www.cs.nyu.edu/roweis/data.html).

The PaviaU dataset [57] was obtained by the ROSIS sensor during a flight campaign over Pavia University. There are 103 bands without noise left for our experiment. The dataset contains 9 ground truth classes: asphalt, meadows, gravel, trees, metal sheets, soil, bitumen, bricks and shadows. Fig. 10 illustrates the sample image in false color as well as the corresponding ground truth.



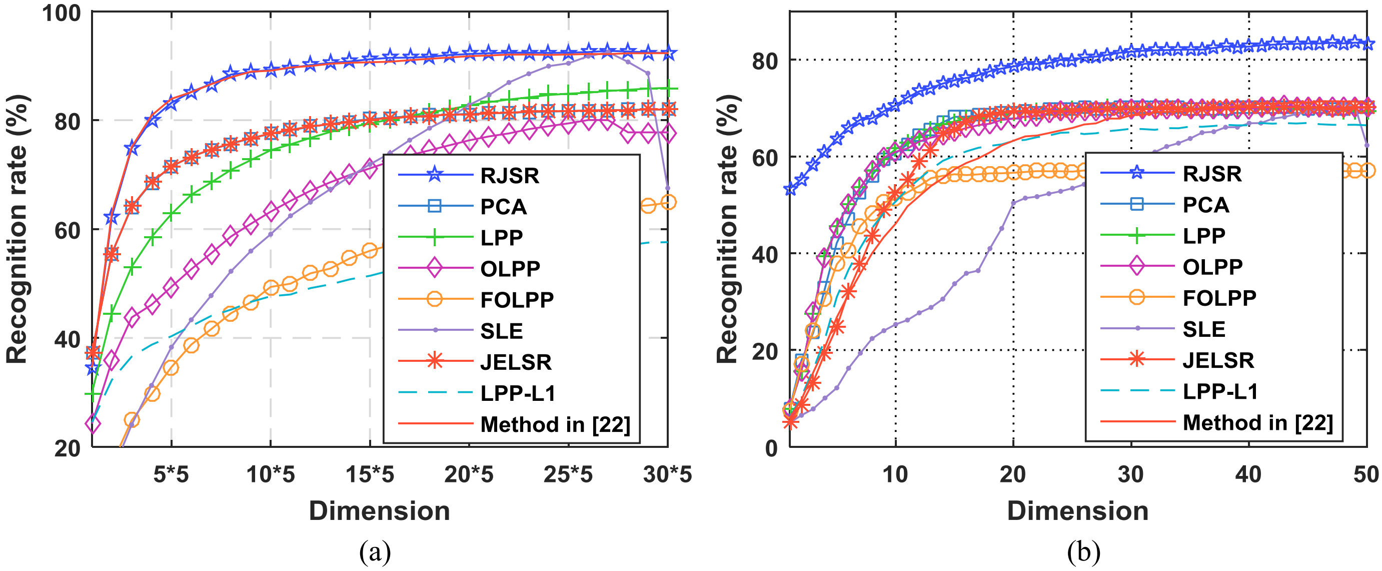
(a) (b)

Fig. 6. The recognition rate versus the variation of (a) regularization term coefficients (b) *K*-neighbors.



(a) (b)

Fig. 7. Experimental results on facial image datasets. (a) Yale database (*l*= 4), (b) ORL database (*l*= 4) .



(a) (b)

Fig. 8. Experimental results on image datasets. (a) AR database (*l*= 4), (b) Binary database (*l*= 20) .

Table 1
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on Yale datasets.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
4	92.76 ± 13.95 24	84.76 ± 14.77 26	89.81 ± 30.34 29	77.52 ± 10.90 12	94.48 ± 12.91 36	91.81 ± 26.41 40	59.05 ± 8.21 7	96.57 ± 12.74 36	**97.14 ± 12.46** **31**
5	93.44 ± 12.96 20	90.00 ± 13.86 24	92.00 ± 24.90 30	79.67 ± 12.04 14	95.22 ± 12.74 37	93.00 ± 26.95 40	64.00 ± 8.32 9	96.78 ± 13.84 36	**97.11 ± 12.44** **37**
6	95.20 ± 12.87 20	95.07 ± 14.72 20	94.40 ± 15.74 32	83.47 ± 11.77 12	97.20 ± 12.52 34	95.20 ± 26.88 40	87.60 ± 12.90 39	97.60 ± 11.69 40	**98.27 ± 11.70** **30**

Table 2
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on ORL datasets.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
4	92.54 ± 9.91 30	90.17 ± 11.05 29	91.46 ± 29.17 24	75.33 ± 7.47 6	93.50 ± 16.02 28	92.54 ± 9.98 30	86.79 ± 10.59 30	94.38 ± 14.36 26	**94.50 ± 13.69** **26**
5	94.80 ± 9.59 28	93.95 ± 11.12 30	94.30 ± 20.48 26	79.75 ± 8.05 7	95.70 ± 15.51 28	94.75 ± 9.57 25	90.45 ± 10.13 30	96.50 ± 14.82 28	**96.75 ± 13.50** **27**
6	96.31 ± 9.57 26	95.88 ± 10.01 30	96.13 ± 10.67 28	78.19 ± 7.84 6	97.19 ± 16.12 27	96.31 ± 9.57 23	92.75 ± 10.35 28	97.44 ± 14.35 30	**97.50 ± 14.36** **27**

Table 3
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on AR datasets.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
4	79.45 ± 8.07 150	82.59 ± 13.13 150	77.46 ± 21.94 125	61.77 ± 15.30 150	90.39 ± 24.03 135	79.45 ± 8.07 150	56.54 ± 10.21 150	90.65 ± 11.65 145	**90.76 ± 11.49** **135**
5	81.96 ± 9.50 150	85.94 ± 13.83 150	80.03 ± 18.23 130	64.82 ± 15.77 150	92.51 ± 24.15 135	81.98 ± 9.50 145	57.59 ± 10.80 150	92.31 ± 10.28 150	**92.74 ± 10.91** **135**
6	84.55 ± 8.92 150	89.46 ± 12.61 150	83.90 ± 12.09 140	64.67 ± 16.09 150	93.79 ± 24.39 135	84.55 ± 8.90 150	58.40 ± 8.66 150	93.85 ± 10.61 150	**93.90 ± 10.08** **140**

Fig. 9. Sample images on Binary database.

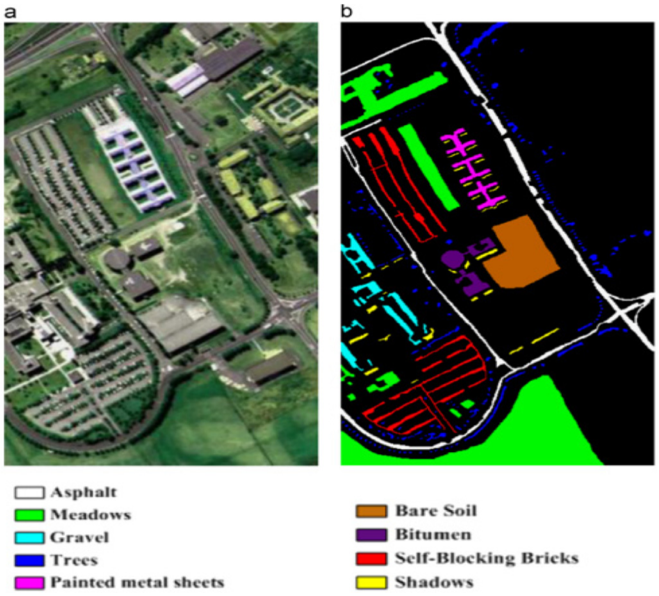


Fig. 10. Sample images on Pavia University dataset. (a) three-feature color composite image. (b) Ground truth: asphalt, meadows, gravel, trees, painted metal sheets, bare soil, bitumen, self- blocking bricks, and shadows.

On Binary alpha dataset, $l(l = 10, 15, 20)$ samples of each class are randomly selected to form the training set while the remaining images are used for testing. The recognition rate of different methods versus the variation of dimension is shown in Fig. 8. (b) with $l = 20$. Table 4 lists the best performance and the corresponding dimension as well as standard deviation of different methods.

On PaviaU dataset, $l(l = 5, 7, 9)$ images of each class are randomly selected as the training while the rest images compose the testing set. The recognition rate versus the dimensionality reduction is demonstrated in Fig. 11. (a) with $l = 9$. The best recognition rate versus the dimension and the standard deviation are shown in Table 5.

From Fig. 8. (b) and Table 6, we can easily know that the proposed method is able to obtain about at least 10% more than all the comparative methods. Fig. 11(a) and Table 5 indicate that the proposed method is also more effective and able to obtain much higher recognition rate than the comparative methods for hyperspectral images. All the results prove the effectiveness of the proposed RJSR in the non-face images.

5.4. The evaluation of robustness

In order to evaluate the robustness of the proposed method in the case when face images are corrupted by block or noise, we conduct a series of experiments on face database.

(1) Images corrupted by gaussian noise

In this subsection, the gaussian noise is added to all images including training set and testing set on Yale and ORL database. Figs. 3 (b) and 4 (b) present the facial images corrupted by Gaussian noise on ORL, Yale database, respectively.

Tables 6 and 7 list the best recognition rate and the corresponding dimension as well as the standard deviation of different methods. From the results, we can know that the proposed method is more robust than other methods in most cases.

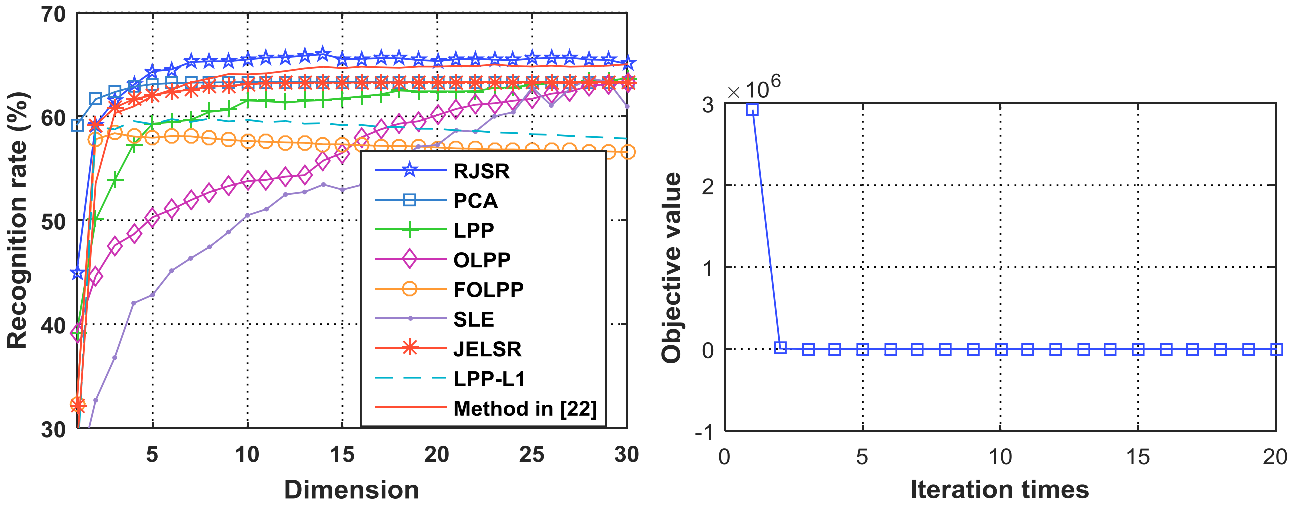
(2) Images with random block corruption

To evaluate the robustness of RJSR in terms of block corruption, we add random block on each of the facial image with size 5 × 5, 10 × 10 and 15 × 15. Fig. 4. (c) shows the sample images on Yale database with block size 15 × 15. The best performance of the proposed RJSR and the comparative methods are listed in Table 8. Figs. 11(b)–13 demonstrate the convergence curve of RJSR on face database and non-face database, respectively. The experimental results clearly illustrate the superiority and the fast convergence of the proposed method.

Therefore, we can conclude that RJSR is able to perform better than the locality learning methods (i.e. LPP, OLPP and FOLPP), the sparse graph embedding methods (i.e. SLE and JELSR) in most cases of our experiments.

Table 4
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on binary datasets.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
10	64.31 ± 12.98 26	61.48 ± 11.98 35	63.41 ± 12.30 46	48.94 ± 9.17 24	63.20 ± 16.20 48	64.30 ± 15.19 27	59.20 ± 13.08 49	64.26 ± 15.12 47	**72.27 ± 9.38** **49**
15	68.07 ± 13.70 42	67.40 ± 12.97 36	67.95 ± 13.06 39	52.56 ± 9.75 26	68.31 ± 18.69 48	68.19 ± 16.92 34	64.47 ± 14.69 32	68.78 ± 16.59 49	**78.65 ± 9.17** **48**
20	70.28 ± 14.07 33	70.23 ± 13.55 30	70.48 ± 13.37 43	57.24 ± 11.19 28	70.88 ± 20.36 48	70.31 ± 17.31 39	66.93 ± 14.96 45	71.43 ± 17.50 47	**83.51 ± 5.05** **49**



(a) (b)

Fig. 11. (a) Experimental results on PaviaU dataset (l= 9). (b) The convergence curve of RJSR on (a) PaviaU database.

Table 5
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on PaviaU datasets.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
5	60.66 ± 29.19 20	54.53 ± 3.40 30	60.66 ± 4.71 30	56.94 ± 4.92 3	60.67 ± 6.88 29	60.66 ± 4.93 30	58.19 ± 3.50 7	61.69 ± 4.42 30	**63.19 ± 2.45** **15**
7	60.07 ± 0.33 11	57.08 ± 5.07 27	60.07 ± 5.11 30	55.93 ± 4.21 3	60.61 ± 7.24 28	60.08 ± 4.35 23	57.35 ± 3.37 3	61.98 ± 4.18 14	**62.32 ± 2.90** **28**
9	63.33 ± 0.76 12	63.65 ± 3.48 30	63.32 ± 4.80 30	58.38 ± 4.60 3	63.65 ± 9.10 28	63.32 ± 5.30 28	59.79 ± 6.30 8	65.02 ± 5.76 30	**65.98 ± 3.73** **14**

Table 6
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on ORL datasets with Gaussian noise.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
4	86.50 ± 8.83 10	70.33 ± 8.15 10	82.54 ± 25.51 20	71.62 ± 7.31 8	72.87 ± 10.96 29	86.25 ± 8.57 7	80.88 ± 10.08 25	**87.67 ± 18.36** **30**	**87.67 ± 14.25** **30**
5	90.40 ± 9.20 10	81.55 ± 9.57 11	89.25 ± 21.57 18	75.30 ± 8.04 7	82.30 ± 12.23 15	90.00 ± 9.06 9	87.90 ± 11.70 30	91.90 ± 19.27 30	**92.25 ± 17.78** **28**
6	93.06 ± 10.41 21	88.00 ± 9.50 16	92.31 ± 12.15 23	77.56 ± 8.49 6	88.50 ± 12.18 15	91.75 ± 10.21 30	88.81 ± 11.51 28	93.94 ± 18.68 29	**94.13 ± 18.12** **29**

Table 7
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on Yale datasets with Gaussian noise.

Training samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
4	90.00 ± 13.20 19	71.90 ± 12.77 22	80.10 ± 17.09 22	75.33 ± 11.51 11	87.14 ± 11.37 39	89.90 ± 13.20 22	54.38 ± 8.47 9	93.52 ± 17.43 39	**94.00 ± 14.13** **39**
5	90.11 ± 12.55 15	77.89 ± 13.24 16	85.78 ± 18.59 25	77.89 ± 10.57 17	87.89 ± 10.91 19	89.89 ± 12.41 18	58.22 ± 7.18 9	93.78 ± 17.56 37	**94.00 ± 13.81** **33**
6	92.27 ± 13.01 22	86.13 ± 14.26 16	90.53 ± 19.32 20	82.27 ± 12.85 11	91.07 ± 11.87 16	91.73 ± 12.78 16	87.07 ± 13.49 39	95.33 ± 14.09 38	**96.27 ± 13.84** **36**

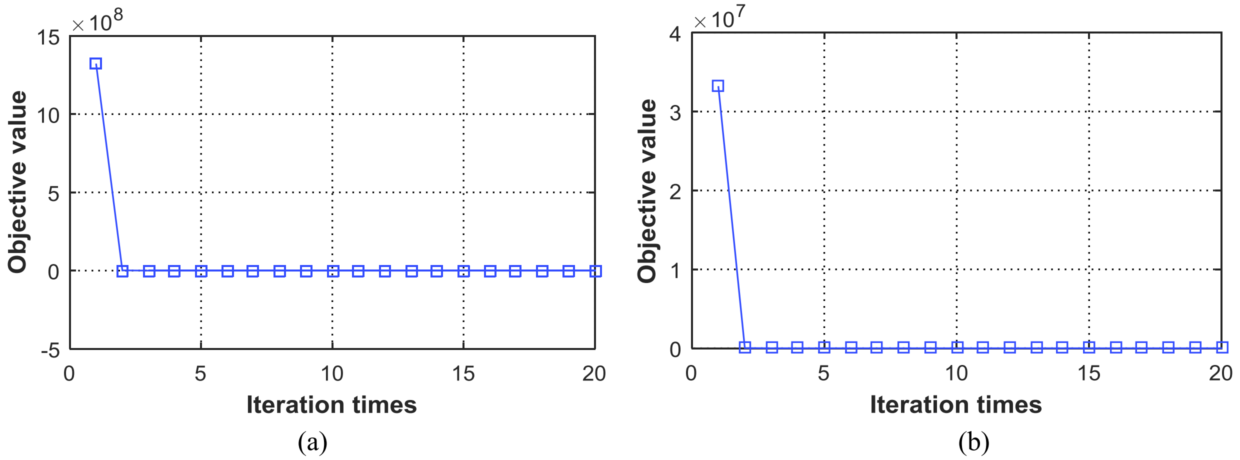
5.5. Experimental results and discussions

According to the experimental results of the proposed method and the comparative methods on face databases and the Binary alpha database as well as the hyperspectral database, we have the following interesting points:

1) RJSR obtains the best performance in almost all experiments. The potential reason for this phenomenon is that RJSR considers the locality of the original data, the orthogonality and the joint sparsity of projections. Based on the groundwork, it can obtain more discriminative information for effective feature selection and extraction.
2) Although RJSR and OLPP as well as FOLPP take both of locality and orthogonality of the projection direction into consideration, RJSR surpass the other two methods in most cases. One reason is that RJSR utilizes the joint $L_{2,1}$-norm on the regularization term to perform jointly sparse feature selection for efficient pattern recognition or classification. The other reason is that the added noise has heavy impact on the FOLPP algorithm, leading to low recognition rates in all the databases.
3) Moreover, since RJSR uses $L_{2,1}$-norm instead of L_2-norm as the basic measurement on the loss function, it is less sensitive to

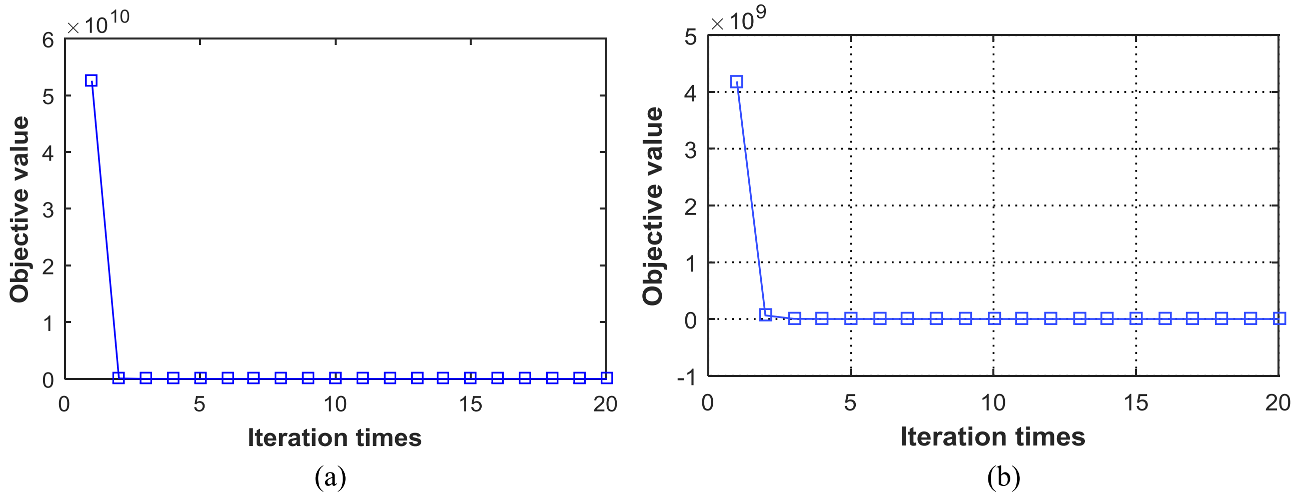
Table 8
Best recognition rate, standard deviation and the corresponding feature dimensions of different algorithms on Yale datasets with gaussian noise and block cover.

Block Size	Training Samples	PCA	LPP	OLPP	FOLPP	SLE	JELSR	LPP-L1	Method in [32]	RJSR
5*5	4	88.67 ± 13.31 21	70.29 ± 12.63 22	88.57 ± 30.37 25	72.86 ± 10.38 13	88.29 ± 13.18 39	88.00 ± 13.35 32	61.57 ± 6.58 8	92.76 ± 20.25 39	**93.43 ± 16.17** **38**
	5	90.56 ± 14.00 23	81.22 ± 13.67 24	90.19 ± 26.68 26	72.56 ± 10.90 12	91.11 ± 14.01 28	90.00 ± 13.84 21	71.78 ± 7.25 8	91.22 ± 17.62 39	**93.89 ± 15.99** **35**
	6	92.53 ± 14.65 20	87.87 ± 16.52 23	89.67 ± 19.61 28	79.73 ± 13.42 14	93.20 ± 14.05 24	91.07 ± 14.45 26	85.60 ± 14.66 32	93.20 ± 15.40 39	**95.60 ± 14.95** **37**
10*10	4	74.76 ± 14.05 22	50.19 ± 9.13 8	66.83 ± 23.00 25	53.43 ± 7.94 13	74.86 ± 11.45 39	73.81 ± 13.92 31	59.14 ± 4.77 6	84.67 ± 17.64 40	**87.71 ± 16.52** **38**
	5	80.22 ± 15.42 26	65.44 ± 13.28 18	73.97 ± 22.97 27	58.67 ± 9.00 15	79.11 ± 13.29 27	78.22 ± 15.60 34	62.22 ± 4.97 7	89.33 ± 18.54 32	**89.44 ± 15.87** **33**
	6	82.80 ± 15.95 37	77.60 ± 14.06 20	82.10 ± 17.81 28	64.40 ± 11.51 12	83.47 ± 13.07 39	82.13 ± 15.80 34	73.20 ± 13.45 27	91.87 ± 17.88 37	**92.00 ± 16.16** **33**
15*15	4	61.90 ± 12.04 40	46.57 ± 8.51 31	51.16 ± 20.56 25	38.38 ± 6.10 17	64.10 ± 12.26 39	61.90 ± 12.16 40	52.90 ± 2.30 5	**82.00 ± 18.46** **40**	79.81 ± 17.41 37
	5	64.67 ± 14.07 35	55.56 ± 10.28 27	61.11 ± 18.27 29	44.33 ± 6.79 18	67.67 ± 13.34 35	64.00 ± 14.30 38	57.78 ± 3.10 8	**84.44 ± 18.08** **39**	83.89 ± 18.04 35
	6	67.07 ± 15.24 31	61.73 ± 12.64 23	66.67 ± 16.00 30	45.73 ± 6.43 14	72.40 ± 14.69 39	67.73 ± 15.29 33	59.33 ± 12.50 35	86.80 ± 18.68 40	**87.33 ± 16.19** **32**



(a) (b)

Fig. 12. The convergence curve of RJSR on (a) ORL database, (b) Yale database.



(a) (b)

Fig. 13. The convergence curve of RJSR on (a) AR database, (b) Binary database.

outliers. The elastic factor on the loss function also avoids the overfitting in the regression model and the experimental results on face images corrupted with noise or block prove the robustness and flexibility of RJSR.

4) As shown in Fig. 8. (b), the curve of the recognition rate keeps growing more than 36 dimensions, which means that RJSR breaks through the small-class problem and obtains more than 36 (the number of class) projections to perform effective feature selection. The conclusion also agrees with the results shown in Table 4.

5) All the experimental results show that the method in [32] and the proposed RJSR can obtain better performance than other comparative methods. The potential reason is that they both use $L_{2,1}$-norm as the basic measurement and consider the overfitting problem. However, RJSR outperforms the method in [32] in most cases. That is because RJSR can obtain orthogonality property of the projection and simultaneously preserve the neighborhood structure of the data.

6. Conclusion

In this paper, we propose a robust jointly sparse regression for effective feature selection. By combining the locality of the manifold structure of the original data, the orthogonality and the joint sparsity of the projection, RJSR is able to obtain more discriminative information for image recognition or classification tasks. In addition, RJSR can also release the small-class problem to obtain more projections via the designed loss function. The proposed optimization problem can be solved by an iterative algorithm. The theoretical analysis including the convergence of the proposed algorithm and the computational complexity are presented. Experiments on face images, hyperspectral images and digits and letters images are conducted to evaluate the performance of RJSR. The experimental results indicate that RJSR can outperform the locality based methods (LPP, OLPP, FOLPP), the joint sparsity learning methods (JELSR) and the L_1-norm based methods (SLE, LPP-L1) with strong robustness.

The proposed RJSR is an iteration algorithm and the computational complexity is higher than the traditional methods, such as PCA, LPP and RR. Therefore, it would be meaningful to reduce the computation cost of RJSR.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (Grant 61573248, Grant 61732011 and Grant 61802267), Research Grant of The Hong Kong Polytechnic University (Project Code: G-UA2B), the Guangdong Natural Science Foundation (Project 2017A030313367 and Project 2017A030310067), and Shenzhen Municipal Science and Technology Innovation Council (No. JCYJ20170302153434048, No. JCYJ20180305124834854 and No. JCYJ20160429182058044).

Appendix

Proof of Theorem 2. In (20), suppose the SVD of $\mathbf{S}$ is $\mathbf{S}=\overset{\frown}{\mathbf{U}}\overset{\frown}{\mathbf{D}}\overset{\frown}{\mathbf{U}}^T$, then we have the following equality

$$\begin{aligned}&tr\left(\mathbf{M}^T\mathbf{X}^T\mathbf{S}\mathbf{X}\mathbf{M}-2\mathbf{M}^T\mathbf{X}^T\mathbf{Y}\right)\\&=tr\left(\mathbf{M}^T\mathbf{X}^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)\mathbf{X}\mathbf{M}-2\mathbf{M}^T\mathbf{X}^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\left(\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\right)^{-1}\mathbf{Y}\right)\end{aligned}\tag{29}$$

From (29), we have

$$\left\|\left(\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\right)^{-1}\mathbf{Y}-\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)\mathbf{X}\mathbf{M}\right\|_2^2=tr\begin{bmatrix}\mathbf{Y}^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^{-1}\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^{-T}\mathbf{Y}-2\mathbf{M}^T\mathbf{X}^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\left(\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\right)^{-1}\mathbf{Y}\\+\mathbf{M}^T\mathbf{X}^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)\mathbf{X}\mathbf{M}\end{bmatrix}\tag{30}$$

Since $\mathbf{Y}^T(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T)^{-1}(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T)^{-T}\mathbf{Y}$ is a constant term and it can be ignored. From (29) and (30), we can know that the optimal solution in (20) is equal to that in (21). That is,

$$\min_{\mathbf{M}}\left\|\left(\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)^T\right)^{-1}\mathbf{Y}-\left(\overset{\frown}{\mathbf{D}}^{1/2}\overset{\frown}{\mathbf{U}}^T\right)\mathbf{X}\mathbf{M}\right\|_2^2\Leftrightarrow\min_{\mathbf{M}}tr\left(\mathbf{M}^T\mathbf{X}^T\mathbf{S}\mathbf{X}\mathbf{M}-2\mathbf{M}^T\mathbf{X}^T\mathbf{Y}\right)$$

□

Proof of Theorem 3. We define the objective function in (13) as $F(\mathbf{A},\bar{\mathbf{B}},\mathbf{h},\mathbf{D},\bar{\mathbf{D}})$ for simplicity. Since at the $(t-1)$-th iteration, $\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_{t-1},\mathbf{D}_{t-1}$ and $\bar{\mathbf{D}}_{t-1}$ are obtained, according to (15), we have

$$F\left(\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_t,\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\le F\left(\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_{t-1},\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\tag{31}$$

As shown in (19), $\mathbf{A}_t$ can be obtained by singular value decomposition of $(\mathbf{h}_t\mathbf{1}^T-\mathbf{Y}^T)\mathbf{D}_{t-1}\mathbf{X}\mathbf{C}\bar{\mathbf{B}}_{t-1}$, the objective function value will be further decreased, then it goes

$$F\left(\mathbf{A}_t,\bar{\mathbf{B}}_{t-1},\mathbf{h}_t,\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\le F\left(\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_{t-1},\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\tag{32}$$

From (26), $\bar{\mathbf{B}}_t$ can be obtained by the approach of OMP, we have

$$F\left(\mathbf{A}_t,\bar{\mathbf{B}}_t,\mathbf{h}_t,\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\le F\left(\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_{t-1},\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1}\right)\tag{33}$$

Since $\mathbf{A}_t,\bar{\mathbf{B}}_t,\mathbf{h}_t$ were obtained, we have the following inequality from (13)

$$tr\left(\mathbf{L}_t^T\mathbf{D}_{t-1}\mathbf{L}_t\right)+\alpha tr\left(\bar{\mathbf{B}}_t^T\mathbf{C}^T\bar{\mathbf{D}}_{t-1}\mathbf{C}\bar{\mathbf{B}}_t\right)\le tr\left(\mathbf{L}_{t-1}^T\mathbf{D}_{t-1}\mathbf{L}_{t-1}\right)+\alpha tr\left(\bar{\mathbf{B}}_{t-1}^T\mathbf{C}^T\bar{\mathbf{D}}_{t-1}\mathbf{C}\bar{\mathbf{B}}_{t-1}\right)\tag{34}$$

where $\mathbf{L}_t=\mathbf{X}\mathbf{C}\bar{\mathbf{B}}_t\mathbf{A}_t^T+\mathbf{1}\mathbf{h}_t^T-\mathbf{Y}$, $\mathbf{L}_{t-1}=\mathbf{X}\mathbf{C}\bar{\mathbf{B}}_{t-1}\mathbf{A}_{t-1}^T+\mathbf{1}\mathbf{h}_{t-1}^T-\mathbf{Y}$.

From the definition of $\mathbf{D}$ and $\bar{\mathbf{D}}$ in (10), (9), we have

$$\sum_i\frac{\left\|\mathbf{L}^i{}_t\right\|_2^2}{2\left\|\mathbf{L}^i{}_{t-1}\right\|_2}+\alpha\sum_i\frac{\left\|(\mathbf{CB})^i{}_t\right\|_2^2}{2\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2}\le\sum_i\frac{\left\|\mathbf{L}^i{}_{t-1}\right\|_2^2}{2\left\|\mathbf{L}^i{}_{t-1}\right\|_2}+\alpha\sum_i\frac{\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2^2}{2\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2}\tag{35}$$

According to Lemma 1, we have

$$\begin{aligned}&\sum_i\left\|\mathbf{L}_t^i\right\|_2-\left(\sum_i\left\|\mathbf{L}_t^i\right\|_2-\sum_i\frac{\left\|\mathbf{L}^i{}_t\right\|_2^2}{2\left\|\mathbf{L}^i{}_{t-1}\right\|_2}\right)+\alpha\sum_i\left\|(\mathbf{CB})_t^i\right\|_2-\alpha\left(\sum_i\left\|(\mathbf{CB})_t^i\right\|_2-\sum_i\frac{\left\|(\mathbf{CB})^i{}_t\right\|_2^2}{2\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2}\right)\\&\le\sum_i\left\|\mathbf{L}_{t-1}^i\right\|_2-\left(\sum_i\left\|\mathbf{L}_{t-1}^i\right\|_2-\sum_i\frac{\left\|\mathbf{L}^i{}_{t-1}\right\|_2^2}{2\left\|\mathbf{L}^i{}_{t-1}\right\|_2}\right)+\alpha\sum_i\left\|(\mathbf{CB})_{t-1}^i\right\|_2-\alpha\left(\sum_i\left\|(\mathbf{CB})_{t-1}^i\right\|_2-\sum_i\frac{\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2^2}{2\left\|(\mathbf{CB})^i{}_{t-1}\right\|_2}\right)\end{aligned}\tag{36}$$

According to Lemma 2, it goes

$$\sum_i\left\|\mathbf{L}_t^i\right\|_2+\alpha\sum_i\left\|(\mathbf{CB})_t^i\right\|_2\le\sum_i\left\|\mathbf{L}_{t-1}^i\right\|_2+\alpha\sum_i\left\|(\mathbf{CB})_{t-1}^i\right\|_2\tag{37}$$

From the definition of $L_{2,1}$-norm shown in (4), we finally have

$$\|\mathbf{L}_t\|_{2,1}+\alpha\left\|\mathbf{C}\bar{\mathbf{B}}_t\right\|_{2,1}\le\|\mathbf{L}_{t-1}\|_{2,1}+\alpha\left\|\mathbf{C}\bar{\mathbf{B}}_{t-1}\right\|_{2,1}\tag{38}$$

Namely,

$$F(\mathbf{A}_t,\bar{\mathbf{B}}_t,\mathbf{h}_t,\mathbf{D}_t,\bar{\mathbf{D}}_t)\le F(\mathbf{A}_{t-1},\bar{\mathbf{B}}_{t-1},\mathbf{h}_{t-1},\mathbf{D}_{t-1},\bar{\mathbf{D}}_{t-1})\tag{39}$$

From (39), we can conclude that the objective function in (8) will monotonically decrease and the proposed Algorithm 2 will finally obtain the optimal solution. □

References

[1] J. Xie, J. Yang, J.J. Qian, Y. Tai, H.M. Zhang, Robust nuclear norm-based matrix regression with applications to robust face recognition, IEEE Trans. Image Process 26 (2017) 2286–2295, doi:10.1109/TIP.2017.2662213.

[2] Y. Xu, Z. Zhong, J. Yang, J. You, D. Zhang, A new discriminative sparse representation method for Robust face Recognition via L2-norm rRegularization, IEEE Trans. Neural Netw. Learn. Syst. 28 (2017) 2233–2242, doi:10.1109/TNNLS.2016.2580572.

[3] K. Zheng, X. Wang, Feature selection method with joint maximal information entropy between features and class, Pattern Recognit. 77 (2018) 20–29, doi:10.1016/j.patcog.2017.12.008.

[4] J. Yang, J. Qian, L. Luo, F. Zhang, Y. Gao, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes, IEEE Trans. Pattern Anal. Mach. Intell. Mach. Intell. (2016) 1 –1, doi:10.1109/TPAMI.2016.2535218.

[5] H. Zhao, W.K. Wong, Regularized discriminant entropy analysis, 47 (2014) 806–819.

[6] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2004) 1–30, doi:10.1198/106186006x113430.

[7] B. Peter N., H.Joao P., K.David J., Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 711–720.

[8] X. He, P. Niyogi, Locality preserving projections, Neural Inf. Process. Syst. 16 (2004) 186–197 10.1.1.19.9400.

[9] D. Cai, X. He, J. Han, Isometric projection, in: Proc. Natl. Conf. Artif. Intell., 2007, pp. 528–533.

[10] Xiaofei He, Deng Cai, Shuicheng Yan, Hong-Jiang Zhang, Neighborhood preserving embedding, in: Tenth IEEE Int. Conf. Comput. Vis. Vol. 1. 2, 2005, pp. 1208–1213, doi:10.1109/ICCV.2005.167.

[11] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326 (80-.).

[12] S. Yan, D. Xu, B. Zhang, H. Zhang, H. Kong, Graph embedding: a general framework for dimensionality reduction, Intern. Conf. Comput. Vis. Pattern Recognit 2 (2005) (2005) 830–837.

[13] N. Nguyen, W. Liu, S. Venkatesh, Ridge regression for two dimensional locality preserving projection, Int. Conf. Pattern Recognit. 2 (2008) 9–12.

[14] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, Int. Conf. Comput. Vis. (2007) 1–7.

[15] D. Brown, Locality-regularized Linear regression for face recognition, Int. Conf. Pattern Recognit (2012) 1586–1589.

[16] J. Wright, S. Member, A.Y. Yang, A. Ganesh, S. Member, S.S. Sastry, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.

[17] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288, doi:10.1111/j.1467-9868.2011.00771.x.

[18] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B. 67 (2005) 301–320.

[19] G.-F. Lu, J. Zou, Y. Wang, Z. Wang, L1-norm-based principal component analysis with adaptive regularization, Pattern Recognit 60 (2016) 901–907, doi:10.1016/j.patcog.2016.07.014.

[20] Z. Qiao, L. Zhou, J.Z. Huang, Sparse linear discriminant analysis with applications to high dimensional low sample size data, IAENG Int. J. Appl. Math. 39 (2009) 48–60.

[21] Z. Zheng, Sparse locality preserving embedding, Int. Congr. Image Signal Process. (2009) 1–5.

[22] Z. Zheng, X. Huang, Z. Chen, X. He, H. Liu, J. Yang, Regression analysis of locality preserving projections via sparse penalty, Inf. Sci. (Ny) 303 (2015) 1–14, doi:10.1016/j.ins.2015.01.004.

[23] S. Yi, Z. Lai, Z. He, Y. ming Cheung, Y. Liu, Joint sparse principal component analysis, Pattern Recognit. 61 (2017) 524–536, doi:10.1016/j.patcog.2016.08.025.

[24] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 2010–2023, doi:10.1109/TPAMI.2015.2505311.

[25] W.K. Wong, Z. Lai, Y. Xu, J. Wen, C.P. Ho, Joint tensor feature analysis for visual object recognition, IEEE Trans. Cybern. 45 (2015) 2425–2436.

[26] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint L2,1-norms minimization, Adv. Neural Inf. Process. Systens 23 (2010) 1813–1821. https://papers.nips.cc/paper/3988-efficient-and-robust-feature-selection-via-joint-l21-norms-minimization.pdf.

[27] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, L21-norm regularized discriminative feature selection for unsupervised learning, Int. Jt. Conf. Artif. Intell. (2011) 1589–1594, doi:10.5591/978-1-57735-516-8/IJCAI11-267.

[28] D. Cai, X. He, J. Han, Spectral regression: a unified approach for sparse subspace learning, IEEE Int. Conf. Data Min. (2007) 73–82, doi:10.1109/ICDM.2007.89.

[29] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A.G. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, Multimedia, IEEE Trans. 15 (2013) 1628–1637, doi:10.1109/TMM.2013.2264928.

[30] S. Xiang, F. Nie, G. Meng, C. Pan, C. Zhang, Discriminative least squares regression for multiclass classification and feature selection, IEEE Trans. Neural Netw. Learn. Syst. 23 (2012) 1738–1754.

[31] H. Yan, J. Yang, Sparse discriminative feature selection, Pattern Recognit. 48 (2015) 1827–1835, doi:10.1016/j.patcog.2014.10.021.

[32] D. Mo, Z. Lai, Robust jointly sparse regression for image feature selection, 4th Asian Conf. Pattern Recognition, ACPR 2017, 2017.

[33] Y. Cui, L. Fan, A novel supervised dimensionality reduction algorithm: graph-based Fisher analysis, Pattern Recognit. 45 (2012) 1471–1481, doi:10.1016/j.patcog.2011.10.006.

[34] X. He, S. Yan, Y. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 328–340, doi:10.1109/TPAMI.2005.55.

[35] J. Zhang, J. Yu, J. Wan, Z. Zeng, L21 norm regularized fisher criterion for optimal feature selection, Neurocomputing 166 (2015) 455–463, doi:10.1016/j.neucom.2015.03.033.

[36] C. Ding, D. Zhou, X. He, H. Zha, R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization, in: Proc. 23rd Int. Conf. Mach. Learn. - ICML '06., 2006, pp. 281–288, doi:10.1145/1143844.1143880.

[37] G. Obozinski, B. Taskar, M. Jordan, in: Multi-Task Feature Selection, Tech. Report, Dep. Stat. Univ., California, Berkeley, 2006, pp. 1–15. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.951&rep=rep1&type=pdf.

[38] H. Huang, C. Ding, Robust tensor factorization using R1 norm, 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, 2008, doi:10.1109/CVPR.2008.4587392.

[39] M.A. Davenport, M.B. Wakin, Analysis of orthogonal matching pursuit using the restricted Isometry property, IEEE Trans. Inf. Theory 56 (2010) 4395–4401.

[40] J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (2007) 4655–4666.

[41] R. Rezaiifar, Orthogonal matching pursuit : recursive function approximat ion with applications to wavelet decomposition, in: 27th Asilomar Conf. Signals, Syst. Comput., 1993, pp. 40–44.

[42] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306.

[43] E.J. Candès, Compressive sampling, Marta Sanz Solé 17 (2007) 1433–1452.

[44] J.A. Tropp, S. Member, Greed is good : algorithmic results for sparse approximation, IEEE Trans. Inf. Theory. 50 (2004) 2231–2242.

[45] F. Nie, H. Huang, X. Cai, Chris Ding, Efficient and robust feature selection via joint L2,1 norms minimization, Adv. Neural Inf. Process. Systens 23 (2010) 1813–1821.

[46] V. Sindhwani, P. Niyogi, Linear manifold regularization for large scale semi-supervised learning, Proc Icml Work. Learn. with Partial. Classif. Train. Data (2005) 80–83. http://webia.lip6.fr/∼amini/ICML05/page80-83.pdf.

[47] F. Nie, D. Xu, I.W.H. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, IEEE Trans. Image Process. 19 (2010) 1921–1932, doi:10.1109/TIP.2010.2044958.

[48] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326 (80-.), doi:10.1126/science.290.5500.2323.

[49] E. Kokiopoulou, Orthogonal neighborhood preserving projections : a projection-based dimensionality reduction technique, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 2143–2156.

[50] D. Cai, X. He, J. Han, S. Member, Orthogonal laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (2006) 3608–3614.

[51] X. Liu, J. Yin, Z. Feng, J. Dong, L. Wang, Orthogonal neighborhood preserving embedding for face recognition, IEEE Int. Conf. Image Process. (2007) 133–136.

[52] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (2010) 331–341, doi:10.1016/j.patcog.2009.05.005.

[53] X. Shi, Y. Yang, Z. Guo, Z. Lai, Face recognition by sparse discriminant analysis via joint L2,1-norm minimization, Pattern Recognit. 47 (2014) 2447–2453, doi:10.1016/j.patcog.2014.01.007.

[54] Yale Face DB: Yale Univ.2007. [Online]. Available: cvc.yale.edu/projects/yalefaces/yalefaces.html, (n.d.).

[55] F.S. Samaria, U.K. Cb, Parameterisation of a stochastic model for human face identification, IEEE Work. Appl. Comput. Vis. (1994) 138–142.

[56] A.A. Martinez, R. Benavente, The AR face database, CVC Tech, 1998 Reptort #24.

[57] H. Yuan, Y.Y. Tang, Y. Lu, S. Member, Hyperspectral image classification based on regularized sparse representation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (2014) 2174–2182.

[58] R. Wang, F. Nie, R. Hong, Fast and orthogonal locality preserving projections for dimensionality reduction, IEEE Trans. Image Process. 26 (2017) 5019–5030.

[59] C. Hou, F. Nie, X. Li, D. Yi, Y. Wu, Joint embedding learning and sparse regression: a framework for unsupervised feature selection, IEEE Trans. Cybern. 44 (2014) 793–804.

[60] Z. Lai, W.K. Wong, Y. Xu, J. Yang, D. Zhang, Approximate Orthogonal Sparse Embedding for Dimensionality Reduction, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 723.

[61] Y. Pang, Y. Yuan, Outlier-resisting graph embedding, Neurocomputing 73 (2010) 968–974, doi:10.1016/j.neucom.2009.08.020.

Dongmei Mo received the B.S degree and M.S Degree from Zhaoqing University and Shenzhen University. She is now pursuing PHD degree in the Hong Kong Polytechnic University (e-mail: dongmei_mo@qq.com).

Zhihui Lai received the B.S degree in mathematics from South China Normal University, M.S degree from Jinan University, and the PhD degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a research associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He has published over 100 scientific articles, including more than 30 papers published on top-tier IEEE Transactions. Now he is an associate editor of International Journal of Machine Learning and Cybernetics. For more information, including all the papers and the Matlab codes, please refer to his website: http://www.scholat.com/laizhihui.