

L1-norm based null space discriminant analysis

Gui-Fu Lu¹ · Jian Zou¹ · Yong Wang¹ ·
Zhongqun Wang¹

Received: 30 December 2015 / Revised: 12 June 2016 / Accepted: 15 August 2016 /

Published online: 25 August 2016

© Springer Science+Business Media New York 2016

Abstract Null space based linear discriminant analysis (NSLDA) is a well-known feature extraction method, which can make use of the most discriminant information in the null space of within-class scatter matrix. However, the conventional formulation of NSLDA is based on L2-norm which makes NSLDA be sensitive to outlier. To address the problem of NSLDA, in this paper, we propose a simple and robust NSLDA based on L1-norm (L1-NSLDA). An iterative algorithm for solving L1-NSLDA is also proposed. Compared to NSLDA, L1-NSLDA is more robust than NSLDA since it is more robust to outliers and noise. Experiment results on some image databases confirm the effectiveness of the proposed L1-NSLDA.

Keywords Feature extraction · Dimensionality reduction · Null space linear discriminant analysis · L1-norm based null space linear discriminant analysis

1 Introduction

Feature extraction is a critical issue in the field of pattern recognition [6]. A main goal of feature extraction is to obtain a few low-dimensional representative features for the purpose of discrimination or data visualization. In past decades, many feature extraction algorithms have been proposed in the literatures [10]. The most famous feature extraction algorithms are perhaps principal component analysis (PCA) [6, 7] and linear discriminant analysis (LDA) [1, 6, 7].

PCA, which aims to find a set of orthogonal projection vectors to maximize the variance of the training samples, is an unsupervised feature extraction algorithm. On the contrary, LDA, which aims to find a set of projection vectors on which the training samples of the same class are as near as possible to each other while the training samples of the different classes are as far

✉ Gui-Fu Lu
luguifu_tougao@163.com

¹ School of Computer and Information, Anhui Polytechnic University, WuHu, Anhui 241000, China

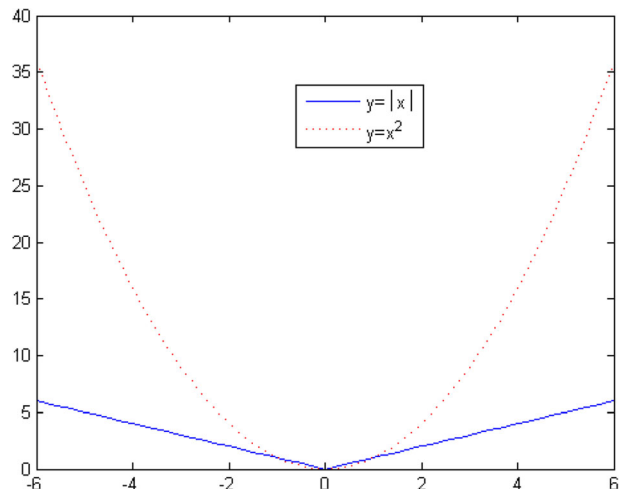
as possible from each other, is a supervised technique. Generally, LDA is more suitable than PCA for objection recognition problems.

The conventional LDA and PCA are both based on L2-norm, which is more sensitive to outliers than L1-norm since the square operation of the L2-norm can magnify the effect of outliers [12]. Then, to address the drawback of L2-norm based feature extraction methods, researchers turned to develop the L1-norm based feature extraction techniques, e.g. robust PCA [5, 11, 12, 18, 22] and robust LDA [15, 19, 20, 23, 25], in recent years. Compared with L2-norm based feature extraction methods, the main advantage of L1-norm based feature extraction methods is that they are less sensitive to effects of the outliers. The relationship between L1-norm and its robustness is explained intuitively in Fig. 1, where the solid line and the dot line correspond to L1-norm and L2-norm, respectively. From Fig. 1, we observe that comparing to L1-norm distance, L2-norm will exaggerate the influence of large errors to some extent, which are usually caused by outliers and noise.

By using maximum likelihood estimation, Ke et al. [11] proposed L1-PCA, which obtains the optimal projection matrix by using convex programming techniques. R1-PCA [5], which is rotational invariant, can combine the advantages of PCA and L1-PCA. However, L1-PCA and R1-PCA are more computationally expensive than the conventional L2-PCA. Recently, Kwak [12] proposed a rotational and robust L1-norm based PCA, i.e., PCA-L1, which can maximize the variance based on L1-norm and learn the optimal projection vectors by using a greedy iteration method. In contrast to L1-PCA and R1-PCA, PCA-L1 can obtain much lower reconstruction error in the facial image reconstruction. In [17], Nie et al. proposed a non-greedy strategy to solve PCA-L1 which can obtain much better projection matrix than that of L1-PCA. By using matrix and tensor techniques, Li et al. [22] and Pang et al. [18], respectively, generalizes L1-PCA to propose L1-norm based 2DPCA and tensor PCA.

For object recognition problems, it is more suitable to choose LDA rather than PCA since LDA can obtain the optimal discriminative projection matrix. By combining maximum margin criterion (WMMC) [13, 14] and R1-PCA, Li et al. proposed a new rotational invariant L1-norm based MMC, called as R1-MMC. However, R1-MMC is computationally expensive since its iterative algorithm to obtain the optimal projection matrix is based on eigenvalue decomposition. Recently, in order to extract robust electroencephalography (EEG) feature,

Fig. 1 Illustration of the exaggeration effect of the L2-norm and comparison with that of the L1-norm



Wang et al. [19] used L1-norm to replace L2-norm and proposed a L1-norm based common spatial patterns (L1-CSP), which can get better performance in the EEG classification experiments. Similar ideas are also appeared in [21, 24, 25], all papers replace L2-norm with L1-norm in the LDA objection function and propose L1-LDA. However, these methods use a greedy strategy to obtain the optimal projection vectors one by one.

Generally, the most discriminant information is contained in the null space of within-class scatter matrix and the method, which uses the discriminant information of null space and is originally proposed to address the small sample size (SSS) problem of LDA, is called null space based LDA (NSLDA) [3, 4, 9, 16]. Motivated by null space based LDA and L1-LDA, in this paper, we propose a L1-norm based null space linear discriminant analysis (L1-NSLDA). By analyzing the objective function of NSLDA, we get the transformed version of its objective function, which is formed using L2-norm. Similar to the other L1-norm based feature extraction methods, it is also very difficult to directly find the global optimal projection matrix of L1-NSLDA. To address the problem, we present a non-greedy iterative algorithm to obtain a local solution of L1-NSLDA.

The remainder of the paper is organized as follows. In section 2, we review briefly the related works on LDA and NSLDA algorithms. In Section 3, we propose the L1-NSLDA method, including its objective function and algorithmic procedure. Section 4 is devoted to the experiments. Finally, we conclude the paper in Section 5.

2 Outline of LDA and NSLDA

Given a data matrix $X = \{x_1, x_2, \dots, x_n\} = [X_1, \dots, X_c] \in \mathbb{R}^{d \times n}$, where $x_i \in \mathbb{R}^d$, for $i = 1, 2, \dots, n$, is the i th training sample in a d dimensional space, $X_i \in \mathbb{R}^{d \times n_i}$, for $i = 1, 2, \dots, c$, is a collection of training samples from the i th class and $\sum_{i=1}^c n_i = n$. Let N_i be the set of column indices that belongs to the i th class, i.e., x_j , for $j \in N_i$, belongs to the i th class. In LDA, the within-class scatter matrix, between-class scatter matrix and total scatter matrix are defined, respectively, as follows:

$$S_w = \sum_{i=1}^c \sum_{j \in N_i} (x_j - m_i)(x_j - m_i)^T = H_w H_w^T \tag{1}$$

$$S_b = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T = H_b H_b^T \tag{2}$$

$$S_t = \sum_{i=1}^n (x_i - m)(x_i - m)^T = H_t H_t^T = S_b + S_w \tag{3}$$

where m_i is the mean of the i th class and is defined as $m_i = \frac{1}{n_i} X_i e_i$, where $e_i = (1, 1, \dots, 1)^T \in \mathbb{R}^{n_i}$, m is the global mean and is defined as $m = \frac{1}{n} A e$, where $e = (1, 1, \dots, 1)^T \in \mathbb{R}^n$, H_b, H_w and H_t are defined, respectively, as

$$\begin{aligned} H_w &= [X_1 - m_1 e_1^T, \dots, X_c - m_c e_c^T] \in \mathbb{R}^{d \times n} \\ H_b &= [\sqrt{n_1}(m_1 - m), \dots, \sqrt{n_c}(m_c - m)] \in \mathbb{R}^{d \times c} \\ H_t &= X - m e^T = (x_1 - m, x_2 - m, \dots, x_n - m) \in \mathbb{R}^{d \times n} \end{aligned} \tag{4}$$

LDA aims to find an optimal projection matrix G that maximizes the following criterion:

$$G = \arg \max_G \text{trace} \left((G^T S_w G)^{-1} (G^T S_b G) \right) \tag{5}$$

where $\text{trace}(\cdot)$ denotes the trace operator. The projection matrix G can be obtained by solving the following generalized eigenvalue problem:

$$S_b g = \lambda S_w g, \quad \lambda \neq 0 \tag{6}$$

whose eigenvectors corresponding to the $c-1$ largest eigenvalues form the columns of G .

When the small sample size problem occurs, S_w is singular and LDA cannot work [8]. NSLDA, which can make use of the discriminant information in the null space of S_w , has been proposed to address the singularity of S_w . In particular, the optimization problem associated with NSLDA [3, 4, 9, 16] is

$$\begin{cases} G = \arg \max \text{trace} (G^T S_b G) \\ \text{trace} (G^T S_w G) = 0, G^T G = I \end{cases} \tag{7}$$

where I is an identity matrix.

3 L1-norm based null space linear discriminant analysis

3.1 Problem formulation

In this subsection, we will present our proposed L1-norm based null space linear discriminant analysis. We first reformulate the objective function of the conventional NSLDA method into a L2-norm based equation and then reveal that NSLDA is based on L2-norm distance criterion. It is well known that the L2-norm distance criterion is sensitive to outliers, which means that atypical samples may affect the desired solution to NSLDA. In literature, L1-norm is usually used as a robust alternative to L2-norm [19, 21, 24, 25]. Then motivated by this idea, we replace L2-norm in NLDSA with L1-norm and present the objective function of L1-NSLDA.

The optimization problem associated with NSLDA, i.e., Eq.(7), can be reformulated as

$$\begin{aligned} G &= \arg \max \text{trace} \left(G^T \left(\frac{1}{n} \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T \right) G \right) \\ \text{s.t. } G^T G &= I, G^T \left(\frac{1}{n} \left(\sum_{i=1}^c \sum_{j \in N_i} (x_j - m_i)(x_j - m_i)^T \right) \right) G = 0 \end{aligned} \tag{8}$$

By simply algebraic transformation, Eq.(7) can be rewritten as

$$\begin{aligned} G &= \arg \max \sum_{i=1}^c \left\| \sqrt{n_i} G^T (m_i - m) \right\|_2^2 \\ \text{s.t. } G^T G &= I, \sum_{i=1}^c \sum_{j \in N_i} \left\| G^T (x_j - m_i) \right\|_2^2 = 0 \end{aligned} \tag{9}$$

where $\|\cdot\|_2$ denotes L2-norm. From Eq. (9), we can find that NSLDA is based on L2-norm measurement, which is sensitive to the effect of outlier. Generally, we can replace L2-norm

with L1-norm to obtain a more robust method. Then, the optimization problem associated with L1-NSLDA is

$$\begin{aligned}
 G &= \operatorname{argmax} \sum_{i=1}^c \left\| \sqrt{n_i} G^T (m_i - m) \right\|_1 \\
 \text{s.t. } G^T G &= I, \sum_{i=1}^c \sum_{j \in N_i} \left\| G^T (x_j - m_i) \right\|_1 = 0
 \end{aligned}
 \tag{10}$$

where $\|\cdot\|_1$ denotes L1-norm. By using Eq. (4), Eq. (10) can be rewritten as

$$\begin{aligned}
 G &= \operatorname{argmax} \left\| G^T H_b \right\|_1 \\
 \text{s.t. } G^T G &= I, \left\| G^T H_w \right\|_1 = 0
 \end{aligned}
 \tag{11}$$

Although it is easy to obtain the global optimal solution of the objection function of conventional NSLDA, i.e., Eq. (7), it is very difficult to directly obtain the global optimal solution to L1-NSLDA, i.e. Eq. (11) since the absolute value operation is nonconvex. In the next subsection, we will propose an iterative algorithm to find a local optimal solution to Eq. (11).

3.2 Algorithm of finding the projection G

In this subsection, we will present an iterative procedure for finding the optimal projection G of L1-NSLDA.

Theorem 1 : Let Q be the matrix whose columns generate the null space of S_t , then we have $\|Q^T H_b\|_1 = 0$ and $\|Q^T H_w\|_1 = 0$.

Proof: Since Q be the null space of S_t , we have $Q^T S_t Q = 0$. Then we have $Q^T S_b Q = 0$ and $Q^T S_w Q = 0$ since S_b and S_w are both positive semi-definite matrices and $S_t = S_b + S_w$. From Eq. (1) and Eq. (2), we have

$$Q^T H_b = 0, \quad Q^T H_w = 0
 \tag{12}$$

Then we can obtain

$$\left\| Q^T H_b \right\|_1 = 0, \quad \left\| Q^T H_w \right\|_1 = 0
 \tag{13}$$

From Theorem 1, we can know that we can first remove the null space of S_t without loss of discriminant information.

Table 1 Computational complexity of each step in L1-NSLDA

No. of Step	1	2	3	4	5	6	7
L1-NSLDA	$O(dn)$	$O(dn^2)$	$O(dn^2)$	$O(n^3)$	$O(n^2c)$	$O(nc^2)$	$O(dnc)$

Table 2 Comparison of recognition rates for the different methods on ORL database without noise

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	89.8 ± 1.6	90.1 ± 1.3	92.0 ± 2.3	92.6 ± 2.1	94.2 ± 1.6	94.5 ± 1.3
5	93.1 ± 1.8	93.5 ± 1.6	93.0 ± 1.0	93.5 ± 1.9	96.3 ± 1.5	96.5 ± 1.4

Let $H_B = (Q^\perp)^T H_b, H_W = (Q^\perp)^T H_w$ and U be the solution of the following optimization problem

$$\begin{aligned}
 U &= \operatorname{argmax} \|U^T H_B\|_1 \\
 \text{s.t. } &U^T U = I, \|U^T H_W\|_1 = 0
 \end{aligned}
 \tag{14}$$

where Q^\perp is a matrix whose columns form an orthogonal basis of S_r . Then we can obtain the projection matrix G by $G = Q^\perp U$. So solving the projection matrix G boils down to solving the matrix U .

Theorem 2 : Let P be the null space of H_w , then $\|P^T H_w\|_1 = 0$ and $\|P^T H_B\|_1 \neq 0$.

Proof: Let $S_T = (Q^\perp)^T S_t Q^\perp, S_B = (Q^\perp)^T S_b Q^\perp$ and $S_W = (Q^\perp)^T S_w Q^\perp$, obviously we have

$$S_T = S_B + S_W \tag{15}$$

Since P is the null space of H_w , we have $P^T H_w = 0, \|P^T H_w\|_1 = 0$ and $P^T S_T P = 0$. From Eq. (15) we can obtain

$$P^T S_B P \neq 0 \tag{16}$$

That is

$$P^T H_B \neq 0 \tag{17}$$

and

$$\|P^T H_B\|_1 \neq 0 \tag{18}$$

From Theorem 2 we can know that solving the projection matrix U boils down to solving the matrix V , where V is the solution of the following optimization problem

$$\begin{aligned}
 V &= \operatorname{argmax} \|V^T \bar{H}_B\|_1 \\
 \text{s.t. } &V^T V = I
 \end{aligned}
 \tag{19}$$

where $\bar{H}_B = P^T H_B$ and $\bar{H}_W = P^T H_W$. Obviously we have $U = PV$.

Table 3 Computing times (S) of each algorithm on ORL database

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	0.236	35.097	0.279	8.728	0.848	1.034
5	0.288	46.221	0.297	9.476	1.057	1.167

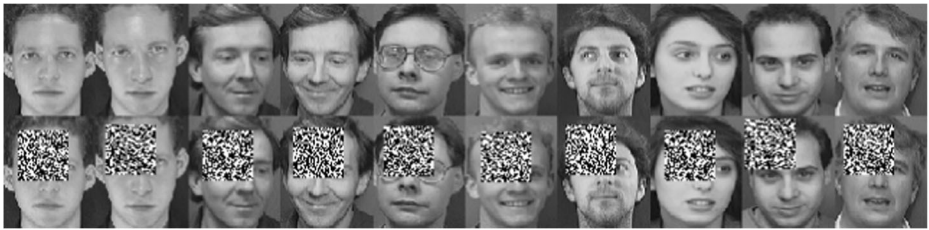


Fig. 2 Some face images with/without occlusion in ORL database

Now we consider how to solve V . Suppose $\alpha_i = \text{sgn}(V^T(\bar{m}_i - \bar{m}))$, $1 \leq i \leq c$, where $\bar{m}_i - \bar{m} = P^T(Q^\perp)^T \sqrt{n_i}(m_i - m)$. Then Eq. (19) can be reformulated as

$$\begin{aligned} V &= \underset{V^T V = I}{\text{argmax}} \|V^T \bar{H}_B\|_1 = \underset{V^T V = I}{\text{argmax}} \sum_{i=1}^c \|V^T (\bar{m}_i - \bar{m})\|_1 \\ &= \underset{V^T V = I}{\text{argmax}} \sum_{i=1}^c \alpha_i^T V^T (\bar{m}_i - \bar{m}) = \underset{V^T V = I}{\text{argmax}} \text{trace}(V^T M) \end{aligned} \tag{20}$$

where $M = \sum_{i=1}^c (\bar{m}_i - \bar{m}) \alpha_i^T$. By using Lagrange multiplier method, Eq. (20) can be rewritten as

$$f(V) = \text{tr}(V^T M) - \frac{1}{2} \text{tr}(L(V^T V - I)) \tag{21}$$

where L is a symmetric Lagrange multiplier matrix. Setting the partial derivative of $f(V)$ with respect to V equal to zero, we obtain

$$M - VL = 0 \tag{22}$$

That is

$$V = ML^{-1} \tag{23}$$

Since $V^T V = I$, we should have

$$(L^{-1})^T M^T M L^{-1} = I \tag{24}$$

Suppose the singular value decomposition (SVD) of M is $M = U_M \Lambda_M V_M^T$ and $L^{-1} = V_M \Lambda_M^{-1} V_M^T$, we have

$$(L^{-1})^T M^T M L^{-1} = (V_M \Lambda_M^{-1} V_M^T)^T (U_M \Lambda_M V_M^T)^T (U_M \Lambda_M V_M^T) (V_M \Lambda_M^{-1} V_M^T) = I \tag{25}$$

Then we can obtain

$$V = ML^{-1} = U_M \Lambda_M V_M V_M \Lambda_M^{-1} V_M^T = U_M V_M^T \tag{26}$$

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

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	80.9 ± 2.7	82.4 ± 2.7	89.8 ± 2.4	90.5 ± 1.9	93.2 ± 1.9	95.1 ± 1.7
5	83.4 ± 2.0	84.7 ± 1.9	93.1 ± 2.0	94.0 ± 1.7	95.2 ± 1.4	97.3 ± 1.4

Note that α_i , $1 \leq i \leq c$, is an unknown variable since it depends on V . We propose an iterative procedure to solve (19) and prove that the iterative procedure converges to a local solution. From the above analysis, the algorithm of solving (11) is described in Algorithm 1.

Algorithm 1: L1-NSLDA.

Input: data matrix X .

Output: transformation matrix G .

Step 1: Form the matrices H_b , H_w and H_t .

Step 2: Compute the economic SVD of H_t and obtain the matrix Q^\perp whose columns are all the left singular vectors corresponding to the nonzero singular values of H_t .

Step 3: Compute the matrices $H_B = (Q^\perp)^T H_b$ and $H_W = (Q^\perp)^T H_w$.

Step 4: Compute the SVD of matrix H_W and obtain the matrix P whose columns are all the left singular vectors corresponding to the zero singular values of H_W .

Step 5: Compute the matrices $\bar{H}_B = P^T H_B$ and $\bar{H}_W = P^T H_W$.

Step 6: Initialize V^1 such that $V^T V = I$, $t=1$

while *not converge* do

1. $\alpha_i = \text{sgn}((V^t)^T (\bar{m}_i - \bar{m}))$, $M^t = \sum_{i=1}^c (\bar{m}_i - \bar{m})(\alpha_i)^T$;

2. Calculate the SVD of M^t as $M^t = U_M^t \Lambda_M^t (V_M^t)^T$ and Let $V^t = U_M^t (V_M^t)^T$;

3. $t=t+1$;

end

Step 7: $G = Q^\perp P V^t$

In [2], Cevikalp et al. pointed out that all samples of the same class will be projected into the same unique vector in NSLDA, which makes NSLDA overfit to data hence it is not stable [4]. Then a small variation of the test samples in the range space of S_w will lead to

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

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	85.1 ± 1.9	85.3 ± 2.0	97.6 ± 1.3	97.7 ± 1.2	98.1 ± 1.2	98.2 ± 1.2
5	85.4 ± 1.7	85.5 ± 1.9	98.3 ± 1.0	98.3 ± 0.8	99.3 ± 1.0	99.4 ± 1.2

misclassification errors. The L1-NSLDA method will inherit the weakness of NSLDA since it is also a null space based method. However, L1-NSLDA finds the optimal projection matrix in the null space of S_w by using a L1-norm based objective function whereas NSLDA finds the optimal projection matrix in the null space of S_w by using a L2-norm based objective function. Then the ultimate projection matrices obtained by L1-NSLDA and NSLDA, respectively, are totally different and L1-NSLDA is not sensitive to the atypical samples. That is, L1-NSLDA can alleviate the effects of the outliers and noise when there exist outliers and noise in the training samples.

The following theorem 3 guarantees the convergence of the Step 6 of Algorithm 1.

Theorem 3 : The Step 6 of Algorithm 1 will monotonically increase the objective value of Eq. (19) in each iteration.

Proof: Obviously, we have

$$\begin{aligned} &\text{trace} \left((V^{t+1})^T \left(\sum_{i=1}^c (\bar{m}_i - \bar{m}) (\alpha_i^t)^T \right) \right) = \text{trace} \left((V^{t+1})^T M^t \right) \\ &\geq \text{trace} \left((V^t)^T M^t \right) = \text{trace} \left((V^t)^T \left(\sum_{i=1}^c (\bar{m}_i - \bar{m}) (\alpha_i^t)^T \right) \right) \end{aligned} \tag{27}$$

Since $\alpha_i^t = \text{sgn} \left((V^t)^T (\bar{m}_i - \bar{m}) \right)$ for each i , we have

$$\left\| (V^{t+1})^T (\bar{m}_i - \bar{m}) \right\|_1 = \alpha_i^{t+1} (V^{t+1})^T (\bar{m}_i - \bar{m}) \geq \alpha_i^t (V^{t+1})^T (\bar{m}_i - \bar{m}) \tag{28}$$

From Eq. (20) and Eq. (28), we have

$$\begin{aligned} \text{trace} \left((V^{t+1})^T M^{t+1} \right) &= \sum_{i=1}^c (\alpha_i^{t+1})^T (V^{t+1})^T (\bar{m}_i - \bar{m}) \geq \sum_{i=1}^c (\alpha_i^t)^T (V^{t+1})^T (\bar{m}_i - \bar{m}) \\ &= \text{trace} \left((V^{t+1})^T M^t \right) \end{aligned} \tag{29}$$

Table 6 Computing times (S) of each algorithm on Yale database

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	0.038	1.773	0.043	0.982	0.154	0.179
5	0.064	2.992	0.074	1.512	0.243	0.256

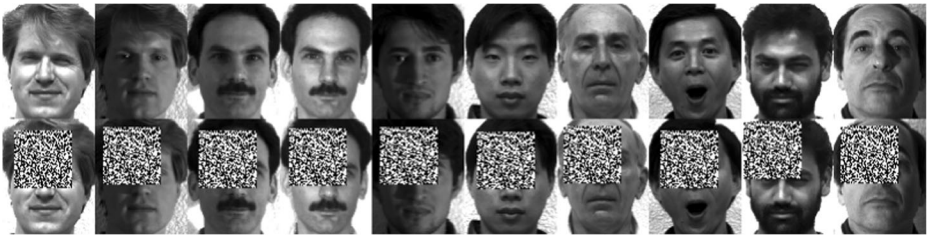


Fig. 3 Some face images with/without occlusion in Yale database

By combining Eq. (27) and Eq. (29), we have

$$\text{trace}\left(\left(V^{t+1}\right)^T M^{t+1}\right) \geq \text{trace}\left(\left(V^t\right)^T M^t\right) \quad (30)$$

Then the Step 6 of Algorithm 1 will monotonically increase the objective value of Eq. (19) in each iteration.

3.3 Computational complexity analysis

In this subsection, we will discuss the computational complexity of the proposed L1-NSLDA method. In Table 1, we list the computational complexity of each step in L1-NSLDA. Note that in Step 6 t denotes the iterative number. Generally, we have $d \gg n \gg c$ when the SSS problem occurs. Then from Table 1 we can know the total computational complexity of L1-NSLDA is in the order of $O(dn^2 + dnc + n^3 + n^2c)$.

4 Experiments and results

In this section, we will compare our proposed L1-NSLDA with PCA [7], L1-PCA [12], LDA [1], L1-LDA [20, 25] and NSLDA [3, 4, 9, 16] on ORL and Yale face databases. In order to overcome the small sample size (SSS) problem of LDA, we firstly use conventional PCA as a processing method. In the PCA phase we keep nearly 98 % image energy. For L1-LDA, the updating parameter β is set to 0.08. In the experiments we use the Euclidean distance based nearest neighbor classifier (1-NN) for classification and the recognition rate is computed by the following formula

$$\text{recognition rate} = \frac{\text{rec}}{\text{tot}} \quad (31)$$

where rec denotes the number of samples in the test set that is corrected labeled as being a given class and tot denotes the total number of samples in the test set. The experiments are implemented on a Mobile DualCore Intel Pentium (1.8GMHz) processor Acer Computer with 4GB RAM and the programming environment is MATLAB 2008.

Table 7 Comparison of recognition rates for the different methods on Yale database with noise

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
4	76.8 ± 3.1	77.9 ± 3.0	93.5 ± 2.0	94.1 ± 2.3	94.9 ± 1.7	96.3 ± 1.8
5	79.2 ± 4.1	80.3 ± 3.9	94.8 ± 1.8	95.6 ± 2.0	96.4 ± 2.3	97.9 ± 1.9

Table 8 Comparison of recognition rates for the different methods on FERET database without noise

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
3	62.6 ± 1.4	63.1 ± 1.5	72.1 ± 0.9	72.6 ± 1.9	85.0 ± 1.6	85.8 ± 1.3

4.1 Experiments on ORL face database

The ORL face database consists of a total of 400 face images, of a total of 40 people (10 samples per person). For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, front position (with tolerance for some side movement). In our experiments, the size of each image in ORL database is 112×92 .

Firstly, we test the performances of L1-NSLDA and the other methods on ORL database without added noise. 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 20 times and the average recognition rates as well as the standard deviation are reported in Table 2. We also reported the computing times of each algorithm in Table 3.

To test the robustness of the proposed L1-NSLDA against outliers, in the following we randomly choose 40 % of the training samples to be contaminated 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. 2. The procedure is repeated 20 times and the average recognition rates as well as the standard deviation are reported in Table 4.

4.2 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 manually cropped and resized to 100×80 .

Firstly, we test the performances of L1-NSLDA and the other methods on Yale database without added noise. 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 20 times and the average recognition rates as well as the standard deviation are reported in Table 5. We also reported the computing times of each algorithm in Table 6.

To test the robustness of the proposed L1-NSLDA against outliers, in the following we randomly choose 40 % of the training samples to be contaminated 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. 3. The procedure is repeated 20 times and the average recognition rates as well as the standard deviation are reported in Table 7.

Table 9 Computing times (S) of each algorithm on FERET database

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
3	1.857	260.2	1.957	67.434	4.307	5.818

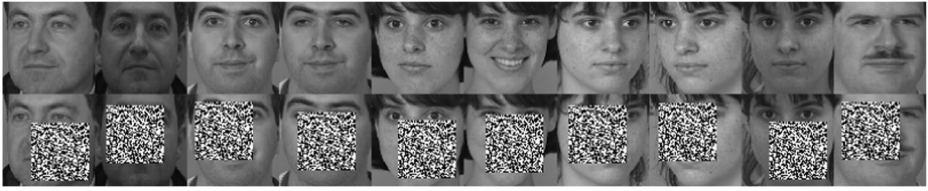


Fig. 4 Some face images with/without occlusion in FERET database

4.3 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 manually cropped and resized to 80×80 .

Firstly, we test the performances of L1-NSLDA and the other methods on FERET database without added noise. In the experiments, we randomly choose i ($i = 3$) samples of each person for training, and the remaining ones are used for testing. The procedure is repeated 20 times and the average recognition rates as well as the standard deviation are reported in Table 8. We also reported the computing times of each algorithm in Table 9.

To test the robustness of the proposed L1-NSLDA against outliers, in the following we randomly choose 40 % of the training samples to be contaminated 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. The procedure is repeated 20 times and the average recognition rates as well as the standard deviation are reported in Table 10.

4.4 Discussions

From the experiment results we can make the following conclusions:

- (1) From Table 2–10, we can find that the recognition rates of the unsupervised methods, e.g. PCA and L1-PCA, are generally lower than those of the supervised methods, e.g. LDA, L1-LDA, NSLDA and L1-NSLDA. This shows that the information of class label is critical to the recognition problems.
- (2) The L1-norm based methods can get higher recognition rates than their L2-norm-based counterparts. This shows that L1-norm is helpful for suppressing the negative effects of outliers indeed.
- (3) Generally, the L1-norm based methods spend more computing times to obtain the projection vectors than their L2-norm-based counterparts.

Table 10 Comparison of recognition rates for the different methods on FERET database with noise

Sample size	PCA	L1-PCA	LDA	L1-LDA	NSLDA	L1-NSLDA
3	51.8 ± 1.4	52.9 ± 1.9	63.4 ± 2.1	63.8 ± 1.7	66.8 ± 1.7	67.2 ± 1.3

- (4) Our proposed L1-NSLDA achieves the highest recognition rates in our experiments. This is attributed to the uses of discriminant information in the null space of within-class scatter matrix and L1-norm.

5 Conclusions

In this paper, a new feature extraction method, called L1-norm based null space linear discriminant analysis (L1-NSLDA), has been proposed. Since L1-norm can suppress the effects of outliers, L1-NSLDA is more robust to outliers than the conventional L2-norm based feature extraction methods. The experiment results on some image databases show the effectiveness of the proposed L1-NSLDA. In the future, we will investigate nonlinear feature extraction methods based on L1-norm.

Acknowledgments This research is supported by NSFC of China (No. 61572033, 71371012, 61672386), the Natural Science Foundation of Education Department of Anhui Province of China (No.KJ2015ZD08), 2014 Program for Excellent Youth Talents in University, the Social Science and Humanity Foundation of the Ministry of Education of China (No.13YJA630098), Anhui Provincial Natural Science Foundation (No.1608085MF147). The authors would like to thank the anonymous reviews and the editor for their helpful comments and suggestions to improve the quality of this paper.

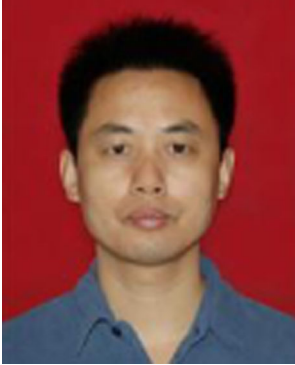
References

1. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE trans Pattern analysis and machine Intelligence* 19(7):711–720
2. Cevikalp H, Neamtu M, Wilkes M, Barkana A (2005) Discriminative common vectors for face recognition. *IEEE trans. Pattern analysis and machine Intelligence* 27(1):4–13
3. Chen LF, Liao HYM, Ko MT, Yu GJ (2000) A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recogn* 33(10):1713–1726
4. Chu D, Thye GS (2010) A new and fast implementation for null space based linear discriminant analysis. *Pattern Recogn* 43(4):1373–1379
5. Ding C, Zhou D, He X, Zha H (2006) R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In: *Proceedings of the 23rd internal conference on machine learning*, June 2006, pp 281–288
6. Duda RO, Hart PE, Stork DG (2000) *Pattern Classification*, 2nd edn. John Wiley & Sons, New York
7. Fukunaga K (1990) *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic Press, Boston, USA
8. Howland P, Wang J, Park H (2006) Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recogn* 39(2):227–287
9. Huang R, Liu Q, Lu H, Ma S (2002) Solving the small size problem of LDA, in: *Proc 16th Int'l Conf Pattern Recognition*. IEEE Computer Society, Quebec, Canada, pp. 29–32
10. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. *IEEE trans. On pattern analysis and machine Intelligence* 22(1):4–37
11. Ke Q, Kanade T (2005) Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming. In: *Proceedings of IEEE conference on computer vision and pattern recogn*, June 2005, pp 1–8
12. Kwak N (2008) Principal component analysis based on L1-norm maximization. *IEEE trans On pattern analysis and machine Intelligence* 30(9):1672–1680
13. Li H, Jiang T, Zhang K (2004) Efficient and robust feature extraction by maximum margin criterion. In: *Adv Neural Inf Proces Syst*. MIT Press, Cambridge, p 97–104
14. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. *IEEE trans On Neural Netw* 17(1):1157–1165
15. Li X, Hua W, Wang H, Zhang Z (2010) Linear discriminant analysis using rotational invariant L1 norm. *Neurocomputing* 13–15(73):2571–2579

16. Lu G-F, Zheng W (2013) Complexity-reduced implementations of complete and null-space-based linear discriminant analysis. *Neural Netw* 46(10):165–171
17. Nie F, Huang H, Ding C, Luo D, Wang H (2011) Principal component analysis with non-greedy L1-norm maximization. In: *The 22nd International joint conference on artificial intelligence (IJCAI)*, Barcelona, 2011, pp 1–6
18. Pang Y, Li X, Yuan Y (2010) Robust tensor analysis with L1-norm. *IEEE trans On circuits and Systems for Video Technology* 20(2):172–178
19. Wang H, Tang Q, Zheng W (2012) L1-norm-based common spatial patterns. *IEEE trans On Biomed Eng* 59(3):653–662
20. Wang H, Lu X, Hu Z, Zheng W (2013) Fisher discriminant analysis with L1-norm. *IEEE Trans on Cybernetics*. doi:[10.1109/TCYB.2013.2273355](https://doi.org/10.1109/TCYB.2013.2273355)
21. Wang H, Lu X, Hu Z, Zheng W (2014) Fisher discriminant analysis with L1-norm. *IEEE Trans on Cybernetics* 44(6):828–842
22. Xuelong L, Pang Y, Yuan Y (2009) L1-Norm-Based 2DPCA. *IEEE Trans on Systems, Man, and Cybernetics - Part B: Cybernetics* 40(4):1170–1175
23. Zheng W, Lin Z, Wang H (2013) L1-norm kernel discriminant analysis via Bayes error bound optimizatin for robust feature extraction. *IEEE Trans on Neural Networks and Learning Systems*. doi:[10.1109/TNNLS.2013.2281428](https://doi.org/10.1109/TNNLS.2013.2281428)
24. Zheng W, Lin Z, Wang H (2014) L1-norm kernel discriminant analysis via Bayes error bound optimizatin for robust feature extraction. *IEEE Trans on Neural Networks and Learning Systems* 25(4):793–805
25. Zhong F, Zhang J (2013) Linear discriminant analysis based on L1-norm maximization. *IEEE Trans Image Process* 22(8):3018–3027



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_tougao@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.