Neural Networks 102 (2018) 36–47

Contents lists available at ScienceDirect

Neural Networks

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

Inter-class sparsity based discriminative least square regression

Jie Wen [a,b], Yong Xu [a,b,*], Zuoyong Li [c], Zhongli Ma [c,d], Yuanrong Xu [a,b]

[a] Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, Guangdong, China
[b] Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, Guangdong, China
[c] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, 350121, Fujian, China
[d] College of Automation, Harbin Engineering University, Harbin, 150001, Heilongjiang, China

ARTICLE INFO

Article history:
Received 31 August 2017
Received in revised form 9 December 2017
Accepted 2 February 2018
Available online 21 February 2018

Keywords:
Least square regression
Inter-class sparsity
Multi-class classification
Supervised learning

ABSTRACT

Least square regression is a very popular supervised classification method. However, two main issues greatly limit its performance. The first one is that it only focuses on fitting the input features to the corresponding output labels while ignoring the correlations among samples. The second one is that the used label matrix, *i.e.*, zero–one label matrix is inappropriate for classification. To solve these problems and improve the performance, this paper presents a novel method, *i.e.*, inter-class sparsity based discriminative least square regression (ICS_DLSR), for multi-class classification. Different from other methods, the proposed method pursues that the transformed samples have a common sparsity structure in each class. For this goal, an inter-class sparsity constraint is introduced to the least square regression model such that the margins of samples from the same class can be greatly reduced while those of samples from different classes can be enlarged. In addition, an error term with row-sparsity constraint is introduced to relax the strict zero–one label matrix, which allows the method to be more flexible in learning the discriminative transformation matrix. These factors encourage the method to learn a more compact and discriminative transformation for regression and thus has the potential to perform better than other methods. Extensive experimental results show that the proposed method achieves the best performance in comparison with other methods for multi-class classification.

1. Introduction

Least squares regression (LSR) has been proved to be an effective technique in the community of pattern classification and computer vision, such as face recognition (Xu et al., 2014), microarray gene classification (Li & Ngom, 2013), cancer classification (Guyon, Weston, Barnhill, & Vapnik, 2002), speech recognition (Kim & Gales, 2011), and image retrieval (Feng, Zhou, & Lan, 2016). LSR aims at learning a transformation to connect the source data and target data with the minimum regression errors. In the past decades, various LSR based methods have been proposed, such as partial LSR (Abdi, 2010), local LSR (Ruppert, Sheather, & Wand, 1995), locally weighted LSR (Ruppert & Wand, 1994), kernel LSR (Gao, Shi, & Liu, 2007), and support vector machine (SVM) (Chang & Lin, 2011; Cherkassky & Ma, 2004). Besides, some representation based classification methods, such as linear regression based classification (LRC) (Naseem, Togneri, & Bennamoun, 2010) and sparsity representation based classification (SRC) (Wright, Yang, Ganesh, Sastry, & Ma, 2009), can be also regarded as the LSR based methods since they use the LSR technique to learn the representation coefficient for classification. Moreover, the very popular subspace learning methods, such as principle component analysis (PCA) , linear discriminant analysis (LDA), locality preserving projections (LPP), and spectral clustering (SC) can be also extended to the LSR framework (Cai, He, & Han, 2007; De la Torre, 2012; Tibshirani, 2011; Wang & Gao, 2015; Wen et al., 2018; Ye, 2007; Zou, Hastie, & Tibshirani, 2006). Compared with the conventional subspace learning methods, the LSR-type methods are more favorable since it is flexible to introduce various meaningful regularizations to improve their interpretability and performances. Moreover, the LSR-type methods can overcome the small-sample-size problem and greatly improve the computational efficiency (Fang, Xu, Li, Lai, Teng et al., 2017; Tibshirani, 2011).

Linear regression (LR) is one of the most popular supervised LSR methods. It has been applied in various classification tasks owing to its good performance and computational efficiency. For multi-class classification tasks, the standard LR first defines a label matrix according to the class labels and then seeks for a transformation matrix that can perfectly transform the samples into their corresponding labels. Under a mild condition, LR is equivalent to the well-known discriminative feature extraction method, *i.e.*, LDA, for multi-class classification (Ye, 2007). LDA seeks for a linear

* Corresponding author at: Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, Guangdong, China.
E-mail address: yongxu@ymail.com (Y. Xu).

https://doi.org/10.1016/j.neunet.2018.02.002
0893-6080/

projection that can pull the samples of same class together and push the samples of different classes far away in the discriminative subspace (Li, Chen, Nie, & Wang, 2017; Wang, Meng, & Li, 2017). Compared with LDA, LR is more flexible and efficient. For example, the sparsity techniques, such as the lasso constraint (l_1 norm) and row-sparsity constraint ($l_{2,1}$ norm) can be simply introduced into the model of LR to improve the interpretability and effectiveness. Introducing the sparsity technique also allows the learned transformation matrix to select the most discriminative features for classification, which is beneficial to improve the performance (Tibshirani, 2011; Xiang, Nie, Meng, Pan, & Zhang, 2012).

However, many issues still exist in the above LR based methods. The first issue is that the target matrix, *i.e.*, zero–one label matrix, is inappropriate for classification (Cai, Ding, Nie, & Huang, 2013; Wang & Pan, 2017; Xiang, Nie et al., 2012; Zhang, Lai et al., 2017; Zhang, Wang, Xiang, & Liu, 2015). For the strict zero–one label matrix, the Euclidean distances of regression responses between samples from different classes are a constant value, *i.e.*, $\sqrt{2}$. This is contrary to the expectation that samples from different classes should be as far as possible after transformation. The second issue is that these LR based methods only focus on fitting the samples to the corresponding labels while ignoring the relationships among samples, which may destroy the underlying structure of data and lead to the overfitting problem (Argyriou, Evgeniou, & Pontil, 2008; Bunea, She, & Wegkamp, 2011; Cai, Ding et al., 2013; Xiang, Zhu, Shen, & Ye, 2012). To solve these problems, many methods have been developed. For example, many researchers proposed to perform the regression on the relaxed label matrix rather than the strict zero–one matrix, in which the most representative works are the discriminative LSR (DLSR) (Xiang, Nie et al., 2012), margin scalable discriminative LSR (MSDLSR) (Wang, Zhang, & Pan, 2016), and retargeted LSR (ReLSR) (Zhang et al., 2015). DLSR introduces the ε-dragging technique to enlarge the distances of regression targets of different classes (Xiang, Nie et al., 2012). Based on DLSR, MSDLSR further imposes a l_1 norm constraint on the dragging matrix to explicitly control the margin of DLSR (Wang et al., 2016). Different from DLSR and MSDLSR, ReLSR does not use ε-dragging technique to relax the label matrix. It directly leans the regression targets from the data by introducing a margin constraint, where the margin between the true and false targets are enforced to be larger than one (Zhang et al., 2015). To emphasize the correlations among samples, the graph regularization term is introduced to the LR, which allows to learn a more compact representations and avoids the overfitting problem (Fang, Xu, Li, Lai, Wong et al., 2017; Xue, Chen, & Yang, 2009). Some researchers also proposed the low-rank linear regression (LRLR) models, in which the rank constraint, *i.e.*, nuclear norm, is imposed on the transformation matrix to explore the correlations among samples (Argyriou et al., 2008; Bunea et al., 2011; Cai, Ding et al., 2013; Xiang, Zhu et al., 2012).

Both the techniques mentioned above are useful and have the potential to improve the classification performance. However, relaxing the label matrix by introducing the ε-dragging technique or margin constraint will also enlarge the distances of the regression responses between samples from the same class, which is harmful to the classification. In this paper, a new relaxed label regression method named inter-class sparsity based discriminative least square regression (ICS_DLSR) is proposed to learn a more discriminative transformation. Different from the above methods, ICS_DLSR aims to preserve the row-sparsity consistency property of samples from the same class such that the distances of regression responses between samples from the same class can be greatly reduced, and thus can obtain a better performance. To this end, a novel inter-class sparsity regularization term is imposed on the transformation. Meanwhile, a sparsity error term with $l_{2,1}$ norm is introduced to the LSR model to relax the strict target label matrix for regression. Several experimental results show that ICS_DLSR can greatly improve the classification accuracies in comparison with the state-of-the-art methods. In brief, the proposed method has the following properties.

(1) The inter-class sparsity constraint is for the first time integrated into the LSR to exploit the relationships among samples. In particular, ICS_DLSR can learn a more compact and discriminative transformation that allows the transformed samples to have a common structure in each class.

(2) ICS_DLSR introduces a sparsity error term with $l_{2,1}$ norm to compensate the regression errors, which is beneficial to learn a more flexible transformation.

The rest of the paper is organized as follows. Section 2 introduces some notations and related works. In Section 3, the formulation and the optimal solution of the proposed method are presented. Then we analyze the proposed method in Section 4. Some experiments are conducted in Section 5 to prove the effectiveness of the proposed method. Section 6 offers the conclusion.

2. Related work

This section briefly introduces some related linear regression methods. For convenience, we first introduce some notations which are used throughout the paper. Let $X = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$ be the training set with n training samples from c classes, where m is the feature dimension of each sample. We use $X_i \in R^{m\times n_i}$ and n_i to denote the sub-training set and the number of samples of the ith class, respectively. For a vector $z = [z_1, z_2, \ldots, z_n]$, its l_2 norm is calculated as $\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2}$. For a matrix $W \in R^{c\times m}$, its l_1-norm, $l_{2,1}$-norm, and 'Frobenius' norm (l_F-norm) are calculated as $\|W\|_1 = \sum_{j=1}^{c}\sum_{i=1}^{m}|W_{ij}|$, $\|W\|_{2,1} = \sum_{i=1}^{c}\sqrt{\sum_{j=1}^{m} W_{ij}^2}$, $\|W\|_F^2 = \sum_{i=1}^{c}\sum_{j=1}^{m} W_{ij}^2$, respectively. $W_{i,j}$ denotes the (i,j)th element of matrix W. W^{-1} denotes the inverse matrix of matrix W. W^T is the transposed matrix of matrix W. We use a zero–one matrix $Y = [y_1, y_2, \ldots, y_n] \in R^{c\times n}$ to represent the label matrix corresponding to the training set X, where each column vector $y_i \in R^{c\times 1}$ is simply defined as follows: if training sample x_i comes from the kth class, then the kth element of column vector y_i is 1 while the remaining elements are 0. I is the identity matrix. Note that, the matrix based features such as image are pre-transformed into the column vector by stacking the matrix columns.

2.1. Standard LR (StLR) and low-rank LR (LRLR)

Given a training set $X \in R^{m\times n}$ and the corresponding label matrix $Y \in R^{c\times n}$, StLR aims at jointly learning a projection that can well transform the given training samples into their respective class labels as follows:

$$\min_{Q} \|Y - QX\|_F^2 + \lambda \|Q\|_F^2 \tag{1}$$

where $Q \in R^{c\times m}$ is the transformation matrix, λ is the regularization parameter with a small positive value. Problem (1) can be easily solved and has a closed solution as $Q = YX^T\left(XX^T + \lambda I\right)^{-1}$. For a test sample z, StLR predicts its label as $k = \mathrm{argmax}_i(Qz)_i$, where $(Qz)_i$ denotes the ith element of vector Qz.

To exploit the correlations reside in the high-dimensional data, LRLR replaces the l_F-norm regularization term with a low-rank constraint as follows:

$$\min_{Q} \|Y - QX\|_F^2 + \lambda \|Q\|_* \tag{2}$$

where $\|Q\|_*$ denotes the nuclear norm (trace norm) of matrix Q and is calculated as the sum of all singular values of matrix Q (Cai, Ding et al., 2013; Zhang, Lai et al., 2017). Compared with StLR, LRLR can discover the low-rank structures of data such that a more discriminative and compact transformation can be learned, and thus has the potential to obtain a better performance.

2.2. DLSR and ReLSR

Different from StLR and LRLR which use a strict zero–one label matrix as regression targets, DLSR and ReLSR are performed on the relaxed label matrix (Xiang, Nie et al., 2012; Zhang et al., 2015). The regression model of DLSR is formulated as follows (Xiang, Nie et al., 2012):

$$\min_{Q,M,b} \|Y + B \odot M - QX - be\|_F^2 + \lambda \|Q\|_F^2 \; s.t. M \geq 0 \tag{3}$$

where Q is the transformation matrix, λ is the regularization parameter with a small positive value, $b \in R^{c\times 1}$ is the bias vector, $e = [1, 1, \ldots, 1] \in R^{1\times n}$ is a row vector with all 1s. $\odot$ is a Hadamard product operator.[1] $M \geq 0$ means that all elements of matrix M are non-negative, *i.e.*, $M_{i,j} \geq 0$, where $M_{i,j}$ is the (i, j)th element of matrix M. Matrix B is defined as follows:

$$B_{i,j} = \begin{cases} +1, & if\ Y_{i,j} = 1 \\ -1, & otherwise \end{cases} \tag{4}$$

where $B_{i,j}$ is the (i, j)th element of matrix B. By introducing the label relaxed term, *i.e.*, $B \odot M$, into the regression loss function, the Euclidean distance of regression responses between two samples from different classes will be larger than $\sqrt{2}$. In addition, it is easy to observe that the margin between the true and false classes is also enlarged (Zhang et al., 2015).

Different from DLSR, ReLSR provides a more intuitive approach to enlarge the margin between the true and false classes as follows:

$$\min_{Q,T,b} \|T - QX - be\|_F^2 + \lambda \|Q\|_F^2 \; s.t. T_{r_j,j} - \max_{i\neq r_j} T_{i,j} \geq 1, \quad j = 1, 2, \ldots, n \tag{5}$$

where $T \in R^{c\times n}$ and $Q \in R^{c\times m}$ are the target matrix and transformation matrix need to be learned, respectively. r_j is the true label of sample x_j. By introducing the novel margin constraint, *i.e.*, $T_{r_j,j} - \max_{i\neq r_j} T_{i,j} \geq 1$, ReLSR can learn a more flexible label matrix and discriminative transformation matrix with the minimum regression loss, which enables to obtain a better classification performance.

3. ICS_DLSR for multi-class classification

In this section, we mainly present the motivation and formulation of the proposed method. We also provide an efficient approach to optimize the model.

3.1. Problem formulation and learning model

As mentioned previously, exploiting the correlations among samples are useful to learn a compact and discriminative transformation matrix (Cai, Ding et al., 2013; Xue et al., 2009). Inspired by this motivation, in this paper we propose a novel approach to exploit these correlations and learn a more discriminative transformation matrix. Different from the existing methods, the proposed method does not focus on preserving the structure of data, but tries to make the transformed samples of the same class have the common sparsity structure. Fig. 1 vividly shows the purpose of the proposed method. For this goal, we introduce a novel class sparsity constraint into the StLR as follows:

$$\min_{Q} \frac{1}{2} \|Y - QX\|_F^2 + \frac{\lambda_1}{2} \|Q\|_F^2 + \lambda_2 \sum_{i=1}^{c} \|QX_i\|_{2,1} \tag{6}$$

where λ_1 and λ_2 are the regularization parameters.

[1] For any two matrices $A \in R^{m\times n}$ and $B \in R^{m\times n}$ with the same dimension, the Hadamard product operator is defined as $(A \odot B)_{i,j} = A_{i,j} \cdot B_{i,j}$, where $(A \odot B)_{i,j}$, $A_{i,j}$, and $B_{i,j}$ denote the (i, j)th element of matrices $(A \odot B)$, A, and B, respectively.

Proposition 1. *Introducing constraint $\sum_{i=1}^{c} \|QX_i\|_{2,1}$ is able to make the transformed features QX have a common sparsity structure in each class.*

At the beginning, we should point out that minimizing $\sum_{i=1}^{c} \|QX_i\|_{2,1}$ is equivalent to minimizing $\|QX_i\|_{2,1}$ $(i = 1, \ldots, c)$ separately. Define $F = QX$, then $F_i \in R^{c\times n_i}$ is the transformed features of the subset of the ith class. Since $\|F_i\|_{2,1} = \left\|R^i\right\|_1$, where $R^i = \left[\left\|f_1^i\right\|_2, \left\|f_2^i\right\|_2, \ldots, \left\|f_c^i\right\|_2\right]^T$, f_j^i $(j = 1, \ldots, c)$ is the jth row vector of F_i. Obviously, minimizing $\|QX\|_{2,1}$ is equivalent to minimizing $\left\|R^i\right\|_1$, which will lead some elements of $\left\|R^i\right\|_1$ to be approximately zeros due to the sparsity selection property of the l_1 norm. Specifically, if the jth element of R^i is enforced to zero, *i.e.*, $\left\|f_j^i\right\|_2^2 = \left(f_{j1}^i\right)^2 + \left(f_{j2}^i\right)^2 + \cdots + \left(f_{jn_i}^i\right)^2 = 0$, then all elements corresponding to the jth row will be forced to zeros, where f_{jk}^i denotes the jth element of the kth sample from the ith class. So minimizing the constraint $\sum_{i=1}^{c} \|QX_i\|_{2,1}$ will make the transformed features QX have the same row-sparsity structure in each class.

Considering that the zero–one label matrix Y is too strict and inappropriate for classification (Wang & Pan, 2017), we introduce a sparsity error term to relax it as follows:

$$\min_{Q,E} \frac{1}{2} \|Y + E - QX\|_F^2 + \frac{\lambda_1}{2} \|Q\|_F^2 + \lambda_2 \sum_{i=1}^{c} \|QX_i\|_{2,1} + \lambda_3 \|E\|_{2,1} \tag{7}$$

where E denotes the errors and λ_3 is also the regularization parameter.

Since the extract features, *i.e.*, Qx_i, have very promising class sparsity property which contains natural distinguishability across samples, we utilize the extracted features to perform classification. That is, after calculating the transformation matrix Q, we perform the nearest neighbor classifier on the transformed features, *i.e.*, Qx_i, to obtain the final classification result.

3.2. Solution to ICS_DLSR

From the optimization problem (7), there are two unknown variables in one equation, which indicates that there is no analytical solution to the proposed method. In this section, we exploit the alternating direction method (ADM) (Lin, Chen, & Ma, 2010; Lin, Liu, & Su, 2011; Yang & Yuan, 2013) to optimize (7). We first introduce an extra variable F to make the optimization problem (7) separable as follows:

$$\min_{Q,F,E} \frac{1}{2} \|Y + E - QX\|_F^2 + \frac{\lambda_1}{2} \|Q\|_F^2 + \lambda_2 \sum_{i=1}^{c} \|F_i\|_{2,1} + \lambda_3 \|E\|_{2,1} \; s.t. F = QX. \tag{8}$$

Then we reformulate (8) into the following augmented Lagrangian function

$$\min_{Q,F,E} \frac{1}{2} \|Y + E - QX\|_F^2 + \frac{\lambda_1}{2} \|Q\|_F^2 + \lambda_2 \sum_{i=1}^{c} \|F_i\|_{2,1} + \lambda_3 \|E\|_{2,1} + \frac{\mu}{2} \left\|F - QX + \frac{C}{\mu}\right\|_F^2 \tag{9}$$

where C and μ are the Lagrange multiplier and penalty parameter, respectively. For problem (9), we can alternately solve each variable of Q, F, E with other variables fixed. The detailed solution steps are as follows.

Step 1. Update Q: By fixing variables F and E, Q can be obtained by minimizing the following objective:

$$L(Q) = \frac{1}{2} \|Y + E - QX\|_F^2 + \frac{\lambda_1}{2} \|Q\|_F^2 + \frac{\mu}{2} \left\|F - QX + \frac{C}{\mu}\right\|_F^2. \tag{10}$$

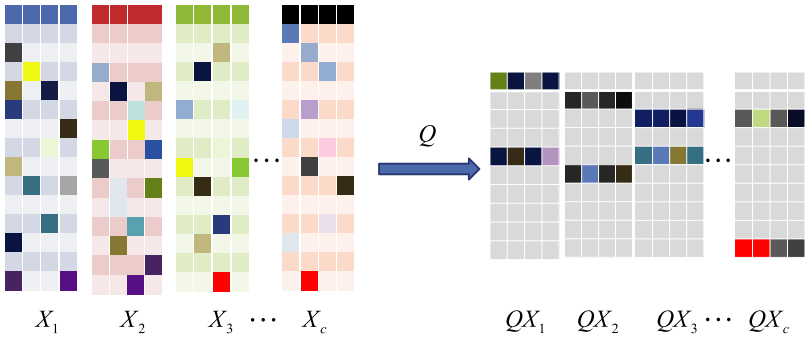


Fig. 1. Illustration of the proposed method. Note: in this figure, each column can be regarded as a sample. $X_1, X_2, \ldots, X_c$ denote the sample set of the 1th-cth classes. Q is the transformation matrix, $QX_1, QX_2, \ldots, QX_c$ are the projected samples.

The optimal Q can be obtained by setting the derivation of $L(Q)$ with respect to Q to zero. That is:

$$\begin{aligned}\frac{\partial L(Q)}{\partial Q} &= (QX - Y - E)X^T + \lambda_1 Q + \mu\left(QX - F - \frac{C}{\mu}\right)X^T \\ &= 0 \\ \Rightarrow Q &= (G_1 + \mu G_2)X^T\left((1+\mu)XX^T + \lambda_1 I\right)^{-1}\end{aligned} \tag{11}$$

where $G_1 = Y + E, G_2 = F + C/\mu$.

Step 2. Update F: Fixing variables Q and E, variable F can be obtained by minimizing the following formula:

$$\min_F \lambda_2 \sum_{i=1}^{c} \|F_i\|_{2,1} + \frac{\mu}{2}\left\|F - QX + \frac{C}{\mu}\right\|_F^2. \tag{12}$$

Define $H = QX - C\mu$, then the optimization problem (12) is equivalent to the following minimization problem

$$\begin{aligned}\min_F \sum_{i=1}^{c} \lambda_2 \|F_i\|_{2,1} + \frac{\mu}{2}\|F_i - H_i\|_F^2 \Leftrightarrow \sum_{i=1}^{c} \min_{F_i} \lambda_2 \|F_i\|_{2,1} \\ + \frac{\mu}{2}\|F_i - H_i\|_F^2\end{aligned} \tag{13}$$

where F_i and H_i are the ith subset of F and H corresponding to the samples of the ith class, respectively. From (13), it is obvious that solving F is equivalent to solving each subset F_i independently (Cai, Nie, & Huang, 2013; Liu, Ji, & Ye, 2009; Liu et al., 2013). According to Theorem 5 in Liu et al. (2009), we can obtain the optimal solution F_i as follows

$$[F_i]_{j,:} = \begin{cases} \dfrac{\left\|[H_i]_{j,:}\right\|_2 - \lambda_2\mu}{\left\|[H_i]_{j,:}\right\|_2}[H_i]_{j,:}, & if\ \left\|[H_i]_{j,:}\right\|_2 > \lambda_2\mu \\ 0, & otherwise \end{cases} \tag{14}$$

where $[F_i]_{j,:}$ and $[H_i]_{j,:}$ denote the jth row vector of F_i and H_i, respectively. When each subset F_i is calculated by (14), we can obtain the optimal F.

Step 3. Update E: Fixing variables Q and F, we can obtain E by minimizing the following $l_{2,1}$ norm based objective function:

$$\min_E \frac{1}{2}\|Y + E - QX\|_F^2 + \lambda_3 \|E\|_{2,1}. \tag{15}$$

Define $U = QX - Y$, E is obtained as follows:

$$E_{j,:} = \begin{cases} \dfrac{\left\|U_{j,:}\right\|_2 - \lambda_3}{\left\|U_{j,:}\right\|_2}U_{j,:}, & if\ \left\|U_{j,:}\right\|_2 > \lambda_3 \\ 0, & otherwise \end{cases} \tag{16}$$

where $E_{j,:}$ and $U_{j,:}$ are the jth row vector of variable E and U, respectively.

Step 4. Update C, μ: Lagrange multiplier C, and penalty parameter μ are respectively updated as follows:

$$C = C + \mu(F - QX) \tag{17}$$

$$\mu = \min(\rho\mu, \mu_{\max}) \tag{18}$$

where ρ and $\mu_{\max}$ are constants.

The proposed algorithm is summarized in Algorithm 1. After obtaining the regression matrix, we use the approach listed in Algorithm 2 to classify the test sample.

Algorithm 1 : ICS_DLSR (solving problem (8))

Input: Data matrix $X \in R^{m\times n}$, label matrix $Y \in R^{c\times n}$, parameters $\lambda_1, \lambda_2, \lambda_3$.
Initialization: Matrix $Q \in R^{c\times m}$ with random values; $F = QX$, $C = 0, E = 0, \mu = 10^{-8}, \rho = 1.01, \mu_{\max} = 10^8$.
while not converged **do**
1. Update Q by using (11);
2. Update F by using (14);
3. Update E by using (16);
4. Update C, μ by (17) and (18), respectively.

end while
Output: Q, E

Algorithm 2 : Classification based on ICS_DLSR

Input: Training data $X \in R^{m\times n}$ with label matrix $Y \in R^{c\times n}$, test sample t.
Output: Predicted label of test sample t.

Step 1. Normalize each training and test samples into a unit vector by $x_i = x_i/\|x_i\|_2$.

Step 2. Project training and test samples onto Q by $\hat{X} = QX$, $\hat{t} = Qt$.

Step 3. Calculate the Euclidean distance between the projected test sample $\hat{t}$ and each projected training sample $\hat{x}_i$, then classify the test sample to the class with respect to the nearest neighbor training sample.

4. Analysis of the proposed method

4.1. Computational complexity

From the summarization of Algorithm 1, there are four main steps. The computational cost of step 4 can be ignored since this step only contains matrix multiplication and addition operations. From (14) and (16), it is obvious that the approach to solve the $l_{2,1}$

norm based optimization problem is also very simple, and thus the computational costs of steps 2 and 3 can also be ignored. For step 1, the main computational cost is the inverse operation, which is $O\left(m^3\right)$ to a $m \times m$ matrix. Thus the total computational complexity of ICS_DLSR is about $O\left(\tau m^3\right)$, where τ is the iteration number.

4.2. Convergence analysis

As presented in the previous section, the ADMM-style optimization approach is adopted to solve problem (9). Since the optimization problem (9) totally has three blocks, it is difficult to theoretically prove the strong convergence property of our optimization approach (Hong & Luo, 2012; Lin et al., 2010; Liu et al., 2013). Nevertheless, the following Theorem guarantees a weak convergence property of the proposed optimization approach (Fang, Teng et al., 2017; Kim, Lee, & Oh, 2015; Zhang, 2010).

Theorem 1. *Let* $\Gamma^t = \left(Q^t, F^t, E^t, C^t\right)$ *be the solution of problem (9) at the* t*th iteration. Assume sequence solution* $\left\{\Gamma^t\right\}_{t=1}^{\infty}$ *is bounded and satisfies the condition* $\lim_{t\to\infty}\left(\Gamma^{t+1} - \Gamma^t\right) = 0$*, then the accumulation point of sequence* $\left\{\Gamma^t\right\}_{t=1}^{\infty}$ *is a Karush–Kuhn–Tucker (KKT) point of problem (9). Whenever* $\left\{\Gamma^t\right\}_{t=1}^{\infty}$ *converges, it converges to a KKT point.*

Proof. Please refer to the Appendix for the detailed proof of Theorem 1.

Theorem 1 provides an assurance to the convergence property of the optimization approach listed in Algorithm 1 to some extent. We also use some experiments to prove the convergence property. Fig. 2 shows the objective function value and the classification accuracy versus the number of iterations of the proposed methods on the Extended Yale B face database and AR face database. The objective value is calculated as $L = \frac{1}{2}\|Y + E - QX\|_F^2 + \frac{\lambda_1}{2}\|Q\|_F^2 + \lambda_2\sum_{i=1}^{c}\|QX_i\|_{2,1} + \lambda_3\|E\|_{2,1}$. It is obvious that the objective value monotonically decreases to a stable point within a few iteration steps. This indicates that the optimization approach can find the local optima of the ICS_DLSR and converges fast. In summary, both of the theoretical analysis and experimental results prove that the proposed method has good convergence property.

4.3. Connections to other methods

(1) Connections to LRLR (Cai, Ding et al., 2013): From the objective function of LRLR and the proposed method, we can find that the above two methods adopt two different approaches to exploit the relationships of data for transformation learning. In LRLR, the low-rank constraint is imposed on the transformation matrix to exploit the unknown or hidden information of data for regression. Different from LRLR, the proposed method utilize the prior knowledge of data, *i.e.*, class information, to improve the discriminability of the transformation matrix. Obviously, compared with the unknown information, utilizing the prior knowledge is more possible to learn the optimal transformation for classification. So the proposed method can obtain a better performance than LRLR for multi-class classification.

(2) Connections to DLSR (Xiang, Nie et al., 2012) and ReLSR (Zhang et al., 2015): DLSR, ReLSR and the proposed method all utilize the relaxed label matrix for regression. From the objective functions of DLSR and ReLSR, *i.e.*, problems (3) and (5), we can find that these two methods only focus on enlarging the margins of samples from different classes, they cannot reduce the margins of samples from the same class. Intuitively, the margins of regression targets from the same class are 0 in the strict label matrix. However, for DLSR and ReLSR, the margins of targets from the same class are usually larger than 0 and may be greatly enlarged when the margins of different classes are enlarged. For example, for DLSR, let $\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$ be the standard label matrix Y of three samples, then its relaxed label matrix $(Y + B \odot M)$ can be expressed as $\begin{bmatrix}1+M_{1,1} & 1+M_{1,2} & -M_{1,3}\\ -M_{2,1} & -M_{2,2} & 1+M_{2,3}\end{bmatrix}$, where $M_{i,j} \geq 0$, $i, j \in \{1, 2, 3\}$. The margin of relaxed labels from the same class is calculated as $Mr_s = \sqrt{\left(M_{1,1} - M_{1,2}\right)^2 + \left(M_{2,1} - M_{2,2}\right)^2}$, the margin of relaxed labels from different classes is calculated as $Mr_d = \sqrt{\left(1 + M_{1,1} + M_{1,3}\right)^2 + \left(1 + M_{2,1} + M_{2,3}\right)^2}$. We can find that when one element such as $M_{1,1}$ is enlarged (other elements are fixed), Mr_d and Mr_s will be also enlarged. Enlarging the margins of the labels from the same class is harmful to the classification tasks. So DLSR and ReLSR are not the perfect regression models. Different from DLSR and ReLSR, the proposed method introduces an inter-class sparsity constraint to the model, which enforces the transformed samples of the same class to have the same sparsity structure. In this way, the margins of the transformed samples from the same class will be reduced as much as possible and those of transformed samples from different classes will be simultaneously enlarged, which is beneficial to classification. Therefore, the proposed method has the potential to perform better than DLSR and ReLSR.

5. Experiments and analysis

In this section, we conduct several experiments to prove the effectiveness of the proposed method. In particular, the proposed method and some popular least square regression based methods, including linear regression classification (LRC) (Naseem et al., 2010), sparse representation based classification (SRC) (Wright et al., 2009), collaborative representation based classification (CRC) (Zhang, Yang, Feng, Ma, & Zhang, 2012), support vector machine (SVM) (Chang & Lin, 2011), low-rank linear regression (LRLR) (Cai, Ding et al., 2013), low-rank ridge regression (LRRR) (Cai, Ding et al., 2013), sparse low-rank regression (SLRR) (Cai, Ding et al., 2013), DLSR, and ReLSR, are fairly evaluated on the real datasets and synthetic datasets. For each group of experiments on the real datasets, all methods are repeated 10 times with the random combinations of training and test samples. Then their mean classification accuracies are reported for fairly comparing.

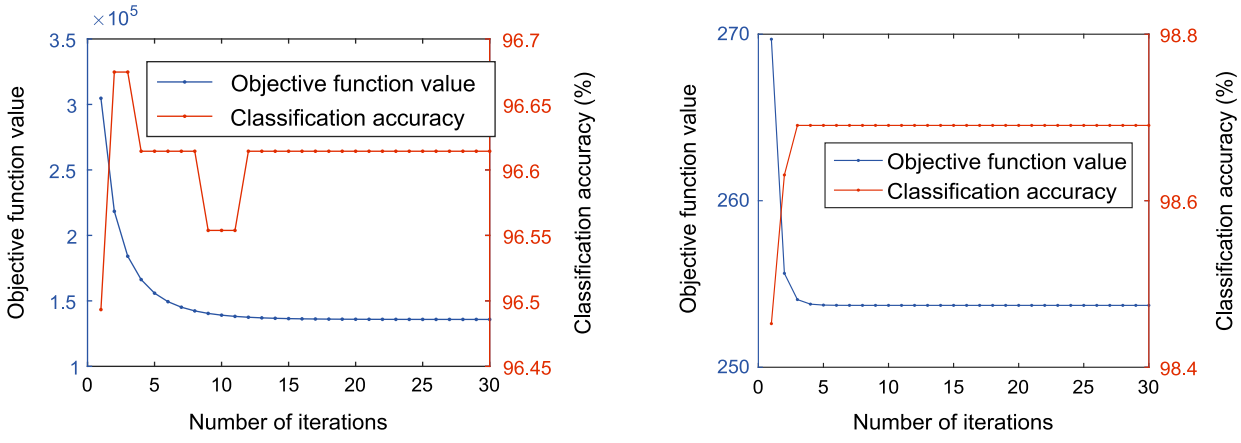
5.1. Experiments on face databases

Three popular face databases, *i.e.*, the Extended Yale B database[2] (Georghiades, Belhumeur, & Kriegman, 2001), the Labeled Faces in the Wild (LFW) database (Learned-Miller, Huang, RoyChowdhury, Li, & Hua, 2016), and the AR database[3] (Martinez, 1998) are adopted to evaluate different classification methods. Detailed information of these three databases is as follows:

(1) The Extended Yale B face database (Georghiades et al., 2001): There are 2414 images provided by 38 persons. Each class has 59–64 frontal images with different illuminations. Fig. 3(a) shows some images of the Extended Yale B database. All images used in the experiments were resized to 32×32 in advance. And then we randomly selected 10, 15, 20, and 25 samples from each class as the training set and treated the remaining samples as the test set. All compared methods were repeated 10 times for each group. Table 1 shows the mean classification accuracies of different methods on this database. It is obvious that the proposed method obtains the best performance.

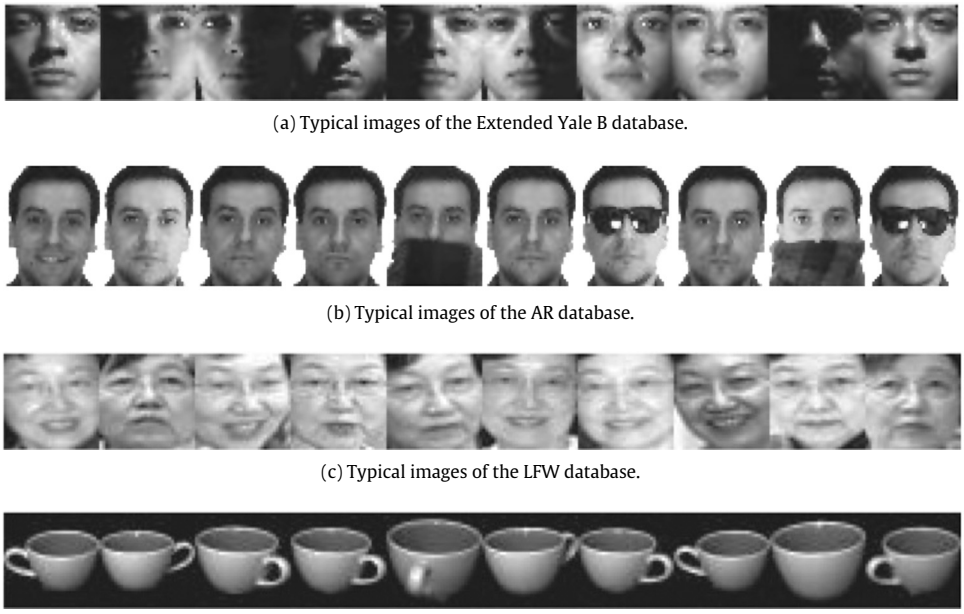
[2] The Extended Yale B dataset is available at: http://vision.ucsd.edu/iskwak/ExtYaleDatabase/ExtYaleB.html.

[3] The AR database is available at: http://www2.ece.ohio-state.edu/aleix/ARdatabase.html.



(a) Extended Yale B. (b) AR.

Fig. 2. Objective function value and classification accuracy versus the number of iterations of the proposed method on the (a) Extended Yale B and (b) AR datasets, in which 20 and 12 samples are randomly selected from each class as training samples, respectively. Note: it is acceptable that the recognition accuracy has the fluctuation of about 0.05% during the first few iterations.



(a) Typical images of the Extended Yale B database.

(b) Typical images of the AR database.

(c) Typical images of the LFW database.

(d) Typical images of the COIL20 database.

Fig. 3. Typical images of different databases.

Table 1
Mean classification accuracies (%) of different methods on the Extended Yale B face database.

No.	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
10	81.65	88.76	85.37	87.35	84.63	87.76	87.95	85.29	87.53	**89.44**
15	88.92	92.77	91.17	89.64	86.31	91.09	89.75	91.14	92.98	**93.89**
20	91.74	94.97	94.48	93.65	88.93	93.19	92.58	93.85	95.36	**96.80**
25	93.78	96.04	96.54	96.12	90.98	95.51	94.24	95.71	96.20	**97.62**

Note: bold numbers denote the best results.

(2) The AR face database (Martinez, 1998): There are over 4000 images of 126 people in the original AR face dataset. In the experiment, we chose a subset with 3120 images of 120 persons to evaluate different classification methods. Each class has 26 frontal face images with different facial expressions, illumination conditions, and occlusions by sun glasses and scarf. All images were resized to 50 × 40 in advance. Fig. 3(b) shows some images of the AR face database. Then we randomly selected 4, 6, 8, and 12 images from each class as the training set and treated the remaining samples as the test set. Similarly, all methods were also repeatedly performed 10 times and their mean values of the classification accuracies are reported for comparison. Experimental results on the AR face database are reported in Table 2. We can find that the proposed method consistently outperforms the other methods, which proves the effectiveness of the proposed method.

(3) The LFW face database (Learned-Miller et al., 2016): This database is designed for the unconstraint face recognition. Images in this database were collected from the web. The original LFW

Table 2
Mean classification accuracies (%) of different methods on the AR face database.

No.	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
4	68.88	88.66	84.39	68.76	83.39	87.98	85.78	86.34	88.34	**91.63**
6	80.09	93.18	89.63	82.98	85.21	90.71	91.25	91.35	93.03	**95.81**
8	87.20	95.02	93.49	89.91	87.02	93.54	93.47	94.78	95.62	**97.39**
12	95.14	96.91	95.30	95.61	90.14	96.38	95.82	97.27	97.57	**98.75**

Note: bold numbers denote the best results.

Table 3
Mean classification accuracies (%) of different methods on the LFW face database.

No.	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
5	29.73	30.12	29.38	26.04	30.24	33.37	30.57	27.90	31.81	**37.64**
6	32.18	31.44	32.51	29.52	33.29	35.24	34.15	30.80	34.45	**40.63**
7	34.53	32.51	33.64	30.60	34.96	35.59	34.36	33.73	37.70	**42.58**
8	37.23	34.55	35.12	33.14	35.59	36.52	35.64	36.80	40.37	**44.47**

Note: bold numbers denote the best results.

Table 4
Mean classification accuracies (%) of different methods on the COIL20 database.

No.	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
10	92.07	87.71	91.22	93.11	83.06	85.24	83.87	92.84	91.93	**93.97**
15	95.59	91.35	94.46	95.65	86.93	90.26	90.18	96.22	95.42	**97.81**
20	97.18	92.22	96.22	97.69	85.67	92.50	90.87	97.35	96.62	**98.44**
25	98.19	94.19	97.23	97.98	89.15	94.04	93.83	97.82	97.05	**98.80**

Note: bold numbers denote the best results.

database contains more than 13 000 images. In this paper, we use a subset of LFW which contains 1251 face images from 86 people to evaluate these classification methods (Wang et al., 2012). Fig. 3(c) shows some images of the LFW database. Each class has 11–20 images which were resized to 32 × 32 in advance. We randomly selected 5, 6, 7, and 8 samples from each class as the training set and treated the remaining samples as the test set. Experimental results on this database are reported in Table 3. From the experimental results, we can find that the classification accuracies of all methods are very low. This is mainly because that images of the LFW database are acquired in different conditions, such as different image acquirement devices, different environments, different poses, and different sizes. These uncontrolled factors lead to a very complex data distribution that is harmful to the classification. Nevertheless, the proposed method still achieves the highest classification accuracies on this database. From Table 3, we can find that the classification accuracy of the proposed method is at least about four percent higher than that of the second best competitor method, *i.e.*, ReLSR.

In conclusion, the above experimental results and analyses verify the effectiveness of the proposed method for face recognition.

5.2. Experiments on the object databases

In this subsection, we evaluate the proposed method on the Columbia Object Image Library (COIL20) database[4] (Nene, Nayar, Murase, et al., 1996). This database totally contains 1440 gray-scale images and 20 classes. Each class has 72 images which are taken at pose intervals of 5 degrees. Typical images of the COIL20 database are shown in Fig. 3(d). In the experiment, each image was resized to a 32 × 32 matrix in advance. Then we randomly select 10, 15, 20, and 25 samples from each class as the training set and treat the remaining samples as the test set to perform the experiment. All methods are repeated 10 times, and their corresponding mean classification accuracies are reported for fairly comparing. Experimental results of different methods on the COIL20 database are shown in Table 4. Obviously, the proposed method obtains competitive performance with LRC, SVM, DLSR, and ReLSR, and performs much better than the remaining methods. This proves the effectiveness of the proposed method for the object classification.

5.3. Experiments on the scene database

In this subsection, all supervised methods are compared on the Fifteen Scene Categories database[5] (Lazebnik, Schmid, & Ponce, 2006) to prove the effectiveness of the proposed method. The Fifteen Scene Categories database totally contains 4485 natural images of 15 categories, such as bedroom, industrial, coast, street, and building. Each category has 210–410 samples. The original typical images of the database are shown in Fig. 4. Followed by the experimental settings presented in Jiang, Lin, and Davis (2013), we use the spatial pyramid features of the Fifteen Scene Categories database[6] (Jiang et al., 2013) (Scene15_SPM database) to evaluate these compared methods. We first randomly selected 10, 20, 30, and 40 samples from each category of the Scene15_SPM database as the training set and treated the remaining samples as the test set. Then all methods are performed 20 times in each experimental group, and their mean classification accuracies (%) are reported for comparison. Experimental results of different methods are listed in Table 5. It is obvious that the proposed method obtains the best performance, which proves the effectiveness of the proposed method in dealing with the scene classification task.

5.4. Experiments on the synthetic dataset with manifold structure

In this subsection, we conduct some experiments to prove the effectiveness of the proposed in dealing with the data with manifold structure. The three-ring data is one of the most representative data with clear manifold structure. Following the experimental settings in Li et al. (2017), we also synthesize two three-ring data with different noises for evaluation.[7] Three-ring data contains three classes. In the experiment, we generate 1000 points for

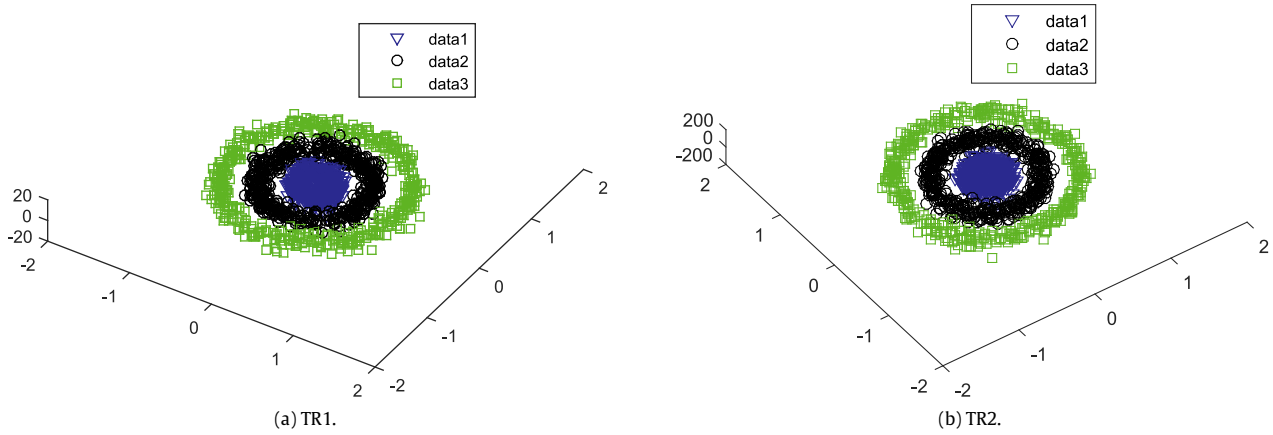
[4] The COIL20 dataset is available at: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[5] The Fifteen Scene Categories database is available at: http://www-cvr.ai.uiuc.edu/ponce_grp/data/.

[6] The spatial pyramid features of the Fifteen Scene Categories database is available at: http://www.umiacs.umd.edu/zhuolin/projectlcksvd.html.

[7] The code to generate the three-ring data is available at: http://www.escience.cn/people/fpnie/papers.html.

Fig. 4. Typical images of the Fifteen Scene Categories database, in which images of each column from the left to the right are from the category of the bedroom, industrial, coast, street, and building, respectively.



(a) TR1. (b) TR2.

Fig. 5. The two types of three-ring data.

Table 5
Mean classification accuracies (%) of different methods on the Scene15_SPM database.

No.	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
10	87.75	87.64	87.60	86.26	81.08	86.02	84.44	87.77	88.04	**89.73**
20	92.21	92.02	91.99	91.30	89.49	88.24	89.53	91.49	92.04	**93.40**
30	93.64	94.02	92.89	93.37	90.59	89.72	89.75	93.50	93.36	**95.27**
40	94.97	95.64	94.65	94.49	91.38	90.34	91.07	94.22	95.79	**96.39**

Note: bold numbers denote the best results.

Table 6
Classification accuracies (%) of different methods on the two three-ring data.

No.	NNC	LRC	CRC	SRC	SVM	LRLR	LRRR	SLRR	DLSR	ReLSR	ICS_DLSR
TR1	93.13	33.00	33.33	39.40	99.20	36.13	35.67	36.13	63.60	63.40	**99.93**
TR2	63.07	32.20	33.33	38.33	**83.45**	32.13	31.93	33.00	58.67	57.87	81.93

Note: NNC is the abbreviation of the most popular unsupervised classify, *i.e.*, nearest neighbor classify (NNC) (Bishop, 2006).

each class and randomly select 500 points from each class as the training set, and treat the remaining points as the test set. Each point of the three-ring data contains three features, in which the first two dimensions are distributed in concentric circles, while the other one dimension can be regarded as noises since it is a randomly value located in the range of $[-20, 20]$ and $[-200, 200]$, respectively. For simplicity, we refer to the two different three-ring data as TR1 and TR2, respectively. The two data are plotted in Fig. 5. Experimental results of different methods on the two kinds of three-ring data are enumerated in Table 6. From Table 6, we can find that the proposed method obtains better performance than the other regression based methods.

5.5. Experimental analysis

Tables 1–6 list the experimental results of the compared methods on different classification tasks. From these tables, we have the following observations:

(1) In most cases, the representation based regression methods, *i.e.*, LRC, CRC, and SRC, achieve better performance than the strict label matrix based regression methods, *i.e.*, LRLR, LRRR, and SLRR. Compared with all the other methods, the proposed method obtains the best performance on the above five databases, which proves the effectiveness of the proposed method for multi-class classification tasks.

(2) The relaxed label based regression methods, *i.e.*, DLSR, ReLSR, and ICS_DLSR, generally perform much better than those methods with strict label matrix, *i.e.*, LRLR, LRRR, and SLRR. This proves that using the relaxed label matrix is beneficial to improve the classification performance. This is mainly because that the relaxed label matrix has larger margins than the strict zero–one label matrix between different classes, which enables to learn a more discriminative transformation for classification.

(3) The proposed method consistently outperforms the other two relaxed label based regression methods, *i.e.*, DLSR and ReLSR, on the above five databases. This indicates that the proposed

(a) Original features. (b) Transformed features.

Fig. 6. t-SNE visualization of (a) the original features and (b) the transformed features obtained by our method on the Extended Yale B database, in which 25 samples per class are randomly selected as training samples to learn the transformation. All samples, including training and test samples, are visualized.

method can learn a more discriminative transformation than the other two methods. This is mainly because that DLSR and ReLSR only focus on enlarging the margin of samples from different classes while ignoring to reduce the distance of samples from the same class. Different from DLSR and ReLSR, the proposed method encourages the samples from the same class to have the same sparsity structure by introducing an inter-class sparsity regularization term. Preserving the inter-class sparsity structure is meaningful. It can reduce the margin of the intra-class as much as possible and simultaneously has the potential to enlarge the margin of the inter-class such that a better classification performance can be obtained. In addition, in Fig. 6, we have visualized the margins of the transformed samples obtained by our method by using the t-SNE[8] (Maaten & Hinton, 2008; Zhang, Shao, Xu, Liu, & Yang, 2017). From this figure, we can clearly see that our method can guarantee the samples of the same class distribute closely and push samples of different classes far away as much as possible.

(4) From the experimental results of Table 6, we can find that most of conventional methods except SVM all obtain very bad performance on the data with manifold structure. In particular, their performances are worse than the popular unsupervised classification method, *i.e.*, nearest neighbor classify (NNC) (Bishop, 2006). This indicates that these methods fail to extract the most discriminative features from the data with special manifold structure. Compared with the other methods, SVM and the proposed method obtain the satisfactory performance, especially on the TR2 data. This proves the effectiveness of the proposed method in dealing with the classification task on the data with such manifold structure.

5.6. Parameter sensitivity and selection

In the proposed method, there are three tuned parameters, *i.e.*, λ_1, λ_2, and λ_3, which are used to balance the importance of the corresponding constraint terms. The first constraint term is used to avoid the trivial solution of projection Q. The second constraint term guarantees the inter-class sparsity structure of data. The third constraint term relaxes the label matrix to adaptively fit the transformed data. The three constraint terms are all meaningful and beneficial to learn the optimal transformation matrix.

To analyze the parameter sensitivity of the proposed method, we first define a candidate set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$ for these parameters and then perform the proposed method with different combinations of these parameters. Figs. 7 and 8 show the relationships of the classification accuracy and the three parameters on the Extended Yale B and AR face databases. From Figs. 7(b) and 8(b), it is obvious that the classification accuracy is almost constant with respect to different values of parameter λ_2, which indicates that the proposed method is insensitive to the parameter λ_2. We can also find that the proposed method can obtain satisfactory performance when parameters λ_1 and λ_1 respectively locate in the range of $[10^{-5}, 10^{-3}]$ and $[10^{-3}, 10^{1}]$.

As far as we know, it is still an open problem to adaptively select the optimal regularization parameters for different databases. In this paper, we use the following strategy to find the optimal parameters for the proposed method (Fang, Teng et al., 2017). From the above analyses, the proposed method is insensitive to the selection of parameters λ_2 to some extent. Thus we can fix parameters λ_2 at first and define a candidate parameter range $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$ for parameters λ_1 and λ_3. By performing the proposed method with different combinations of parameters λ_1 and λ_3 selected from the candidate range, we can obtain a best combination of these two parameters. Then we fix the two parameters with the obtained best values and use the similar approach to find the optimal value from the candidate parameter range for parameter λ_2. Finally, we can find the optimal combinations of these three parameters from the 3D space formed by the candidate parameters of λ_1, λ_2, and λ_3. Then we perform the proposed method 10 times with the optimal parameters and report the mean classification accuracy for comparison.

6. Conclusion

In this paper, a novel least square error based supervised learning method is proposed for multi-class classification. Different from other methods which focus on fitting the samples to the strict zero–one label matrix, the proposed method tries to learn a more discriminative projection from a relaxed label matrix and pursues the transformed samples of the same class to have the same row-sparsity structure. By introducing the inter-class sparsity constraint, the proposed method can greatly reduce the margin of intra-class and simultaneously enlarge the margin of inter-class such that a better performance is guaranteed. Extensive experiments on the face, object, and scene databases prove the effectiveness of the proposed method.

[8] The code of t-SNE is available at: http://lvdmaaten.github.io/tsne/.

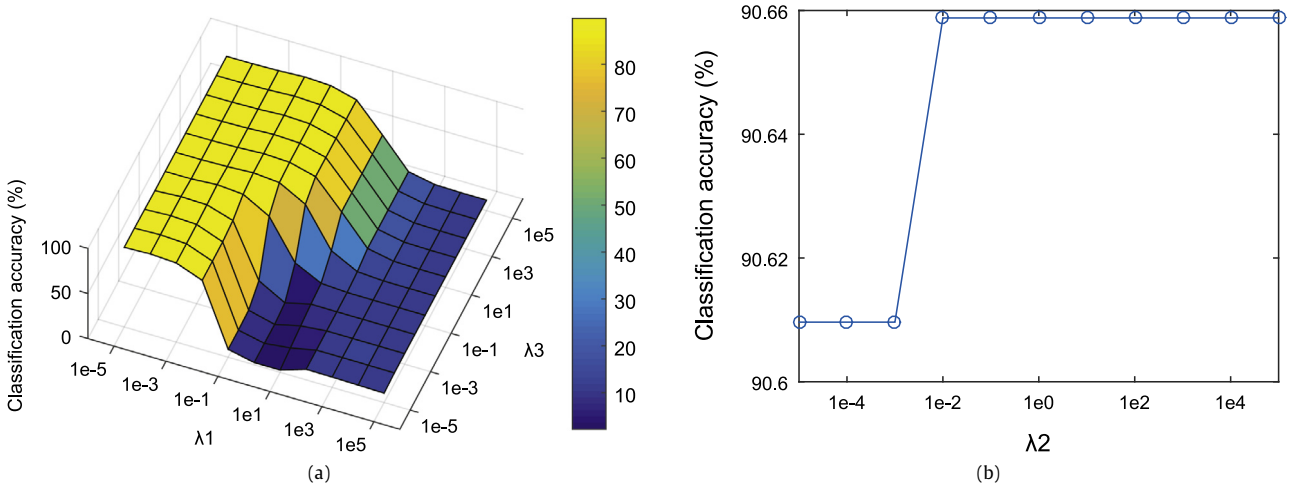


Fig. 7. Relationships of the classification accuracy (%) and different combinations of parameters on the Extended Yale B dataset, in which 10 samples of each class are selected as the training samples. (a) Parameter λ_2 is fixed, (b) parameters λ_1 and λ_3 are fixed.

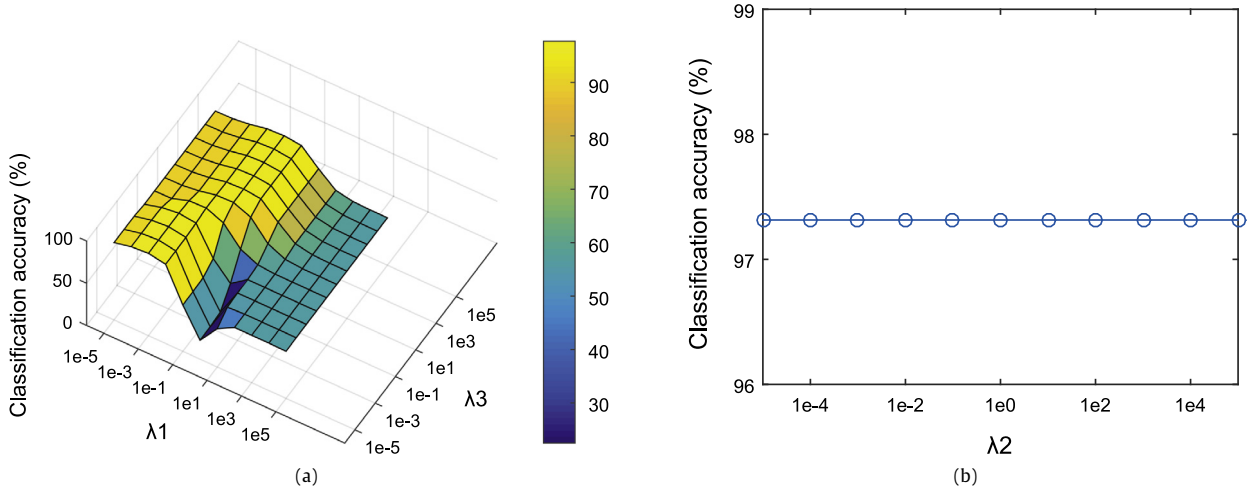


Fig. 8. Relationships of the classification accuracy (%) and different combinations of parameters on the AR dataset, in which 8 samples of each class are selected as the training samples. (a) Parameter λ_2 is fixed, (b) parameters λ_1 and λ_3 are fixed.

Acknowledgments

This paper is partially supported by the National Natural Science Foundation of China (Grant nos. 61370163 and 61772254), Guangdong Province high-level personnel of special support program (Grant no. 2016TX03X164), Key Project of College Youth Natural Science Foundation of Fujian Province (Grant no. JZ160467), Shenzhen Fundamental Research fund (Grant no. JCYJ20160331185006518), Natural Science Foundation of Guangdong Province (Grand no. 2017A030313384), Natural Science Foundation of Heilongjiang Province (Grant no. E2017017), Fujian Provincial Leading Project (Grand no. 2017H0030), Fuzhou Science and Technology Planning Project (Grand no. 2016-S-116), and Program for New Century Excellent Talents in Fujian Province University (Grand no. NCETFJ).

Appendix

Proof of Theorem 1. For simplicity, we use L to denote the optimization problem (9). The KKT conditions of the optimization problem (9) are as follows:

$$F = QX \tag{19}$$

$$\frac{\partial L}{\partial Q} = (QX - G_1)X^T + \lambda_1 Q - CX^T = 0 \tag{20}$$

$$\frac{\partial L}{\partial E} = E - U + \lambda_3 \frac{\partial \left(\|E\|_{2,1}\right)}{\partial E} = 0 \tag{21}$$

$$\frac{\partial L}{\partial F_i} = C_i + \lambda_2 \frac{\partial \left(\|F_i\|_{2,1}\right)}{\partial F_i} = 0, \forall i = 1, \ldots, c \tag{22}$$

where $G_1 = Y + E, U = QX - Y$.

Let $\Gamma = (Q, F, E, C)$ and $\Gamma^+ = \left(Q^+, F^+, E^+, C^+\right)$ be the solution of our optimization problem (9) at the current iteration and next iteration in the solution sequence $\left\{\Gamma^t\right\}_{t=1}^{\infty}$, respectively. From the previous section, the Lagrange multiplier C is updated as follows:

$$C^+ = C + \mu (F - QX). \tag{23}$$

It is obvious that if variable $\left\{C^t\right\}_{t=1}^{\infty}$ converges to a stationary point, *i.e.*, $\left(C^+ - C\right) \to 0$, then we obtain $(F - QX) \to 0$. So, the first KKT condition, *i.e.*, (19), is proved.

For the second KKT condition, we have the following equation from (11):

$$Q^+ - Q = (G_1 + \mu G_2) X^T \left((1+\mu) XX^T + \lambda_1 I\right)^{-1} - Q. \tag{24}$$

From (24), we can obtain:

$$\begin{aligned}&\left(Q^+ - Q\right)\left((1+\mu) XX^T + \lambda_1 I\right)\\ &= (G_1 + \mu G_2) X^T - Q\left((1+\mu) XX^T + \lambda_1 I\right)\\ &= \left[(G_1 - QX) X^T + CX^T - \lambda_1 Q\right] + \mu (F - QX) X^T.\end{aligned} \tag{25}$$

We have proved that when variable $\left\{C^t\right\}_{t=1}^{\infty}$ converges, the equation $F - QX = 0$ holds. From (25), it is obvious that when variable $\left\{Q^t\right\}_{t=1}^{\infty}$ converges to a stationary point, *i.e.*, $\left(Q^+ - Q\right) \to 0$, we can obtain $\left((G_1 - QX) X^T + CX^T - \lambda_1 Q\right) \to 0$. So the second KKT condition, *i.e.*, (20) is proved.

For the third and fourth KKT conditions, *i.e.*, (21) and (22), we can find that these two equations are equivalent to the following equations, respectively:

$$E_{j,:} - U_{j,:} + \lambda_3 \frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = 0, \forall j = 1, \ldots, c \tag{26}$$

$$[C_i]_{j,:} + \lambda_2 \frac{\partial\left(\left\|[F_i]_{j,:}\right\|_2\right)}{\partial [F_i]_{j,:}} = 0, \forall i, j \in [1, c] \tag{27}$$

where $\frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = \begin{cases} \frac{E_{j,:}}{\left\|E_{j,:}\right\|_2}, & E_{j,:} \neq 0 \\ \left\{y \in R^{1\times n} \mid \|y\|_2 \leq 1\right\}, & E_{j,:} = 0 \end{cases}$, $\frac{\partial\left(\left\|[F_i]_{j,:}\right\|_2\right)}{\partial [F_i]_{j,:}} = \begin{cases} \frac{[F_i]_{j,:}}{\left\|[F_i]_{j,:}\right\|_2}, & [F_i]_{j,:} \neq 0 \\ \left\{y \in R^{1\times n_i} \mid \|y\|_2 \leq 1\right\}, & [F_i]_{j,:} = 0 \end{cases}$ (Liu et al., 2009).

From (16), we have the following equation:

$$E_{j,:}^+ - E_{j,:} = \begin{cases} U_{j,:} - \lambda_3 \frac{U_{j,:}}{\left\|U_{j,:}\right\|_2} - E_{j,:}, & if\ \left\|U_{j,:}\right\|_2 > \lambda_3 \\ -E_{j,:}, & otherwise. \end{cases} \tag{28}$$

When variable $\left\{E^t\right\}_{t=1}^{\infty}$ converges, *i.e.*, $E_{j,:}^+ - E_{j,:} = 0, \forall j = 1, \ldots, c$, we obtain:

$$E_{j,:}^+ - E_{j,:} = \begin{cases} U_{j,:} - \lambda_3 \frac{U_{j,:}}{\left\|U_{j,:}\right\|_2} - E_{j,:} = 0, & if\ \left\|U_{j,:}\right\|_2 > \lambda_3 \\ -E_{j,:} = 0, & otherwise. \end{cases} \tag{29}$$

From (29), when $\left\|U_{j,:}\right\|_2 > \lambda_3$, we can deduce $E_{j,:} - U_{j,:} + \lambda_3 \frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = 0$ since $\frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = \frac{E_{j,:}}{\left\|E_{j,:}\right\|_2} = \frac{U_{j,:}}{\left\|U_{j,:}\right\|_2}$; when $\left\|U_{j,:}\right\|_2 \leq \lambda_3$ (otherwise), the equation $E_{j,:} - U_{j,:} + \lambda_3 \frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = 0$ also holds since $\frac{U_{j,:}}{\lambda_3} \in \frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}}$. So for any columns of E, the equation $E_{j,:} - U_{j,:} + \lambda_3 \frac{\partial\left(\left\|E_{j,:}\right\|_2\right)}{\partial E_{j,:}} = 0$ always holds. Thus the third KKT condition, *i.e.*, (21), is proved.

From (14), when variable $\left\{F^t\right\}_{t=1}^{\infty}$ converges, *i.e.*, $[F_i]_{j,:}^+ - [F_i]_{j,:} = 0, \forall i, j \in [1, c]$, we have the following equation:

$$[F_i]_{j,:}^+ - [F_i]_{j,:} = \begin{cases} [H_i]_{j,:} - \frac{\lambda_2 [H_i]_{j,:}}{\mu \left\|[H_i]_{j,:}\right\|_2} - [F_i]_{j,:} = 0, & \left\|[H_i]_{j,:}\right\|_2 > \frac{\lambda_2}{\mu} \\ -[F_i]_{j,:} = 0, & otherwise. \end{cases} \tag{30}$$

Similarly, from (30), when $\left\|[H_i]_{j,:}\right\|_2 > \frac{\lambda_2}{\mu}$, we can deduce that $[C_i]_{j,:} + \lambda_2 \frac{\partial\left(\left\|[F_i]_{j,:}\right\|_2\right)}{\partial [F_i]_{j,:}} = 0$ since $H = QX - C\mu$, $F = QX$, and $\frac{\left\|[H_i]_{j,:}\right\|_2}{[H_i]_{j,:}} = \frac{\left\|[F_i]_{j,:}\right\|_2}{[F_i]_{j,:}}$; when $\left\|[H_i]_{j,:}\right\|_2 \leq \frac{\lambda_2}{\mu}$ (otherwise), the equation $[C_i]_{j,:} + \lambda_2 \frac{\partial\left(\left\|[F_i]_{j,:}\right\|_2\right)}{\partial [F_i]_{j,:}} = 0$ still holds since $\frac{\mu [H_i]_{j,:}}{\lambda_2} \in \frac{\partial\left(\left\|[F_i]_{j,:}\right\|_2\right)}{\partial [F_i]_{j,:}}$ and $[F_i]_{j,:} = [(QX)_i]_{j,:} = [H_i]_{j,:} + [C_i]_{j,:}\mu = 0$. So, (24) always holds in all conditions. Thus, the fourth KKT condition, *i.e.*, (22) is proved as well.

In summary, if the sequence solution $\left\{\Gamma^t\right\}_{t=1}^{\infty}$ of our optimization problem (9) is bounded and satisfies the condition of $\lim_{t\to\infty}\left(\Gamma^{t+1} - \Gamma^t\right) = 0$, the accumulation point of sequence $\left\{\Gamma^t\right\}_{t=1}^{\infty}$ is a Karush–Kuhn–Tucker (KKT) point that satisfies the four KKT conditions mentioned above. Thus we complete the proof.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdiscip. Rev. Comput. Stat.*, *2*(1), 97–106.

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Mach. Learn.*, *73*(3), 243–272.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York, Inc.

Bunea, F., She, Y., & Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 1282–1309.

Cai, X., Ding, C., Nie, F., & Huang, H. (2013). On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1124–1132).

Cai, D., He, X., & Han, J. (2007). Spectral regression: A unified approach for sparse subspace learning. In *IEEE international conference on data mining* (pp. 73–82).

Cai, X., Nie, F., & Huang, H. (2013). Exact top-k feature selection via l 2,0 -norm constraint. In *International joint conference on artificial intelligence* (pp. 1240–1246).

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

Cherkassky, V., & Ma, Y. (2004). Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, *17*(1), 113–126.

De la Torre, F. (2012). A least-squares framework for component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, *34*(6), 1041–1055.

Fang, X., Teng, S., Lai, Z., He, Z., Xie, S., & Wong, W. K. (2017). Robust latent subspace learning for image classification. *IEEE Transactions on Neural Networks & Learning Systems*, *PP*(99), 1–14.

Fang, X., Xu, Y., Li, X., Lai, Z., Teng, S., & Fei, L. (2017). Orthogonal self-guided similarity preserving projection for classification and clustering. *Neural Netw.*, *88*, 1–8.

Fang, X., Xu, Y., Li, X., Lai, Z., Wong, W. K., & Fang, B. (2017). Regularized label relaxation linear regression. *IEEE Transactions on Neural Networks and Learning Systems*.

Feng, Q., Zhou, Y., & Lan, R. (2016). Pairwise linear regression classification for image set retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4865–4872).

Gao, J., Shi, D., & Liu, X. (2007). Significant vector learning to construct sparse kernel regression models. *Neural Netw.*, *20*(7), 791–798.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 643–660.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.

Hong, M., & Luo, Z. Q. (2012). On the linear convergence of the alternating direction method of multipliers. *Math. Program.*, *162*(1–2), 1–35.

Jiang, Z., Lin, Z., & Davis, L. S. (2013). Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(11), 2651–2664.

Kim, D., & Gales, M. (2011). Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 315–325.

Kim, E., Lee, M., & Oh, S. (2015). Elastic-net regularization of singular values for robust subspace learning. In *IEEE conference on computer vision and pattern recognition* (pp. 915–923).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition, Vol. 2* (pp. 2169–2178). IEEE.

Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., & Hua, G. (2016). Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis* (pp. 189–248). Springer.

Li, X., Chen, M., Nie, F., & Wang, Q. (2017). Locality adaptive discriminant analysis. In *Twenty-sixth international joint conference on artificial intelligence* (pp. 2201–2207).

Li, Y., & Ngom, A. (2013). Nonnegative least-squares methods for the classification of high-dimensional biological data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, *10*(2), 447–456.

Lin, Z., Chen, M., & Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.

Lin, Z., Liu, R., & Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems* (pp. 612–620).

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient l 2, 1 -norm minimization. In *Conference on uncertainty in artificial intelligence* (pp. 339–348).

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(1), 171–184.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, *9*(2605), 2579–2605.

Martinez, A. M. (1998). The ar face database. CVC technical report.

Naseem, I., Togneri, R., & Bennamoun, M. (2010). Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(11), 2106–2112.

Nene, S. A., Nayar, S. K., & Murase, H. et al., (1996). Columbia object image library (coil-20).

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, *90*(432), 1257–1270.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 1346–1370.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, *73*(3), 273–282.

Wang, J. J.-Y., & Gao, X. (2015). Max–min distance nonnegative matrix factorization. *Neural Netw.*, *61*, 75–84.

Wang, Q., Meng, Z., & Li, X. (2017). Locality adaptive discriminant analysis for spectral-spatial classification of hyperspectral images. *IEEE Geoscience & Remote Sensing Letters*, *14*(11), 2077–2081.

Wang, L., & Pan, C. (2017). Groupwise retargeted least-squares regression. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, S.-J., Yang, J., Sun, M.-F., Peng, X.-J., Sun, M.-M., & Zhou, C.-G. (2012). Sparse tensor discriminant color space for face verification. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(6), 876–888.

Wang, L., Zhang, X.-Y., & Pan, C. (2016). Msdlsr: Margin scalable discriminative least squares regression for multicategory classification. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(12), 2711–2717.

Wen, J., Fang, X., Cui, J., Fei, L., Yan, K., Chen, Y., et al. (2018). Robust sparse linear discriminant analysis. *IEEE Trans. Circuits Syst. Video Technol.*.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 210–227.

Xiang, S., Nie, F., Meng, G., Pan, C., & Zhang, C. (2012). Discriminative least squares regression for multiclass classification and feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(11), 1738–1754.

Xiang, S., Zhu, Y., Shen, X., & Ye, J. (2012). Optimal exact least squares rank minimization. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 480–488). ACM.

Xu, Y., Fang, X., Zhu, Q., Chen, Y., You, J., & Liu, H. (2014). Modified minimum squared error algorithm for robust classification and face recognition experiments. *Neurocomputing*, *135*(C), 253–261.

Xue, H., Chen, S., & Yang, Q. (2009). Discriminatively regularized least-squares classification. *Pattern Recognit.*, *42*(1), 93–104.

Yang, J., & Yuan, X. (2013). Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, *82*(281), 301–329.

Ye, J. (2007). Least squares linear discriminant analysis. In *Proceedings of the 24th international conference on machine learning* (pp. 1087–1093). ACM.

Zhang, Y. (2010). An alternating direction algorithm for nonnegative matrix factorization. Technical preprint.

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Trans. Image Process.*, *26*(3), 1466–1481.

Zhang, Z., Shao, L., Xu, Y., Liu, L., & Yang, J. (2017). Marginal representation learning with graph structure self-adaptation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, X.-Y., Wang, L., Xiang, S., & Liu, C.-L. (2015). Retargeted least squares regression algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(9), 2206–2213.

Zhang, L., Yang, M., Feng, X., Ma, Y., & Zhang, D. (2012). Collaborative representation based classification for face recognition. arXiv preprint arXiv:1204.2358.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.