Multimed Tools Appl (2018) 77:16155–16175
DOI 10.1007/s11042-017-5193-9

Sparse L1-norm-based linear discriminant analysis

Gui-Fu Lu[1,2] · **Jian Zou**[1] · **Yong Wang**[1] · **Zhongqun Wang**[1]

Received: 24 November 2016 / Revised: 20 July 2017 / Accepted: 4 September 2017 /
Published online: 13 September 2017

Abstract Linear discriminant analysis (LDA) is a well-known feature extraction method, which has been widely used for many pattern recognition problems. However, the objective function of conventional LDA is based on L2-norm, which makes LDA sensitive to outliers. Besides, the basis vectors learned by conventional LDA are dense and it is often hard to explain the extracted features. In this paper, we propose a novel sparse L1-norm-based linear discriminant analysis (SLDA-L1) which not only replaces L2-norm in conventional LDA with L1-norm, but also use the elastic net to regularize the basis vectors. Then L1-norm used in SLDA-L1 is for both robust and sparse modelling simultaneously. We also propose an efficient iterative algorithm to solve SLDA-L1 which is theoretically shown to arrive at a locally maximal point. Experiment results on some image databases demonstrate the effectiveness of the proposed method.

Keywords Feature extraction · Dimensionality reduction · L1-norm · L2-norm · elastic net regularization

1 Introduction

Feature extraction plays a critical role in a lot of pattern recognition and machine learning problems [8, 10–13, 32, 42]. A main goal of feature extraction is to find some low-dimensional expressive or discriminative features for high-dimensional data. Many feature extraction techniques have been proposed in the literatures [14]. Among them, principal component analysis (PCA) [8, 9] and linear discriminant analysis (LDA) [4, 8, 9] may be two of the most well-known linear ones.

✉ Gui-Fu Lu
luguifu_tougao@163.com

[1] School of Computer and Information, AnHui Polytechnic University, WuHu, AnHui 241000, China

[2] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

PCA is an unsupervised feature extraction method. It aims to seek a set of orthogonal projection vectors to maximize the variance of the given data. To improve the performance of PCA, some supervised PCA methods have been proposed. Bair et al. [2] and Barshan et al. [3], respectively, proposed the supervised principal component (SPCA) method which can be applied to regression problems where the number of features greatly exceeds the number of samples. Xia et al. [33]proposed a supervised probabilistic principal component analysis (SPPCA) which has successfully been used for feature extraction in hyperspectral remote sensing imagery. By generating face relevance maps, Kawulok et al. [16] improves the discriminating capability of Eigenfaces, which is based on PCA. Different from conventional PCA, LDA is a supervised feature extraction method. It aims to seek a projection transformation matrix on which the given data of the same class are as near as possible to each other while the given data of the different classes are as far as possible from each other. Generally, LDA is superior to PCA for classification problems. Recently, researchers proposed some other feature extraction methods which need not to be trained in advance. In [21], Leng et al. proposed a method called two-directional two-dimensional random projection ($(2D)^2$RP) for feature extraction of biometrics. $(2D)^2$RP can avoid the problem of small sample size (SSS) and overfitting. Dynamic weighted DPA (DWDPA) is proposed in [19, 20] to improve the discrimination power of the selected discrete cosine transform coefficients without premasking window.

Although PCA and LDA have been widely used in many fields, they are sensitive to outliers since their objective functions are based on L2-norm in which the square operation exaggerates the effect of outliers [18]. Generally, L1-norm is more robust to outliers than L2-norm. Therefore, some L1-norm-based feature extraction techniques, e.g. robust PCA [7, 17, 18, 26, 27, 34] and robust LDA [24, 30, 31, 38, 39], are developed to address the drawback of L2-norm-based feature extraction methods.

By using maximum likelihood estimation to the given data, Ke et al. [17] proposed L1-PCA. The objective function of L1-PCA is solved via convex programming techniques. However, L1-PCA is not rotational invariant. R1-PCA [7], in which the rotational invariant L1-norm is used, is rotational invariant and can combine the advantages of PCA and L1-PCA. However, due to the absolute operator, the optimizations of L1-PCA and R1-PCA are more difficult than the conventional PCA.

Recently, to address the drawbacks of L1-PCA and R1-PCA, Kwak [18] proposed the PCA-L1 method, which is rotational invariant and can be solved by a greedy iteration algorithm. PCA-L1 is computational efficiently. Experiments on some face databases demonstrate that PCA-L1 can obtain much lower reconstruction error than L1-PCA and R1-PCA. To obtain much better projection matrix of PCA-L1, Nie et al. [26] proposed a non-greedy strategy to solve PCA-L1 which can obtain all the projection vectors simultaneously. Motivated by PCA-L1 and 2DPCA [36], Li et al. [34] proposed the L1-norm-based 2DPCA method (2DPCA-L1). Further, Pang et al. [27] proposed L1-norm-based tensor PCA (TPCA-L1).

Similar to conventional PCA, L1-norm-based PCA is also unsupervised and cannot find the optimal discriminative projection vectors. Then for object recognition problems, LDA, which is a supervised feature extraction method, may be a better choice. Motivated by R1-PCA and maximum margin criterion (WMMC) [22, 23], Li et al. [24] proposed a new rotational invariant L1-norm based MMC (R1-MMC). The iterative algorithm for solving R1-MMC is based on eigenvalue decomposition, and then R1-MMC is computational expensive. Recently, Wang et al. [30] proposed a robust version of common spatial patterns (CSP), i.e., L1-norm-

based CSP (CSP-L1), which replace L2-norm in CSP with L1-norm. An efficient iterative algorithm for solving CSP-L1 is also proposed. Similar ideas are also appeared in [31, 38, 39] where L2-norm is replaced with L1-norm in the LDA objection function and LDA-L1 is proposed.

However, the projection vectors learned by the above L1-norm-based PCA and LDA are still dense, which makes it hard to explain the obtained features. To address this problem, sparse feature extraction methods have been developed in recent years. Zou et al. [40] proposed sparse PCA (SPCA), where PCA is firstly reformulated as a regression problem and then the elastic net is used to regularize the bases obtained by PCA. Jenatton et al. [15] further proposed the structured sparse PCA algorithm, which is a generalization of SPCA. Motivated by the graph embedding framework [35], Cai et al. [5, 6] proposed a unified sparse subspace learning (USSL) framework with spectral regression. Similarly, Tao et al. [41] proposed so-called manifold elastic net, which is also a unified sparse dimension reduction framework with patch alignment technique. Wang [28] proposed structured sparse linear graph embedding (SSLGE), which use the structured sparsity-inducing norm to regularize the bases obtained by linear graph embedding. However, the objective function of the above sparse learning methods is still based on L2-norm with certain sparsity constraint, which makes these methods prone to the influence of noises. In [25], Meng et al. extended PCA-L1 with sparsity (PCA-L1S). In PCA-L1S, the objective function is based on L1-norm. Besides, the project vectors are also regularized with L1-norm. Similarly, Wang et al. [29] extended 2DPCA-L1 with sparsity (2DPCA-L1S).

In order to improve the robustness of LDA-L1 and interpretability of the vectors obtained by LDA-L1 further, in this paper, we propose a novel sparse L1-norm-based linear discriminant analysis (SLDA-L1) which not only replace L2-norm in conventional LDA with L1-norm, but also use the elastic net to regularize the basis vectors. Then L1-norm used in SLDA-L1 is for both robust and sparse modeling simultaneously. An efficient iterative algorithm to solve SLDA-L1, which is theoretically shown to arrive at a locally maximal point, is also proposed.

To show the robustness of SLDA-L1, here, we will presents an experiment on artificial dataset. Two Gaussian classes with covariance matrices being [0.05 0; 0 2] and means being [−2 0] and [2 0] respectively. Each class consists of 20 2D samples as depicted in Fig. 1a and Fig. 1b, respectively, where two classes are respectively specified by “o” and “x”. Firstly, we extract the projection vectors using LDA-L2, R1-MMC, LDA-L1 and SLDA-L1 on the above data without any outlier. The obtained projection vectors are shown in Fig. 1a. We can find that the obtained projection vectors are very similar. In order to compare the robustness of different methods, an additional outlier, i.e. [4] specified by red “o” is introduced in Fig. 1b. The projection vectors extracted by different methods are also shown in Fig. 1b. From Fig. 1b, we can know the projection vectors obtained by SLDA-L1 is better than the other three methods.

It is worth highlighting the novelty of our proposed SLDA-L1 method.

1) We propose a novel sparse L1-norm-based linear discriminant analysis (SLDA-L1), in which L1-norm is for both robust and sparse modeling simultaneously.
2) We also propose an efficient iterative algorithm to solve SLDA-L1, which is theoretically shown to arrive at a locally maximal point.

The remainder of the paper is organized as follows. In section 2, we review briefly the conventional LDA technique. In Section 3, we propose the SLDA-L1 method, including its

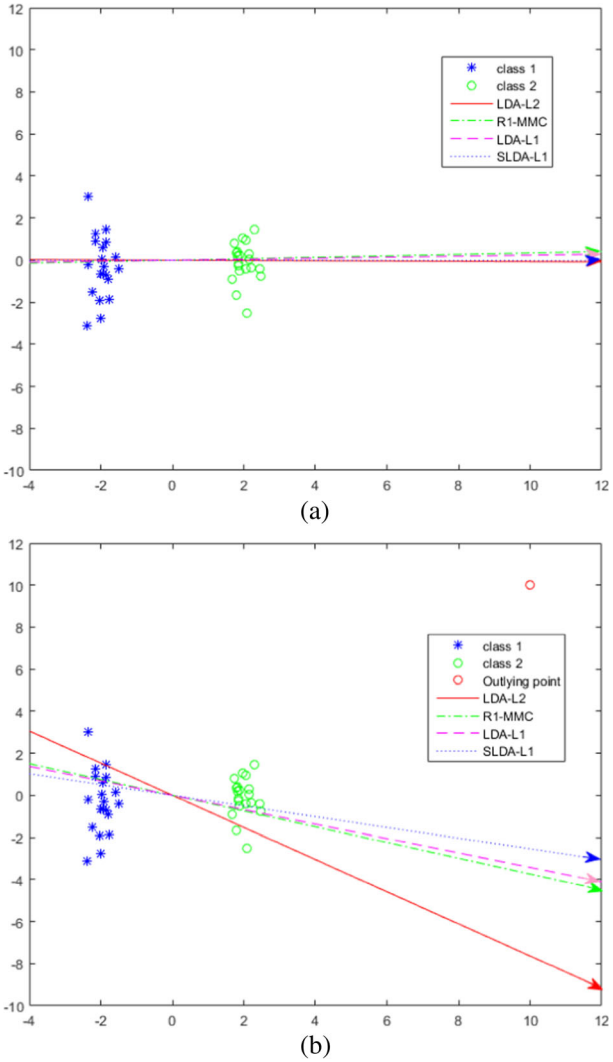


Fig. 1 Projection vectors learned by different methods on an artificial dataset

objective function and algorithmic procedure. Section 4 is devoted to the experiments. Finally, we conclude the paper in Section 5.

2 Related work

Let $X = \left\{\mathbf{x}_j^i, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k\right\}$ be a d-dimensional real sample set with n elements, where k is the number of the classes, n_i is the number of the samples of ith class, $n = \sum_{i=1}^{k} n_i$ is the number of the data set, and $\mathbf{x}_j^i$ is the jth samples of the ith class. In LDA

(termed as LDA-L2), between-class scatter matrix and within-class scatter matrix, are respectively defined as follows:

$$S_b = \sum_{i=1}^{k} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{1}$$

$$S_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{x}_j^i - \mathbf{m}_i\right)\left(\mathbf{x}_j^i - \mathbf{m}_i\right)^T \tag{2}$$

where $\mathbf{m}_i = (1/n_i)\sum_{j=1}^{n_i} \mathbf{x}_j^i$ is the mean of the ith class and $\mathbf{m} = (1/n)\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathbf{x}_j^i$ is the global mean of the data set.

Define the matrices

$$H_b = [\sqrt{n_1}(\mathbf{m}_1 - \mathbf{m}),\ \sqrt{n_2}(\mathbf{m}_2 - \mathbf{m}),\ \cdots,\ \sqrt{n_k}(\mathbf{m}_k - \mathbf{m})] \tag{3}$$

$$H_w = \left[\ \mathbf{x}_1^1 - \mathbf{m}_1,\ \cdots\ \mathbf{x}_{n_1}^1 - \mathbf{m}_1,\ \cdots,\ \mathbf{x}_1^k - \mathbf{m}_k,\ \cdots,\ \mathbf{x}_{n_k}^k - \mathbf{m}_k\ \right] \tag{4}$$

The optimal projection vector $\mathbf{w} \in R^d$ of LDA can be obtained by maximizing the following so-called Fisher criterion:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \tag{5}$$

The projection vector $\mathbf{w}$ is the leading generalized eigenvector of $S_b\mathbf{w} = \lambda S_w\mathbf{w}$.

3 Sparse L1-norm-based linear discriminant analysis (SLDA-L1)

3.1 Problem formulation

In this subsection, we will present our proposed sparse L1-norm-based linear discriminant analysis. By simply transforming, Eq. (5) can be reformulated as

$$J(\mathbf{w}) = \frac{\|\mathbf{w}^T H_b\|_2^2}{\|\mathbf{w}^T H_w\|_2^2} \tag{6}$$

where $\|\cdot\|_2$ denotes L2-norm. Obviously, the objective function of LDA-L2 is based on L2-norm. It is generally believed that L1-norm based feature extraction methods are more robust to outliers than L2-norm based feature extraction methods [31, 39]. Besides, in the sparsity-inducing modeling, L1-norm is often used to regularize the bases obtained by the feature extraction methods. Then, the objective function of SLDA-L1 is defined as follows:

$$F(\mathbf{w}) = \frac{\|\mathbf{w}^T H_b\|_1}{\|\mathbf{w}^T H_w\|_1} - \lambda\|\mathbf{w}\|_1 - \frac{\eta}{2}\|\mathbf{w}\|_2^2 \tag{7}$$

where $\lambda > 0$ and $\eta > 0$ are tuning parameters. The first term in Eq. (7) is the robust design of the objective function of LDA-L2. The second and the third terms in Eq. (7) are the

elastic net, which can circumvent potential limitations of lasso [40]. That is, it can not only result in sparsity, but also can improve the grouping effectiveness in regression. The optimal $\mathbf{w}$ can be obtained by maximizing Eq. (7). It, however, is difficult to obtain the global optimal solution of variable of $\mathbf{w}$ since the absolute value operation in Eq. (7) is not differentiable. In the following subsection, we will propose an iterative algorithm to find a local optimal $\mathbf{w}$.

3.2 Algorithm of finding the projection vector w

We firstly introduce the following notations.

Let $\mathbf{w}(t)$ be the basis vector in the t-th iteration. Remove the zero entries of $\mathbf{w}(t)$ and denote the new vector as $\underline{\mathbf{w}}(t)$. Remove the entries of A whose indices is the indices of the zero entries of $\mathbf{w}(t)$ and the new vector is denoted as $\underline{A}$. For example, let $\mathbf{w}(t)=(2,0,0,3,0,5)^T$and $\mathbf{x}_j^i=(1,3,2,6,8,4)^T$. Then $\underline{\mathbf{w}}(t)=(2,3,5)^T$ and $\underline{\mathbf{x}}_j^i=(1,6,4)^T$. Let$\mathbf{w}(t+1)$ be the vector which is formed by inserting zero entries into $\underline{\mathbf{w}}(t+1)$ and the indices of the inserted zero entries of $\mathbf{w}(t+1)$ are just the indices of the zero entries of $\underline{\mathbf{w}}(t)$. For example, if $\underline{\mathbf{w}}(t+1)=(3,5,4)$, then $\mathbf{w}(t+1)=(3,0,0,5,0,4)$.

Now the iterative algorithm of SLDA-L1 is stated as follows.

Algorithm1: SLDA-L1

Input: data matrix X

Output: projection vector $\mathbf{w}$

Step 1: Set t=0. Select any unit d-dimensional vector as $\mathbf{w}(t)$;

Step 2: Define two polarity functions

$$p_i(t)=\begin{cases}1, & if\ \mathbf{w}^T(t)(\mathbf{m}_i-\mathbf{m})\ge 0\\ -1, & if\ \mathbf{w}^T(t)(\mathbf{m}_i-\mathbf{m})<0\end{cases},\quad i=1,\ldots,k \tag{8}$$

and

$$q_j^i(t)=\begin{cases}1, & if\ \mathbf{w}^T(t)(\mathbf{x}_j^i-\mathbf{m}_i)\ge 0\\ -1, & if\ \mathbf{w}^T(t)(\mathbf{x}_j^i-\mathbf{m}_i)<0\end{cases},\quad i=1,\ldots,k,\ j=1,\ldots,n_i \tag{9}$$

Step 3: Let

$$\underline{\mathbf{d}}(t)=\frac{\sum_{i=1}^{k}\sqrt{n_i}\,p_i(t)(\underline{\mathbf{m}}_i-\underline{\mathbf{m}})}{\left\|\underline{\mathbf{w}}(t)^T\underline{H}_w\right\|_1}-\frac{\left(\sum_{i=1}^{k}\sqrt{n_i}\left|\underline{\mathbf{w}}(t)^T(\underline{\mathbf{m}}_i-\underline{\mathbf{m}})\right|\right)\sum_{i=1}^{k}\sum_{j=1}^{n_i}q_j^i(t)(\underline{\mathbf{x}}_j^i-\underline{\mathbf{m}}_i)}{\left(\left\|\underline{\mathbf{w}}(t)^T\underline{H}_w\right\|_1\right)^2}$$
$$-\lambda\left(sign(\underline{w}_1(t)),\ldots,sign(\underline{w}_p(t))\right)-\eta\underline{\mathbf{w}}(t) \tag{10}$$

where $\underline{w}_i(t)$, $i=1,\ldots,p$ is the ith element of $\underline{\mathbf{w}}(t)$ and p is the number of the nonzero elements of $\underline{\mathbf{w}}(t)$. Then update $\underline{\mathbf{w}}(t+1)$ as

$$\underline{\mathbf{w}}(t+1)=\underline{\mathbf{w}}(t)+\beta\underline{\mathbf{d}}(t) \tag{11}$$

where $\beta>0$ denotes the learning rate.

Step 4: If $F(\mathbf{w}+1)$ does not increase significantly, then stop the iteration and output $\mathbf{w}(t+1)$ as the local optimal vector. Otherwise, Let $t\leftarrow t+1$ and go to Step 2.

The convergence of the above iterative algorithm is theoretically justified by the following Theorem 1.

Theorem 1 With the above iterative procedure of SLDA-L1, the objective function of $F(\mathbf{w})$ is nondecreasing with each iteration.

Proof At each iteration t, we have

$$F(\mathbf{w}(t)) = \frac{\left\|\mathbf{w}(t)^T H_b\right\|_1}{\left\|\mathbf{w}(t)^T H_w\right\|_1} - \lambda\|\mathbf{w}(t)\|_1 - \frac{\eta}{2}\|\mathbf{w}(t)\|_2^2 \tag{12}$$

Obviously Eq. (12) can be rewritten as

$$F(\mathbf{w}(t)) = \frac{\left\|\underline{\mathbf{w}}(t)^T \underline{H}_b\right\|_1}{\left\|\underline{\mathbf{w}}(t)^T \underline{H}_w\right\|_1} - \lambda\left\|\underline{\mathbf{w}}(t)\right\|_1 - \frac{\eta}{2}\left\|\underline{\mathbf{w}}(t)\right\|_2^2 \tag{13}$$

The denominator of the first term in Eq. (13) can be rewritten as

$$\left\|\underline{\mathbf{w}}(t)^T \underline{H}_w\right\|_1 = \frac{1}{2}\underline{\mathbf{w}}(t)^T \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\left(\overset{i}{\underline{\mathbf{x}}}_j - \underline{\mathbf{m}}_i\right)\left(\overset{i}{\underline{\mathbf{x}}}_j - \underline{\mathbf{m}}_i\right)^T}{\left|\underline{\mathbf{w}}(t)^T \left(\overset{i}{\underline{\mathbf{x}}}_j - \underline{\mathbf{m}}_i\right)\right|} \right) \underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{w}}(t)^T \underline{H}_w\right\|_1$$

$$= \frac{1}{2}\underline{\mathbf{w}}(t)^T \underline{V}(t)\underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1 \tag{14}$$

where

$$\underline{V}(t) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\left(\overset{i}{\underline{\mathbf{x}}}_j - \underline{\mathbf{m}}_i\right)\left(\overset{i}{\underline{\mathbf{x}}}_j - \underline{\mathbf{m}}_i\right)^T}{\left|\overset{i}{\underline{z}}_j(t)\right|} \tag{15}$$

$\underline{z}_j^i(t) = \underline{\mathbf{w}}(t)^T\left(\underline{\mathbf{x}}_j^i - \underline{\mathbf{m}}_i\right)$ and $\underline{\mathbf{z}}(t)$ is the vector having the elements $\left\{\underline{z}_j^i(t)\right\}_{j=1,\ldots,n_i;i=1,\ldots,k}$.

The numerator of the first term in Eq. (13) can be rewritten as

$$\left\|\mathbf{w}(t)^T H_b\right\|_1 = \left\|\underline{\mathbf{w}}(t)^T \underline{H}_b\right\|_1 = \underline{\mathbf{w}}(t)^T \sum_{i=1}^{k} \sqrt{n_i} p_i(t)\left(\underline{\mathbf{m}}_i - \underline{\mathbf{m}}\right) \qquad = \underline{\mathbf{w}}(t)^T \underline{\mathbf{u}}(t) \tag{16}$$

where

$$\underline{\mathbf{u}}(t) = \sum_{i=1}^{k} \sqrt{n_i} p_i(t)\left(\underline{\mathbf{m}}_i - \underline{\mathbf{m}}\right) \tag{17}$$

The second term of Eq. (13) can be rewritten as

$$\|\mathbf{w}(t)\|_1 = \frac{1}{2}\underline{\mathbf{w}}(t)^T \underline{U}(t)\underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{w}}(t)\right\|_1 \tag{18}$$

where $\underline{U}(t) = diag\left(\left|\underline{w}_1(t)\right|^{-1}, \ldots, \left|\underline{w}_p(t)\right|^{-1}\right)$ is a diagonal matrix.

Combining Eq. (14), Eq. (16), Eq. (18) and Eq. (13), we have

$$F(\mathbf{w}(t)) = \frac{\underline{\mathbf{w}}(t)^T \underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t)^T \underline{V}(t)\underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1} - \frac{\lambda}{2}\left(\underline{\mathbf{w}}(t)^T \underline{U}(t)\underline{\mathbf{w}}(t) + \left\|\underline{\mathbf{w}}(t)\right\|_1\right) - \frac{\eta}{2}\left\|\underline{\mathbf{w}}(t)\right\|_2^2 \quad (19)$$

Since it is intractable to derive the function $F(\mathbf{w}(t))$ directly, we introduce a surrogate function as follows:

$$L(\xi) = \frac{\xi^T \underline{\mathbf{u}}(t)}{\frac{1}{2}\xi^T \underline{V}(t)\xi + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1} - \frac{\lambda}{2}\left(\xi^T \underline{U}(t)\xi + \left\|\underline{\mathbf{w}}(t)\right\|_1\right) - \frac{\eta}{2}\|\xi\|_2^2 \quad (20)$$

Note that only ξ is the variable while $\underline{\mathbf{u}}(t)$, $\underline{V}(t)$, $\underline{\mathbf{z}}(t)$, $\underline{U}(t)$ and $\underline{\mathbf{w}}(t)$ are all fixed values in the function $L(\xi)$. Compute the gradient of Eq. (20) as follows:

$$\begin{aligned} g(\xi) &= \frac{\partial L(\xi)}{\partial \xi} \\ &= \frac{\left(\frac{1}{2}\xi^T \underline{V}(t)\xi + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1\right)\underline{\mathbf{u}}(t) - \left(\xi^T \underline{\mathbf{u}}(t)\right)\underline{V}(t)\xi}{\left(\frac{1}{2}\xi^T \underline{V}(t)\xi + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1\right)^2} - \lambda\left(\underline{U}(t)\xi\right) - \eta\xi \\ &= \frac{\underline{\mathbf{u}}(t)}{\frac{1}{2}\xi^T \underline{V}(t)\xi + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1} - \frac{\left(\xi^T \underline{\mathbf{u}}(t)\right)\underline{V}(t)\xi}{\left(\frac{1}{2}\xi^T \underline{V}(t)\xi + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1\right)^2} - \lambda\left(\underline{U}(t)\xi\right) - \eta\xi \end{aligned} \quad (21)$$

From Eq. (21), we can obtain the gradient at point $\underline{\mathbf{w}}(t)$:

$$\begin{aligned} g\left(\underline{\mathbf{w}}(t)\right) &= \frac{\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t)^T \underline{V}(t)\underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1} - \frac{\left(\underline{\mathbf{w}}(t)^T \underline{\mathbf{u}}(t)\right)\underline{V}(t)\underline{\mathbf{w}}(t)}{\left(\frac{1}{2}\underline{\mathbf{w}}(t)^T \underline{V}(t)\underline{\mathbf{w}}(t) + \frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1\right)^2} \\ &- \lambda\left(\underline{U}(t)\underline{\mathbf{w}}(t)\right) - \eta\underline{\mathbf{w}}(t) \end{aligned} \quad (22)$$

Substituting $\underline{\mathbf{u}}(t)$, $\underline{V}(t)$, $\underline{z}_j^i(t)$ and $\underline{U}(t)$ into Eq. (22), we have

$$\begin{aligned} & g\left(\underline{\mathbf{w}}(t)\right) \\ &= \frac{\sum_{i=1}^{k} \sqrt{n_i} p_i(t)\left(\underline{\mathbf{m}}_i - \underline{\mathbf{m}}\right)}{\left\|\underline{\mathbf{w}}(t)^T \underline{H}_w\right\|_1} - \frac{\left(\sum_{i=1}^{k} \sqrt{n_i}\left|\underline{\mathbf{w}}(t)^T\left(\underline{\mathbf{m}}_i - \underline{\mathbf{m}}\right)\right|\right)\sum_{i=1}^{k}\sum_{j=1}^{n_i} g_j^i(t)\left(\underline{\mathbf{x}}_j^i - \underline{\mathbf{m}}_i\right)}{\left(\left\|\underline{\mathbf{w}}(t)^T \underline{H}_w\right\|_1\right)^2} \\ &- \lambda\left(sign\left(\underline{w}_1(t)\right), \ldots, sign\left(\underline{w}_p(t)\right)\right) - \eta\underline{\mathbf{w}}(t) \\ &= \underline{\mathbf{d}}(t) \end{aligned} \quad (23)$$

Obviously, $\underline{\mathbf{d}}(t)$is the vector which points to the ascending direction of $g(\xi)$ at point $\underline{\mathbf{w}}(t)$. Let

$$\underline{\mathbf{w}}(t+1) = \underline{\mathbf{w}}(t) + \beta\underline{\mathbf{d}}(t) \quad (24)$$

Then according to Eq. (21), we have

$$L\left(\underline{\mathbf{w}}(t+1)\right)\geq L\left(\underline{\mathbf{w}}(t)\right) \tag{25}$$

That is

$$\begin{aligned}&\frac{\underline{\mathbf{w}}(t+1)^T\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t+1)^T\underline{V}(t)\underline{\mathbf{w}}(t+1)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1}-\frac{\lambda}{2}\left(\underline{\mathbf{w}}(t+1)^T\underline{U}(t)\underline{\mathbf{w}}(t+1)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\\ &\geq\frac{\underline{\mathbf{w}}(t)^T\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t)^T\underline{V}(t)\underline{\mathbf{w}}(t)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1}-\frac{\lambda}{2}\left(\underline{\mathbf{w}}(t)^T\underline{U}(t)\underline{\mathbf{w}}(t)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t)\right\|_2^2=F(\mathbf{w}(t))\end{aligned} \tag{26}$$

In the following we will prove the following inequality

$$\begin{aligned}&\frac{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\right\|_1}{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_w\right\|_1}-\lambda\left\|\underline{\mathbf{w}}(t+1)\right\|_1-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\\ &\geq\frac{\underline{\mathbf{w}}(t+1)^T\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t+1)^T\underline{V}(t)\underline{\mathbf{w}}(t+1)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1}-\frac{\lambda}{2}\left(\underline{\mathbf{w}}(t+1)^T\underline{U}(t)\underline{\mathbf{w}}(t+1)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\end{aligned} \tag{27}$$

Before proving Eq. (27), we firstly introduce the following lemma.

Lemma [15] For any vector $\mathbf{a}=(a_1,\ \ldots,a_n)^T\in R^n$, the following equality hold:

$$\|\mathbf{a}\|_1=\min_{\varsigma\in R_+^n}\frac{1}{2}\sum_{j=1}^{n}\frac{a_j^2}{\varsigma_j}+\frac{1}{2}\|\varsigma\|_1 \tag{28}$$

and the minimum is uniquely reached at $\varsigma=|a_j|, j=1,\ \ldots,n$ where $\varsigma=(\varsigma_1,\ \ldots,\varsigma_n)^T$.

Considering the denominator of the first term of the right hand of Eq. (27), we have

$$\begin{aligned}&\frac{1}{2}\underline{\mathbf{w}}(t+1)^T\underline{V}(t)\underline{\mathbf{w}}(t+1)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1=\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\left(\underline{\mathbf{w}}(t+1)^T\left(\overset{i}{\underline{\mathbf{x}}}_j-\underline{\mathbf{m}}_i\right)\right)^2}{\left|\overset{i}{z}_j(t)\right|}\\ &+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1\geq\min_{\varsigma\in R_+^n}\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\left(\underline{\mathbf{w}}(t+1)^T\left(\overset{i}{\underline{\mathbf{x}}}_j-\underline{\mathbf{m}}_i\right)\right)^2}{\varsigma_j}+\frac{1}{2}\|\varsigma\|_1=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left|\underline{\mathbf{w}}(t+1)^T\left(\overset{i}{\underline{\mathbf{x}}}_j-\underline{\mathbf{m}}_i\right)\right|\\ &=\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_w\right\|_1\geq 0\end{aligned} \tag{29}$$

Considering the numerator of the first term of the right hand of Eq. (27), we have

$$\begin{aligned}&\underline{\mathbf{w}}(t+1)^T\underline{\mathbf{u}}(t)=\underline{\mathbf{w}}(t+1)^T\sum_{i=1}^{k}\sqrt{n_i}p_i(t)\left(\underline{\mathbf{m}}_i-\underline{\mathbf{m}}\right)\leq\underline{\mathbf{w}}(t+1)^T\sum_{i=1}^{k}\sqrt{n_i}p_i(t+1)\left(\underline{\mathbf{m}}_i-\underline{\mathbf{m}}\right)\\ &=\sum_{i=1}^{k}\sqrt{n_i}\left|\underline{\mathbf{w}}(t+1)^T\left(\underline{\mathbf{m}}_i-\underline{\mathbf{m}}\right)\right|=\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\right\|_1\end{aligned} \tag{30}$$

In Eq. (30), the equality is due to the fact that $p_i(t+1)\underline{\mathbf{w}}(t+1)^T(\underline{\mathbf{m}}_i-\underline{\mathbf{m}})$, $i=1,2,\ldots,k$, is always nonnegative since $p_i(t+1)$ is the polarity of $\underline{\mathbf{w}}(t+1)^T(\underline{\mathbf{m}}_i-\underline{\mathbf{m}})$ while $p_i(t)\underline{\mathbf{w}}(t+1)^T(\underline{\mathbf{m}}_i-\underline{\mathbf{m}})$, $i=1,2,\ldots,k$ may be negative for some i.

Since $\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\|_1\geq 0$, then from Eq. (29) and Eq. (30) we have

$$\frac{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\right\|_1}{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_w\right\|_1}\geq\frac{\underline{\mathbf{w}}(t+1)^T\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t+1)^T\underline{V}(t)\underline{\mathbf{w}}(t+1)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1} \tag{31}$$

Considering the second term of the right hand of Eq. (27), we have

$$\begin{aligned}&\frac{1}{2}\left(\underline{\mathbf{w}}(t+1)^T\underline{U}(t)\underline{\mathbf{w}}(t+1)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)=\frac{1}{2}\left(\sum_{i=1}^{p}\frac{\underline{w}_i(t+1)^2}{\left|\underline{w}_i(t)\right|}+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)\\&\geq\frac{1}{2}\left(\sum_{i=1}^{p}\frac{\underline{w}_i(t+1)^2}{\left|\underline{w}_i(t+1)\right|}+\left\|\underline{\mathbf{w}}(t+1)\right\|_1\right)=\left\|\underline{\mathbf{w}}(t+1)\right\|_1\end{aligned} \tag{32}$$

Then from Eq. (32) we have

$$-\lambda\left\|\underline{\mathbf{w}}(t+1)\right\|_1\geq-\frac{\lambda}{2}\left(\underline{\mathbf{w}}(t+1)^T\underline{U}(t)\underline{\mathbf{w}}(t+1)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right) \tag{33}$$

Combining Eq. (31) and Eq. (33), we have

$$\begin{aligned}&\frac{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\right\|_1}{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_w\right\|_1}-\lambda\left\|\underline{\mathbf{w}}(t+1)\right\|_1-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\\&\geq\frac{\underline{\mathbf{w}}(t+1)^T\underline{\mathbf{u}}(t)}{\frac{1}{2}\underline{\mathbf{w}}(t+1)^T\underline{V}(t)\underline{\mathbf{w}}(t+1)+\frac{1}{2}\left\|\underline{\mathbf{z}}(t)\right\|_1}-\frac{\lambda}{2}\left(\underline{\mathbf{w}}(t+1)^T\underline{U}(t)\underline{\mathbf{w}}(t+1)+\left\|\underline{\mathbf{w}}(t)\right\|_1\right)-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\end{aligned} \tag{34}$$

We also have

$$\begin{aligned}F(\mathbf{w}(t+1))&=\frac{\left\|\mathbf{w}(t+1)^TH_b\right\|_1}{\left\|\mathbf{w}(t+1)^TH_w\right\|_1}-\lambda\|\mathbf{w}(t+1)\|_1-\frac{\eta}{2}\|\mathbf{w}(t+1)\|_2^2\\&=\frac{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_b\right\|_1}{\left\|\underline{\mathbf{w}}(t+1)^T\underline{H}_w\right\|_1}-\lambda\left\|\underline{\mathbf{w}}(t+1)\right\|_1-\frac{\eta}{2}\left\|\underline{\mathbf{w}}(t+1)\right\|_2^2\end{aligned} \tag{35}$$

Combining Eq. (35), Eq. (34) and Eq. (26), we have

$$F(\mathbf{w}(t+1))\geq F(\mathbf{w}(t)) \tag{36}$$

3.3 Extension to multiple basis vectors

In subsection 3.2, we have shown how to obtain one optimal projection vector. Generally, one projection vectors is not enough. In the following, we will introduce how to learn multiple projection vectors. The deflation technique is used to extract the other projection vectors. If the first r-1 projection vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{r-1}$ have been obtained, then the rth projection vector $\mathbf{w}_r$ is calculated by using the deflated data

$$\left(\mathbf{x}_j^i\right)^{deflated} = \mathbf{x}_j^i - \sum_{l=1}^{r-1} \mathbf{w}_l \left(\mathbf{w}_l^T \mathbf{x}_j^i\right) \tag{37}$$

After $\left(\mathbf{x}_j^i\right)^{deflated}$ is obtained, we will use $\left(\mathbf{x}_j^i\right)^{deflated}$ to compute H_b and H_w. Then the procedure in subsection 3.2 is used to compute $\mathbf{w}_r$.

4 Experiments and results

In this section, we will compare our proposed SLDA-L1 with LDA-L2 [4], LDA-L1 [31, 39] and R1-MMC [24] on artificial datasets and some image databases, e.g. Yale, FERET, FRGC and COIL-20. In order to overcome the singularity problem of LDA-L2, we use PCA to reduce the dimension of training sample before the use of LDA. In the PCA phase, we keep nearly 98% image energy. For SLDA-L1, the parameters λ and η are determined by cross-validation. In order to select λ and η conveniently, we assign λ and η as the same value. The learning rate β in SLDA-L1 and LDA-L1 is tried from the set [0.01, 0.03, 0.05, 0.07, 0.09] and the value, which can produce the largest objective function value, is selected. The initial projection vectors of LDA-L1 and SLDA-L1 are set as the projection vectors of LDA-L2. In the experiments, we firstly use the compared methods, i.e., SLDA-L1, LDA-L1, LDA-L2 and R1-MMC, to extract the features of intensity images. Then the nearest neighbor classifier (1-NN) with Euclidean distance is used for classification. The programming environment is MATLAB 2008.

4.1 Experiments on Yale face database

The Yale face database contains 165 Gy scale images of 15 individuals, each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). In our experiments, each image in Yale database was cropped and resized to 64 × 64.

In the experiments, we randomly choose i (i = 4,5) samples of each person for training, and the remaining ones are used for testing. The procedure is repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 1. To test the

Table 1 Comparison of recognition rates for the different methods on Yale database without noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
4	86.05 ± 2.1	87.7 ± 2.7	88.3 ± 2.5	88.4 ± 2.2
5	90.4 ± 1.8	91.5 ± 1.8	92.0 ± 1.9	92.2 ± 1.8

Fig. 2 Some face images with/without occlusion in Yale database

robustness of the proposed SLDA-L1 against outliers, we contaminate each image independently with the probability of 0.5 by rectangle noise. The rectangle noise takes white or black dots, its location in face image is random and its size is 40 × 40. Some face images with or without rectangle noise are shown in Fig. 2. Similarly, the procedure is also repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 2. We also illustrate the plot of recognition rate vs. the dimension of reduced space for different methods in Fig. 3.

4.2 Experiments on FERET face database

The FERET face database contains 14,126 images from 1199 individuals. In our experiments, we select a subset which contains 1400 images of 200 individuals (each individual has seven images). The subset involves variations in facial expression, illumination and pose. In our experiments, each image in FERET database was cropped and resized to 80 × 80.

In the experiments, we randomly choose i (i = 3,4) samples of each person for training, and the remaining ones are used for testing. The procedure is repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 3. To test the robustness of the proposed SLDA-L1 against outliers, we contaminate each image independently with the probability of 0.5 by rectangle noise. The rectangle noise takes white or black dots, its location in face image is random and its size is 50 × 50. Some face images with or without rectangle noise are shown in Fig. 4. Similarly, The procedure is also repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 4. We also illustrate the plot of recognition rate vs. the dimension of reduced space for different methods in Fig. 5.

4.3 Experiments on FRGC face database

The FRGC version 2 database contains 12,776 training images, 16,028 controlled target images, and 8014 uncontrolled query images for the FRGC Experiment 4. The controlled images have good image quality, while the uncontrolled images display poor image quality, such as large illumination variations, low resolution of the face region, and possible blurring. It is the uncontrolled factors that pose the grand challenge to face recognition performance. In

Table 2 Comparison of recognition rates for the different methods on Yale database with noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
4	74.8 ± 4.1	77.5 ± 4.1	76.4 ± 4.4	79.5 ± 3.7
5	79.4 ± 4.4	83.5 ± 4.7	82.4 ± 4.3	85.4 ± 2.2

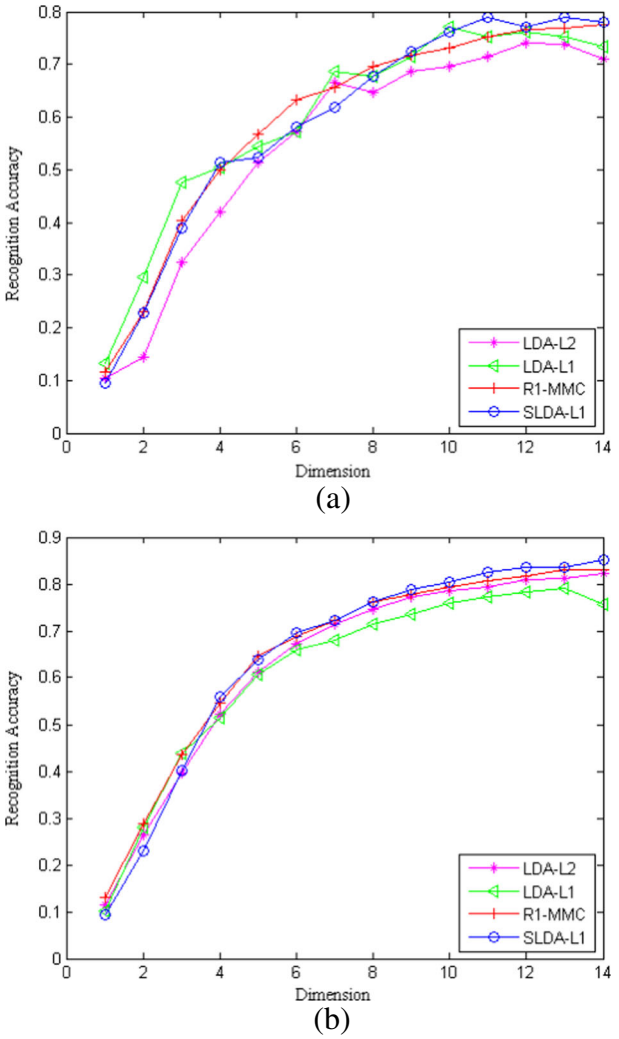


Fig. 3 Recognition rate vs. dimension of reduced space on the Yale database. (a) 4 Train, (b) 5 Train

our experiments, 100 people, each with 24 images, were chosen in the experiments. In the experiments, the first 12 samples of each person for training, and the remaining ones are used for testing. Each image was cropped to the size of 60 × 60. The maximal recognition rates of each method are listed in Table 5 and the recognition rates vs. the variations of the dimensions is shown in Figs. 6 and 7. To test the robustness of the proposed SLDA-L1 against outliers, we

Table 3 Comparison of recognition rates for the different methods on FERET database without noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
3	77.9 ± 1.7	81.6 ± 1.0	82.1 ± 1.1	84.2 ± 1.0
4	87.8 ± 2.0	89.9 ± 1.5	90.4 ± 1.2	93.6 ± 1.0

Fig. 4 Some face images with/without occlusion in FERET database

contaminate each image independently with the probability of 0.5 by rectangle noise. Some face images with or without rectangle noise are shown in Fig. 6. The rectangle noise takes white or black dots, its location in face image is random and its size is 25 × 25. The maximal recognition rates of each method are listed in Table 6.

4.4 Experiments on COIL-20 image database

The COIL-20 data set contains 1440 images of 20 objects. For each object, 72 images were captured with a black background from varying angles. The moving interval of the camera is five degrees. Each image is resized to 32 × 32 in our experiment.

In the experiments, we randomly choose ten images of each object for training, and the remaining ones are used for testing. The procedure is repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 7.To test the robustness of the proposed SLDA-L1 against outliers, we contaminate each image independently with the probability of 0.5 by rectangle noise. The rectangle noise takes white or black dots, its location in face image is random and its size is 16 × 16. Some images with or without rectangle noise are shown in Fig. 8. Similarly, the procedure is repeated 10 times and the average recognition rates as well as the standard deviation are reported in Table 8. We also illustrate the plot of recognition rate vs. the dimension of reduced space for different methods in Fig. 9.

4.5 Experiments based on LBP descriptor

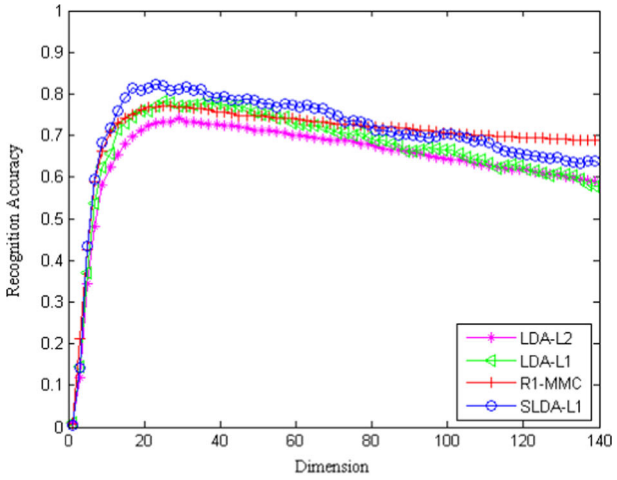
In order to further evaluate the recognition performance of the SLDA-L1 method, we use the method of applying the LBP descriptor [1, 37] to Yale face database. We firstly add noises to Yale face images as in Section 4.1. Then the $LBP^{u2}_{8,2}$ operator is applied to extract the features of images. Finally, the same experiment procedures as those in Section 4.1 are conducted. The average recognition rates as well as the standard deviation are reported in Table 9.

4.6 Parameter sensitiveness

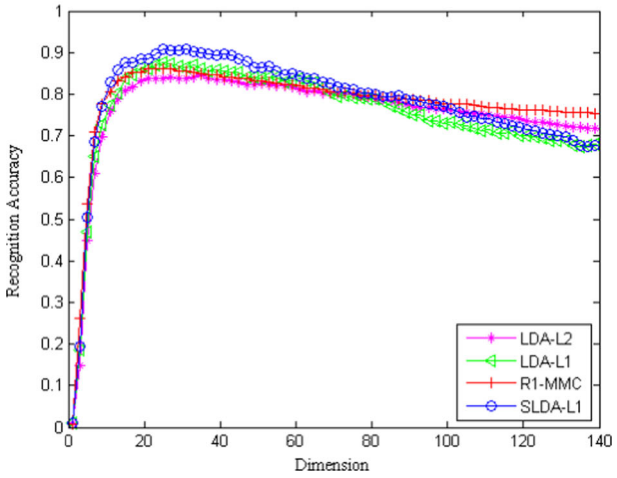
In SLDA-L1, we use the elastic net to regularize the projection vector, and then there are two parameters, i.e., λ and η, to be tuned. In this subsection, we investigate how

Table 4 Comparison of recognition rates for the different methods on FERET database with noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
3	74.2 ± 2.2	77.6 ± 1.3	77.9 ± 1.5	81.3 ± 2.2
4	84.9 ± 0.9	86.3 ± 2.4	87.9 ± 1.1	90.5 ± 1.1



(a)



(b)

Fig. 5 Recognition rate vs. dimension of reduced space on the FERET database. (a) 3 Train, (b) 4 Train

the recognition rates of the SLDA-L1 method depends on λ and η. λ and η are selected from [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.06, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20]. Fig. 10 shows the recognition performance of SLDA-L2 on the Yale database w.r.t. λ and η when four samples of each person are used for training.

Table 5 Comparison of recognition rates for the different methods on FRGC database without noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
12	67.6	68.3	69.6	71.8

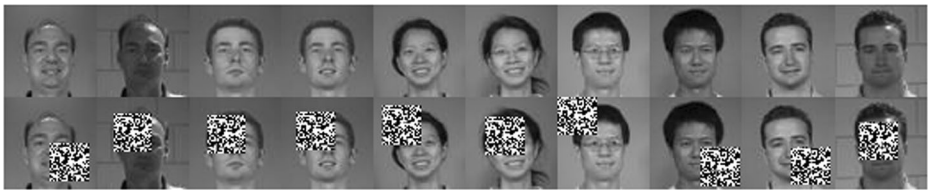


Fig. 6 Some face images with/without occlusion in FRGC database

Table 6 Comparison of recognition rates for the different methods on FRGC database with noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
12	59.5	62.5	64.0	67.2

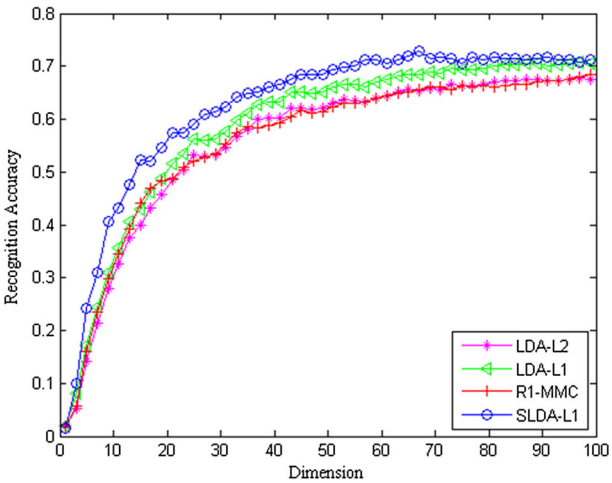


Fig. 7 Recognition rate vs. dimension of reduced space on the FRGC database

4.7 Discussions

From the experiment results we can find that L1-norm-based algorithms, i.e., LDA-L1, R1-MMC and SLDA-L1, can achieve higher classification rates than their L2-norm-based counterpart, i.e., LDA-L2. The reason is that L1-norm embedded in the objective function can suppress the negative

Fig. 8 Some images with/without occlusion in COIL-20 database

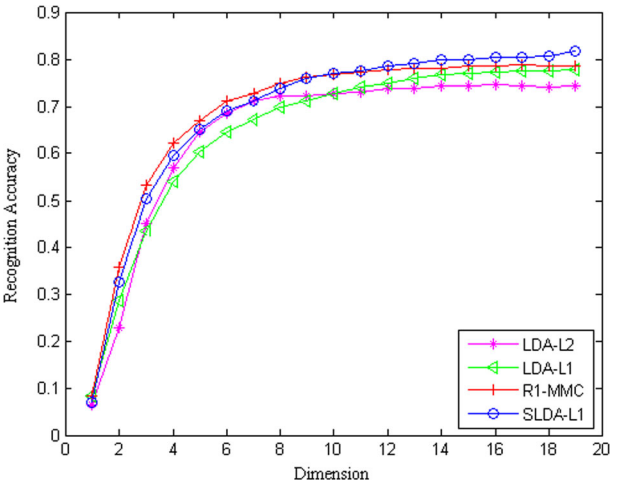


Fig. 9 Recognition rate vs. dimension of reduced space on the COIL-20 database

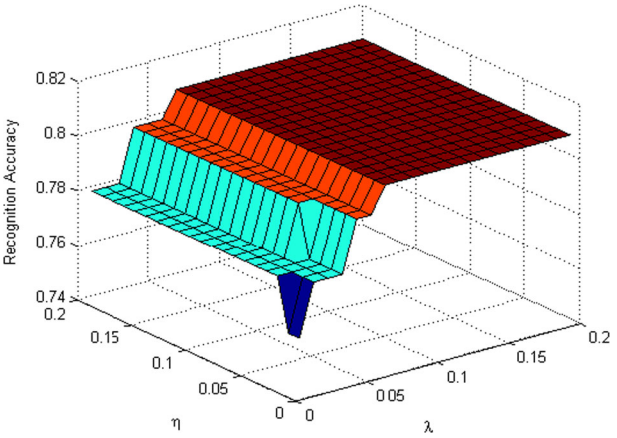


Fig. 10 The recognition performance on Yale database by varying λ and η

Table 7 Comparison of recognition rates for the different methods on COIL-20 database without noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
10	84.4 ± 1.2	88.5 ± 1.1	90.3 ± 0.7	91.7 ± 0.8

Table 8 Comparison of recognition rates for the different methods on COIL-20 database with noise

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
10	74.2 ± 2.5	78.4 ± 1.5	77.9 ± 1.8	81.8 ± 1.6

Table 9 Comparison of recognition rates on Yale database based on LBP descriptor

Sample size	LDA-L2	R1-MMC	LDA-L1	SLDA-L1
4	83.1 ± 4.0	86.0 ± 3.2	86.3 ± 3.1	88.6 ± 2.7
5	81.4 ± 4.3	89.2 ± 3.4	88.8 ± 4.1	91.6 ± 2.4

effects of outliers. Besides, the recognition rates of SLDA-L1 are higher than those of LDA-L1 and R1-MMC, which suggests that introducing the elastic net regularization term into LDA-L1 could improve the recognition performance further. From the experiments in Section 4.5, we can find that using the LBP descriptor can improve the recognition rates of all the algorithms. However, the recognition performance of SLDA-L1 are still higher than those of the other algorithms. Besides, from the experiments in Section 4.6 we can find that the recognition performance of SLDA-L1 depend on the tuning of λ and η. However, when λ and η varies from 0.01 to 0.2, the recognition performance of SLDA-L1 is relatively stable.

5 Conclusions

In this paper, we proposed a novel feature extraction method, called sparse L1-norm-based linear discriminant analysis (SLDA-L1), of subspace based on the L1-norm. First, the proposed SLDA-L1 method seeks projection vectors to maximize the between-class scatter matrix and meanwhile minimized the within-class scatter matrix, both of which are based on L1-norm instead of the conventional L2-norm. Secondly, L1-norm is also used in SLDA-L1 to regularize the basis vectors. Then, L1-norm used in SLDA-L1 is for both robust and sparse modelling simultaneouslyThe experiment results on some image databases show the effectiveness of the proposed SLDA-L1. In the future, we will investigate nonlinear feature extraction methods based on L1-norm.

Acknowledgments This research is supported by supported by NSFC of China (No. 61572033, 71371012), the Natural Science Foundation of Education Department of Anhui Province of China (No.KJ2015ZD08), the Social Science and Humanity Foundation of the Ministry of Education of China (No. 13YJA630098).

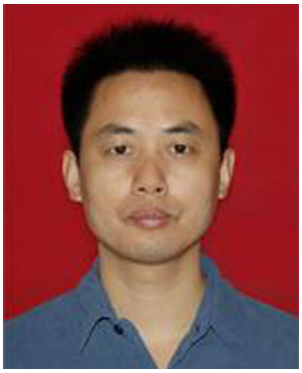
References

1. Ahonen T, Hadid A, Pietikäinen M (2006) Face description with local binary patterns: Application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041
2. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. J Am Stat Assoc 101(473):119–137
3. Barshan E, Ghodsi A, Azimifar Z, Jahromi MZ (2011) Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. Pattern Recogn 44(7):1357–1371
4. Belhumeour PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
5. Cai D, He X, Han J (2007) Spectral regression: a unified approach for sparse subspace learning. Proceeding of the 2007 International Conference on Data Mining (ICDM 07), Omaha, NE, 73–87
6. Cai D, He X, Han J (2008) Sparse projections over graph. Proceedings of the 21st AAAI conference on artifician intelligence
7. Ding C, Zhou D, He X, Zha H (2006) R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. Proceedings of the 23rd Internal Conference on Machine Learning, 281–288
8. Duda RO, Hart PE, Stork DG (2000) Pattern Classification, 2nd edn. John Wiley & Sons, New York
9. Fukunaga K (1990) Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, Boston, USA

10. Gu B, Sheng VS (2016) A robust regularization path algorithm for ν-support vector classification. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2016.2527796
11. Gu B, Sheng VS, Tay KY, Romano W, Li S (2015) Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst 26(7):1403–1416
12. Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S (2015) Incremental learning for ν-support vector regression. Neural Netw 67(7):140–150
13. Gu B, Sun X, Sheng VS (2016) Structural minimax probability machine. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2016.2544779
14. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: A review. IEEE Trans on Pattern Analysis and Machine Intelligence 22(1):4–37
15. Jenatton R, Obozinski G, Bach F (2010) Structured sparse principal component analysis. Proceeding of the 13th international conference on artificial intelligence and statistics, 366–373
16. Kawulok M, Wu J, Hancock ER (2011) Supervised relevance maps for increasing the distinctiveness of facial images. Pattern Recogn 44(4):929–939
17. Ke Q, Kanade T (2005) Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming, Proc IEEE Conf Comput Vis Pattern Recognit, San Diego, CA, USA, 20-26 June, vol. 1, p 1-8
18. Kwak N (2008) Principal component analysis based on L1-norm maximization. IEEE Trans on Pattern Anal Mach Intell 30(9):1672–1680
19. Leng L, Zhang J, Xu J, Khan MK, Alghathbar K (2010) Dynamic weighted discrimination power analysis in DCT domain for face and palmprint recognition. Int Conf Inf Commun Technol Convergence:467–471
20. Leng L, Zhang J, Khan MK, Chen X, Alghathbar K (2010) Dynamic weighted discrimination power analysis: A novel approach for face and palmprint recognition in DCT domain. Int J Phys Sci 5(17):2543–2554
21. Leng L, Zhang J, Chen G, Khan MK, Alghathbar K (2011) Two-directional two-dimensional random projection and its variations for face and palmprint recognition. Int Conf Comput Sci Appl, Santander, Spain, June 20-23, p 458–470
22. Li H, Jiang T, Zhang K (2004) Efficient and robust feature extraction by maximum margin criterion. Advances in Neural Information Processing Systems, Cambridge, MA, 97-104
23. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. IEEE Trans Neural Netw 17(1):1157–1165
24. Li X, Hua W, Wang H, Zhang Z (2010) Linear discriminant analysis using rotational invariant L1 norm. Neurocomputing 13-15(73):2571–2579
25. Meng D, Zhao Q, Xu Z (2012) Improve robustness of sparse PCA by L1-norm maximization. Pattern Recogn 45(1):487–497
26. Nie F, Huang H, Ding C, Luo D, Wang H (2011) Principal component analysis with non-greedy L1-norm maximization. The 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, 1-6
27. Pang Y, Li X, Yuan Y (2010) Robust tensor analysis with L1-Norm. IEEE Trans Circuits Syst Video Technol 20(2):172–178
28. Wang H (2012) Structured sparse linear graph embedding. Neural Netw 27:38–44
29. Wang H, Wang J (2013) 2DPCA with L1-norm for simultaneously robust and sparse modelling. Neural Netw 46(10):190–198
30. Wang H, Tang Q, Zheng W (2012) L1-norm-based common spatial patterns. IEEE Trans Biomed Eng 59(3):653–662
31. Wang H, Lu X, Hu Z, Zheng W (2014) Fisher discriminant analysis with L1-norm. IEEE Trans Cybernetics 44(6):828–842
32. Wen X, Shao L, Xue Y, Fang W (2015) A rapid learning algorithm for vehicle classification. Inf Sci 295(1):395–406
33. Xia J, Chanussot J, Du P, He X (2014) (Semi-) supervised probabilistic principal component analysis for hyperspectral remote sensing image classification. IEEE J Sel Top Appl Earth Observations Remote Sens 7(6):2224–2236
34. Xuelong L, Pang Y, Yuan Y (2009) L1-Norm-Based 2DPCA. IEEE Trans Syst Man Cybern B Cybern 40(4):1170–1175
35. Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin S (2007) Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
36. Yang J, Zhang D, Frangi AF, Yang JY (2004) Two-dimensional PCA: A new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell 26(1):131–137
37. Zhao G, Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressionas. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
38. Zheng W, Lin Z, Wang H (2014) L1-norm kernel discriminant analysis via Bayes error bound optimizatin for robust feature extraction. IEEE Trans Neural Netw Learn Syst 25(4):793–805

39. Zhong F, Zhang J (2013) Linear discriminant analysis based on L1-norm maximization. IEEE Trans Image Process 22(8):3018–3027
40. Zhou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15(2):265–286
41. Zhou T, Tao D, Wu X (2011) Manifold elastic net: A unified framework for sparse dimension reduction. Data Min Knowl Disc 22:340–371
42. Zhou Z, Wang Y, Wu QMJ, Yang C-N, Sun X (2016) Effective and efficient global context verification for image copy detection. IEEE Trans Inf Forensics and Secur. https://doi.org/10.1109/TIFS.2016.2601065

Gui-Fu Lu received the B.S degree in 1997 from Hefei University of Technology, P.R. China, the M.S. degree in 2004 from Hangzhou Institute of Electronics Engineering, and the PhD degree in 2012 from Nanjing University of Science and Technology, P.R. China. Since 2004, he has been teaching in the School of Computer Science and Information, AnHui Polytechnic University, WuHu, AnHui, China. His research interests include computer vision, digital image processing and pattern recognition. E-mail: luguifu_jsj@163.com



Jian Zou received the M.S. degree in applied mathematics from the Department of Mathematics of Nanjing University of Information Science & Technology, Nanjing, China, in 2006. He received the PhD degree in 2013 from Nanjing University of Science and Technology, P.R. China. His scientific interests are in the fields of pattern recognition, manifold learning and information statistics.

Yong Wang received the B.S. and M.S. degrees in computer science from Anhui university technology and science, WuHu, AnHui, China, in 2001, and 2007, respectively. Currently, he is with the School of Computer Science and Information, AnHui Polytechnic University, WuHu, AnHui, China. His research interests include software engineering and machine learning.

Zhongqun Wang is a professor in the School of Management Engineering, AnHui Polytechnic University, WuHu, AnHui, China. His research interests include software engineering and machine learning.