# Discriminative Elastic-Net Regularized Linear Regression

Zheng Zhang, Zhihui Lai, Yong Xu, *Senior Member, IEEE,* Ling Shao, *Senior Member, IEEE,*
Jian Wu, Guosen Xie

***Abstract*—In this paper, we aim at learning compact and discriminative linear regression models. Linear regression has been widely used in different problems. However, most of the existing linear regression methods exploit the conventional zero-one matrix as the regression targets, which greatly narrows the flexibility of the regression model. Another major limitation of theses methods is that the learned projection matrix fails to precisely project the image features to the target space due to their weak discriminative capability. To this end, we present an elastic-net regularized linear regression (ENLR) framework, and develop two robust linear regression models which possess the following special characteristics. First, our methods exploit two particular strategies to enlarge the margins of different classes by relaxing the strict binary targets into a more feasible variable matrix. Second, a robust elastic-net regularization of singular values is introduced to enhance the compactness and effectiveness of the learned projection matrix. Third, the resulting optimization problem of ENLR has a closed-form solution in each iteration, which can be solved efficiently. Finally, rather than directly exploiting the projection matrix for recognition, our methods employ the transformed features as the new discriminate representations to make final image classification. Compared with the traditional linear regression model and some of its variants, our method is much more accurate in image classification. Extensive experiments conducted on publicly available datasets well demonstrate that the proposed framework can outperform the state-of-the-art methods. The MATLAB codes of our methods can be available at http://www.yongxu.org/lunwen.html.**

***Index Terms*—Elastic-Net regularization, discriminative methods, linear regression, image classification**

## I. Introduction

DISCRIMINATIVE methods (e.g., regression models) have a good reputation in both theoretical research and practical applications, and also have been extensively applied to solving many computer vision problems [1], [2]. Different from the probabilistic models, discriminative methods typically project image features to some continuous or discrete targets, and then exploit the projection matrix to make image classification or regression [3], [4]. In addition, discriminative methods can achieve impressive performance when constructing robust projection matrix and providing sufficient training samples [5], [6]. However, the problem of robust discriminative learning has not been exhaustively explored and perfectly solved.

Least square regression (LSR) is a typical and fundamental technique in statistics theory. Due to its mathematically tractable and efficient solution as well as simple yet effective formulation, LSR has been widely used in many other applications such as computer vision and pattern recognition [7]. Many variations have been proposed to enhance the performance of the conventional LSR, such as partial LSR [8], weighted LSR [9], and nonnegative least squares [10]. Moreover, extensive discriminative LSR methods have been developed to improve the robustness and effectiveness of the existing regression approaches. For example, Xiang et al. [7] designed a general framework of discriminative least square regression (DLSR) by introducing the $\varepsilon$-dragging technique for image classification and feature selection, and Zhang et al. [11] introduced a method of retargeted LSR by learning transformed regression. A unified least square framework [12] is constructed to formulate many component analysis methods and generate their regularized and kernel extensions. Thus, LSR model has become a popular technique and also has been widely adopted to deal with recognition and classification tasks [13].

Another important and fundamental variant of LSR is the problem of least absolute shrinkage and selection operator, i.e. LASSO [14], or sparse representation problem [6], [15]. Sparse representation based classification (SRC) method [15] has been extensively applied to addressing the face recognition problem, and the performance is very impressive. Subsequently, numerous representation based classification methods have been proposed to improve its effectiveness, robustness and efficiency of face recognition [16]. For example, linear regression based classification (LRC) [17] method exploits the linear combination of each class of training samples to represent the test sample, and then classifies the test sample to the class which leads to the minimum representation residual. Collaborative representation based classification method (CRC) [18] introduces the $l_2$-norm regularization instead of the $l_1$-norm regularization for efficient face recognition. Specifically, literature [18] demonstrates that SRC theoretically is a special case of the collaborative representation method, and the computational efficiency of CRC is dramatically higher than SRC

Manuscript received *** **, 2016; revised *** **, 2016; accepted *** **, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61370163, and Grant 61233011. (*Corresponding author: Yong Xu.*)

Z. Zhang, Y. Xu and J. Wu are with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail:darrenzz219@gmail.com, yongxu@ymail.com, wujianhitsz@163.com).

Z. Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518055, China, and also with the Institute of Textiles & Clothing, Hong Kong Polytechnic University, Hong Kong (e-mail: lai_zhi_hui@163.com).

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (e-mail: ling.shao@ieee.org).

G. Xie is with the Department of Information Engineering, Henan University of Science and Technology (e-mail: gsxiehm@gmail.com).

without sacrificing much classification accuracies. Moreover, locality-constrained linear coding (LLC) enforces the locality constraints to perform a local embedding of the descriptors [19], [20]. In addition, the representation based technique has been introduced to a wide range of applications.

The low-rank minimization problem has attracted a lot of attention due to its effectiveness on data representation [21]. It is worth noting that robust PCA (RPCA) [22] is one of the most popular methods based on the low-rank minimization. Providing that data are lying in a single subspace, RPCA decomposes the observed data into two components, the low-rank uncorrupted data term and sparse noise term. The low-rank regression model [23] has been studied because of the apparent advantage of the low-rank characteristics [22], [24], [25]. Based on the traditional low-rank linear regression (L-RLR) model, Cai et al. [23] proposed two low-rank regression models, i.e. low-rank ridge regression (LRRR) and sparse low-rank regression (SLRR) methods. Specifically, these three low-rank regression models are equivalent to linear discriminant analysis based regressions [23]. Furthermore, all of them are based on the nature of the low-rank minimization, which can capture the underlying structure of data correlation patterns [22], [24]. Latent low-rank representation (LatLRR) [26] explores the unobserved hidden information of data, and can robustly extract salient features from noise or corrupted data. Subsequently, many variations of the low-rank minimization have been applied to solve different problems [3], [27]–[29]. For example, Li and Fu [3] proposed a supervised regularization-based robust subspace learning method by jointly removing noise term with low-rank constraint and learning a discriminate subspace from the clean data. Wei et al. [27] designed a method of low-rank matrix recovery method by embedding the structure incoherence (LRSI) information for robust face recognition. Li et al. [28] constructed a classwise block-diagonal structure (CBDS) dictionary by imposing the class-wise discriminative structure regularization term to make the samples from different classes be reconstructed with different bases. Benefiting from recent advances on low-rank minimization, a framework of robust regression model [2] was proposed to solve several computer vision problems.

Nonetheless, most existing regression methods in the learning phase only focus on directly projecting the original visual features to conventional zero-one target matrix, which provides too little freedom to fit the strict binary label matrix. Moreover, the projection matrices learned by these methods fail to precisely project the image features to the target fields due to its weak discriminative capability. It is notable that a robust and discriminative regression method should equip with three-fold characteristics, i.e. compact projection matrix, discriminative regression targets and robust to errors in the data. Given these deficiencies, this paper develops a novel elastic-net regularized linear regression (ENLR) framework, and two robust ENLR methods, i.e. discriminative ENLR and marginalized ENLR, are proposed to construct a robust and compact regression model for multi-category image classification. More specifically, the elastic-net regularization term is accumulated to learn a more compact projection matrix, and at the same time, enlarging the margins of different classes is significant and beneficial to the classification tasks. Based on the $\varepsilon$-dragging technique, the discriminative regression targets are further formulated to better fit regression tasks. Moreover, marginalized regression targets are learned directly from data by enforcing a strong constraint on the learned targets between the true and false classes. Furthermore, instead of directly exploiting the projection matrix for classification, the data points under the simple linear transformation using the learned projection matrix are employed to final classification such that the transformed data is more discriminative and robust to errors. In addition, the low-rank model always suffers from heavy computational burden due to singular value decomposition procedure. To efficiently solve it, ENLR introduces an alternative definition of the nuclear-norm with a strong convexity strategy such that our method can be scalable to large data sets. To the best of our knowledge, this is for the first time to unify the elastic-net regularization of singular values and learning discriminative regression targets into one framework, which is a very simple but extraordinarily effective method for image classification. The effectiveness of the ENLR framework is demonstrated on different classification tasks. Therefore, the main contributions of this paper are summarized as follows.

(1) In this paper, the elastic-net regularization of singular values and constructing distinctive regression targets are for the first time integrated into one unified discriminative linear regression framework. The underlying characteristics of the elastic-net regularization of singular values are explicitly uncovered and analyzed such that the elastic-net theory is extended to the elastic-net regularization of singular values.

(2) By virtue of enlarging the margins of different classes, we propose two robust elastic-net regularized linear regression methods as well as the corresponding alternative efficient methods. Specifically, the discriminative ENLR (DENLR) method interpolates the $\varepsilon$-dragging technique into the ENLR framework, and a more flexible marginalized ENLR (MENLR) method is developed by directly learning the marginal regression targets from data, in which a strong marginalized constraint is enforced to make the learned targets distinguishable.

(3) Two efficient algorithms are proposed to solve the resulting optimization problems, and theoretical and experimental analysis are conducted to prove the convexity and convergence of the designed optimization algorithms. Additionally, the theoretical relationships between the proposed ENLR framework and the prevailing LSR models are revealed.

The rest of this paper is organized as follows. We briefly introduce some related works in Section II. Then, we describe the proposed ENLR framework and theoretical analysis in Section III, and optimization algorithm is presented in Section IV. Extensive experiments are reported in Section V. Finally, the conclusion remarks and our future work are summarized in Section VI.

## II. Related Work

### A. Notation

The matrix is denoted by bold uppercase letters, e.g. $\boldsymbol{X}$, and the $i$-th row and $j$-th column element of matrix $\boldsymbol{X}$ is denoted as $\boldsymbol{X_{ij}}$. Column vectors are denoted by bold lower letters, e.g.

$\boldsymbol{x}$. $\|\boldsymbol{X}\|_F^2 = tr(\boldsymbol{X}^T\boldsymbol{X}) = tr(\boldsymbol{X}\boldsymbol{X}^T)$ designates the Frobenius norm of matrix $\boldsymbol{X}$, where $tr(\bullet)$ is the trace operator. $\|\boldsymbol{X}\|_*$ is the nuclear norm of the matrix $\boldsymbol{X}$, i.e. $\|\boldsymbol{X}\|_* = \sum_i |\sigma_i|$ where $\sigma_i$ is the $i$-th singular value of matrix $\boldsymbol{X}$. $\boldsymbol{X}^T$ denotes the transposed matrix of $\boldsymbol{X}$ and $\boldsymbol{I}$ denotes an identity matrix.

### B. Linear regression model

Linear and non-linear regression have been widely applied to many computer vision problems, such as classification [2], [7], [11]. Standard linear regression model for classification is to learn a linear projection matrix in the training stage, and uses it to project the observed image features $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}] \in \Re^{d\times n}$ approximate to the target matrix $\boldsymbol{Y} = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_n}]^{\boldsymbol{T}} \in \Re^{n\times c}$ by minimizing

$$\min_{\boldsymbol{D}} \|\boldsymbol{X^T D} - \boldsymbol{Y}\|_F^2, \tag{1}$$

where $\boldsymbol{X}$ is the given data set, $\boldsymbol{D} \in \Re^{d\times c}$ is the learned projection matrix, and $\boldsymbol{Y}$ is the corresponding binary class indicator matrix. Specifically, $\boldsymbol{y_i} \in \Re^c$ is the label vector of the $i$-th sample $\boldsymbol{x_i}$, and $n$ and $c$ are the number of samples and classes, respectively. A more popular-used regularized linear regression model is formulated by addressing the following optimization problem

$$\min_{\boldsymbol{D},\boldsymbol{b}} \|\boldsymbol{X^T D} + \boldsymbol{e_n b^T} - \boldsymbol{Y}\|_F^2 + \lambda\|\boldsymbol{D}\|_F^2. \tag{2}$$

The general steps of the linear regression model for image classification task are as follows. In the training stage, we learn the projection matrix $\boldsymbol{D}$, and any test point is estimated by $\boldsymbol{D^T x}_{te}$ in the test step.

### C. Low-rank linear regression model

The low-rank linear regression (LRLR) model [23] is a modified version of the standard linear regression model (1). Compared to the conventional linear regression model, a more compact low-rank projection is learned by enforcing the rank minimization constraint to explore the underlying correlation structures between classes [23], and the objective function of LRLR is formulated as

$$\min_{\boldsymbol{D}} \|\boldsymbol{X^T D} - \boldsymbol{Y}\|_F^2 + \lambda rank(\boldsymbol{D}). \tag{3}$$

Because of the discrete property of the rank function, which is a non-convex non-smooth problem, a tractable optimization problem is reformulated by replacing the rank function with the nuclear norm regularization [32], i.e.

$$\min_{\boldsymbol{D}} \|\boldsymbol{X^T D} - \boldsymbol{Y}\|_F^2 + \lambda\|\boldsymbol{D}\|_*. \tag{4}$$

The nuclear norm regularization can effectively discover the hidden structures between classes such that the learned low-rank projection matrix is more compact and discriminative than the traditional projection matrix. The low-rank linear regression model is demonstrated to be equivalent to the linear discriminant analysis based regression [23]. It is worth noting that the low-rank linear regression models can provide better data mining results in comparison with the existing full-rank linear regression models [23].

## III. The proposed ENLR framework

In this section, we focus on learning a compact and discriminative regression model for robust multi-category image classification. For linear regression model, compact projection matrix and discriminative regression targets are both important. We introduce an elastic-net regularization of singular values term to formulate robust projection matrix, and the enlarged slack regression targets are constructed to improve its discriminant. Therefore, an elastic-net regularized linear regression (ENLR) framework and two discriminative linear regression methods are proposed for image classification.

### A. A general framework of elastic-net regularized linear regression model

To learn a compact and discriminative projection matrix, a general framework of elastic-net regularization based linear regression model is formulated as

$$\min_{\boldsymbol{D}} \phi(\boldsymbol{D}) + \lambda_1\|\boldsymbol{D}\|_* + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2, \tag{5}$$

where $\lambda_1$ and $\lambda_2$ are the regularization parameters for balancing respective terms. The most straightforward regression loss function is $\phi(\boldsymbol{D}) = \|\boldsymbol{X^T D} - \boldsymbol{Y}\|_F^2$. For the above objective function (5), we have the following proposition.

***Proposition 1***: *Objective function (5) is a robust regression problem with an elastic-net regularization of singular values.*

The singular value decomposition (SVD) factorizes the linear transformation matrix $\boldsymbol{D}$ into

$$\boldsymbol{D} = \boldsymbol{U\Sigma V}^T = \sum_{i=1}^{r} \boldsymbol{u}_i\boldsymbol{\sigma}_i\boldsymbol{v}_i^T, \tag{6}$$

where $r = \min\{c, d\}$ is the rank of $\boldsymbol{D}$, $\boldsymbol{u}_i \in \Re^d$ and $\boldsymbol{v}_i \in \Re^c$ are respectively the left and right singular vectors of $\boldsymbol{D}$, and $\boldsymbol{\sigma}_i$ is the $i$-th singular value of matrix $\boldsymbol{D}$.

It is notable that the nuclear norm of matrix $\boldsymbol{D}$ can be interpreted as a sum of the singular values, i.e. $\|\boldsymbol{D}\|_* = \sum_i^r |\boldsymbol{\sigma}_i|$, and the Frobenius norm of matrix $\boldsymbol{D}$ is $\|\boldsymbol{D}\|_F^2 = tr(\boldsymbol{DD^T}) = tr(\boldsymbol{U\Sigma V^T V\Sigma U^T}) = tr(\boldsymbol{\Sigma}^2) = \sum_i |\boldsymbol{\sigma}_i|^2$. Thus, by integrating the nuclear norm and the Frobenius norm penalties of matrix $\boldsymbol{D}$, we have the elastic-net regularization of singular values term, i.e. $\|\boldsymbol{D}\|_* + \|\boldsymbol{D}\|_F^2 = \sum_i |\boldsymbol{\sigma}_i| + \sum_i |\boldsymbol{\sigma}_i|^2$.

Typically, the large singular values always highlight the components where the fundamental information lies. It is interesting to note that the singular values can directly reflect the importance of underlying components of data. For example, smaller singular values always come from the redundant or noise-contaminated components when the data contains redundant information or noise. It seems natural to use the measurement of singular values to analyze data. Based on these observations, the elastic-net regularization of singular values provides an advisable approach to removing the redundant components based on the following proposition.

***Proposition 2***: *The elastic-net regularization of singular values can effectively enable automatic grouped variable s-election of principal components and continuous shrinkage of redundant components.*

Based on Proposition 1, we know that the elastic-net regularization of singular values is composed of the $l_2$-norm and $l_1$-norm regularization of singular values. It is known that the $l_2$-norm regularization of singular values tends to shrink a variable towards zero but generally keeps all the components in the model, which may lead to redundant information in predictors. However, this deficiency of the $l_2$-norm regularization fortunately generates the grouping characteristic in the model-fitting procedure. On the contrary, the $l_1$-norm regularization of singular values can produce automatic selection of principal information and continuous shrinkage of redundant information simultaneously [14]. However, one significant limitation of the $l_1$-norm regularization is that when the correlations among a group of variables are very high, it tends to select only one variable from the group, but neglects the remaining ones, which may lead to sub-optimal results. The feasible way of overcoming this deficiency is to regard the highly-correlated group as a whole to make variable selection, i.e. grouped variable selection. Therefore, it is reasonable to mix the $l_2$-norm and $l_1$-norm regularization of singular values, yielding the elastic-net regularization of singular values, which can effectively enable automatic grouped variable selection of principal components and continuous elimination of dependencies and redundancies in data. In this way, the proposed ENLR framework is a succinct and stable linear regression formulation.

Furthermore, given optimal regression $\boldsymbol{D}$, we will project $\boldsymbol{x}$ to the target space (e.g. label space) by

$$\boldsymbol{D}^T\boldsymbol{x} = \sum_{i=1}^{\acute{r}} \boldsymbol{\sigma}_i(\boldsymbol{u}_i^T\boldsymbol{x})\boldsymbol{v}_i, \tag{7}$$

where $\acute{r}$ is the number of selected singular values. Herein the target space can be viewed as a weighted linear combination of target-component vectors $\{\boldsymbol{v}_i\}_{i=1}^{\acute{r}}$, and the $i$-th weight is composed of two terms, i.e. the $i$-th singular value $\boldsymbol{\sigma}_i$, and transformed feature value $\boldsymbol{u}_i^T\boldsymbol{x}$, which is determined by the feature-component vectors $\{\boldsymbol{u}_i\}_{i=1}^{\acute{r}}$. We can see that the optimized selection of singular values can generate the optimal feature-component and target-component vectors such that the importance of the feature correlations and target correlations is simultaneously uncovered.

Moreover, the elastic-net regularization of features has shown its great superiorities in comparison with the ridge regularization [18] and LASSO [14] in many applications such as feature selection [33] and matrix factorization [30]. However, the elastic-net regularization of features can not effectively capture and mine the subtle information from data, whereas exploiting the elastic-net regularization of singular values can attain a more compact and distinctive projection matrix, which improves the performances of linear regression models. Based on the elastic-net regularized linear regression framework in Eq. (5), two robust elastic-net regularized linear regression methods are proposed, i.e. discriminative elastic-net regularized linear regression (DENLR) and marginalized elastic-net regularized linear regression (MENLR).

### *B. Discriminative Elastic-net Regularized Linear Regression*

To enhance the discriminative capability of regression results, the $\varepsilon$-dragging technique is introduced to transform the strict zero-one regression targets into the disjunctive but discriminative ones such that the regression model is more robust. Due to the weak separability of the strict binary regression targets in (1), the $\varepsilon$-dragging technique enforces the regression targets of different classes moving along mutual opposite directions such that the margins between different classes are enlarged and more discriminative regression targets are achieved.

We take an example to introduce the rationale of the $\varepsilon$-dragging technique and demonstrate that the reformulated regression targets are more discriminative than $\boldsymbol{Y}$. Let $x_1$, $x_2$, $x_3$ be three training samples, which are respectively from the third, first and second classes, and then the corresponding binary-class label matrix is defined as $\boldsymbol{Y} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \in \Re^{3\times3}$. However, we expect that the strict binary regression target matrix can be relaxed into some soft extent to fit the data. To this end, a slack variable matrix is constructed by using the $\varepsilon$-dragging technique, which drags these binary outputs far way along different directions. More specifically, if we take the above three samples as an example, the regression target matrix is defined as $\tilde{\boldsymbol{Y}} = \begin{bmatrix} -m_{11} & -m_{12} & 1+m_{13} \\ 1+m_{21} & -m_{22} & -m_{23} \\ -m_{31} & 1+m_{32} & -m_{33} \end{bmatrix}, \ s.t. \ m_{ij} \geq 0.$ Apparently, the distance of each sample in matrix $\boldsymbol{Y}$ is $\sqrt{2}$, while the distance between each sample in $\tilde{\boldsymbol{Y}}$ is bigger than or equal to $\sqrt{2}$ owing to the nonnegative constraint of parameter $m$s. For example, the first and second sample in $\tilde{\boldsymbol{Y}}$ is $\sqrt{(-m_{11}-1-m_{21})^2+(-m_{12}+m_{22})^2+(1+m_{13}+m_{23})}$ $\geq \sqrt{2}$. It is easy to see that the margins of the different classes are enlarged.

By introducing the $\varepsilon$-dragging technique, a discriminative elastic-net regularized linear regression (DENLR) model is developed, and its objective function is formulated as

$$\min_{\boldsymbol{D}} \psi(\boldsymbol{D}) + \lambda_1\|\boldsymbol{D}\|_* + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2, \tag{8}$$

where $\psi(\boldsymbol{D}) = \|\boldsymbol{X^T D} - \tilde{\boldsymbol{Y}}\|_F^2$ and $\tilde{\boldsymbol{Y}}$ is the relaxed regression target matrix.

To obtain an optimal $\tilde{\boldsymbol{Y}}$, an elaborate strategy is devised as follows. Let $\boldsymbol{E}$ be a constant matrix, and the $i$-th row and $j$-th column entry is defined as

$$\boldsymbol{E}_{ij} = \begin{cases} +1 & if \quad \boldsymbol{Y}_{ij} = 1 \\ -1 & if \quad \boldsymbol{Y}_{ij} = 0, \end{cases} \tag{9}$$

and then, we have $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{E} \odot \boldsymbol{M}$, where $\boldsymbol{M} \in \Re^{n\times c}$ is a learned nonnegative matrix. Thus, the proposed DENLR model (8) is rewritten as the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{M}} & \|\boldsymbol{X^T D} - (\boldsymbol{Y} + \boldsymbol{E} \odot \boldsymbol{M})\|_F^2 + \lambda_1\|\boldsymbol{D}\|_* \\ & + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \quad s.t. \quad \boldsymbol{M} \geq 0. \end{aligned} \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2651396, IEEE Transactions on Image Processing

### C. Marginalized Elastic-net Regularized Linear Regression

From problem (10), we can see that the relaxed target space of DENLR is subject to the bound that the regression results should be larger than 1 for true classes and smaller than 0 for false classes. However, this target space is still based on the zero-one label matrix $\boldsymbol{Y}$, which greatly confines the flexibility of the regression model. To this end, we propose to directly learn the regression targets from data, and a marginalized constraint is enforced to make the learned targets distinguishable. We consider the following marginalized elastic-net regularized linear regression (MENLR) problem:

$$\begin{aligned} &\min_{\boldsymbol{D},\boldsymbol{R}} \|\boldsymbol{X}^T\boldsymbol{D}-\boldsymbol{R}\|_F^2+\lambda_1\|\boldsymbol{D}\|_*+\frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \\ &\quad s.t.\ \ \boldsymbol{r}_{iy_i}-\max_{j\neq y_i}\boldsymbol{r}_{ij}\geq C, i=1,\cdots,n, \end{aligned} \tag{11}$$

where $\boldsymbol{R}=[\boldsymbol{r}_1,\cdots,\boldsymbol{r}_n]^T\in\Re^{n\times c}$ is the learned regression targets, and $C$ is a constant. Herein $y_i$ denotes the index of the true class for the $i$-th sample $\boldsymbol{x}_i$. That is, if the $i$-th sample is from the $m$-th class (i.e. $y_i$=$m$), the value of the $m$-th element of the learned target vector $\boldsymbol{r}_i$, i.e. $\boldsymbol{r}_{im}$, should be bigger than the rest of the elements by a fixed margin of $C$. Similar to SVM [43], we simply set the marginal value between the true and the false classes to 1, i.e. $C=1$. Apparently, the marginalized constraint makes the learned regression targets between the true and false classes separable by a fixed distance such that the proposed MENLR is more flexible and discriminative.

### D. Efficient ENLR

For large-scale image classification tasks, the computation complexity of the designed model should be seriously taken into consideration. Thus, the following theorem [34] can make our models appropriate for practical applications.

**Theorem 1.** *For any matrix $\boldsymbol{D}$, we have the following equation:*

$$\|\boldsymbol{D}\|_*=\min_{\boldsymbol{D}=\boldsymbol{A}\boldsymbol{B}}\|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F=\min_{\boldsymbol{D}=\boldsymbol{A}\boldsymbol{B}}\frac{1}{2}(\|\boldsymbol{A}\|_F^2+\|\boldsymbol{B}\|_F^2). \tag{12}$$

***Proof.*** For better flow of the paper, we move the proof of Theorem 1 to Appendix A. □

Based on the Theorem 1, we make an equivalent representation of DENLR as

$$\begin{aligned} &\min_{\boldsymbol{D},\boldsymbol{M},\boldsymbol{A},\boldsymbol{B}}\|\boldsymbol{X^T}\boldsymbol{D}-(\boldsymbol{Y}+\boldsymbol{E}\odot\boldsymbol{M})\|_F^2+\frac{\lambda_1}{2}(\|\boldsymbol{A}\|_F^2 \\ &\quad+\|\boldsymbol{B}\|_F^2)+\frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2\ \ s.t.\ \ \boldsymbol{D}=\boldsymbol{A}\boldsymbol{B},\ \boldsymbol{M}\geq 0, \end{aligned} \tag{13}$$

and MENLR is rewritten as

$$\begin{aligned} &\min_{\boldsymbol{D},\boldsymbol{R}}\|\boldsymbol{X}^T\boldsymbol{D}-\boldsymbol{R}\|_F^2+\frac{\lambda_1}{2}(\|\boldsymbol{A}\|_F^2+\|\boldsymbol{B}\|_F^2) \\ &+\frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2\ \ s.t.\ \ \boldsymbol{D}=\boldsymbol{A}\boldsymbol{B},\boldsymbol{r}_{iy_i}-\max_{j\neq y_i}\boldsymbol{r}_{ij}\geq C. \end{aligned} \tag{14}$$

## IV. Optimization and Algorithm Analysis

In this section, we present two efficient and effective optimization algorithms to solve (13) and (14). In general, the two optimization problems with the low-rank constraint $\boldsymbol{D}=\boldsymbol{A}\boldsymbol{B}$ are both non-convex and non-smooth problems. Fortunately, ALM provides a preferable way to find minimum points of such optimization problems with equality constraints as (13) and (14). To obtain efficient solutions, we utilize the ALM strategy to optimize the resulting problems in an alternative minimization manner, i.e. minimizing the loss with respect to one variable when fixing the rest variables [35].

### A. Optimization of DENLR

The ALM strategy solves the problems by alternatively minimizing the augmented Lagrangian of the original problems and maximizing the dual problems. Here the augmented Lagrangian function of problem (13) is

$$\begin{aligned} &\mathcal{L}(\boldsymbol{D},\boldsymbol{M},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}_1)=\|\boldsymbol{X}^T\boldsymbol{D}-(\boldsymbol{Y}+\boldsymbol{E}\odot\boldsymbol{M})\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \\ &\quad+\frac{\lambda_1}{2}(\|\boldsymbol{A}\|_F^2+\|\boldsymbol{B}\|_F^2)+\langle\boldsymbol{C}_1,\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\rangle+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\|_F^2, \end{aligned} \tag{15}$$

where $\langle\boldsymbol{P},\boldsymbol{Q}\rangle$=$tr(\boldsymbol{P^T}\boldsymbol{Q})$, $\boldsymbol{C}_1$ is a Lagrange multiplier and $\mu>0$ is a penalty parameter. The minimum points of $\mathcal{L}$ with respect to primal variables can be found via the block coordinate descend (BCD) method. The augmented Lagrangian is minimized along one coordinate direction at each iteration. We expand this procedure in more details.

*Updating $\boldsymbol{A}$*: Fix the other variables and update $\boldsymbol{A}$ by solving the following problem.

$$\begin{aligned} \boldsymbol{A}^+ &=\arg\min_{\boldsymbol{A}}\frac{\lambda_1}{2}\|\boldsymbol{A}\|_F^2+<\boldsymbol{C}_1,\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}>+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\|_F^2 \\ &=\arg\min_{\boldsymbol{A}}\frac{\lambda_1}{2}\|\boldsymbol{A}\|_F^2+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}+\frac{\boldsymbol{C}_1}{\mu}\|_F^2, \end{aligned} \tag{16}$$

where the rest terms irrelevant to $\boldsymbol{A}$ in $\mathcal{L}$ are viewed as constants and ignored in the loss since they make no differences in this particular procedure. The resulting problem (16) is a typical regularized least square problem, hence its solution is easily obtained as

$$\boldsymbol{A}^+=(\boldsymbol{C}_1+\mu\boldsymbol{D})\boldsymbol{B}^T(\lambda_1\boldsymbol{I}+\mu\boldsymbol{B}\boldsymbol{B}^T)^{-1}. \tag{17}$$

*Updating $\boldsymbol{B}$*: The variable $\boldsymbol{B}$ plays a symmetric role to that of $\boldsymbol{A}$ in $\mathcal{L}$, hence the updating of $\boldsymbol{B}$ is performed in a symmetric way:

$$\begin{aligned} \boldsymbol{B}^+ &=\arg\min_{\boldsymbol{B}}\frac{\lambda_1}{2}\|\boldsymbol{B}\|_F^2+\langle\boldsymbol{C}_1,\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\rangle+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\|_F^2 \\ &=\arg\min_{\boldsymbol{B}}\frac{\lambda_1}{2}\|\boldsymbol{B}\|_F^2+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}+\frac{\boldsymbol{C}_1}{\mu}\|_F^2. \end{aligned} \tag{18}$$

Similarly,

$$\boldsymbol{B}^+=(\lambda_1\boldsymbol{I}+\mu\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{C}_1+\mu\boldsymbol{D}). \tag{19}$$

**Algorithm 1.** Optimization of DENLR by Exact ALM

```
Require: Feature Matrix $\boldsymbol{X}$; Label Matrix $\boldsymbol{Y}$; Constant matrix $\boldsymbol{E}$; Parameters $\lambda_1$, $\lambda_2$.
Initialization: $\boldsymbol{M} = \boldsymbol{0}$, $\boldsymbol{D} \in \Re^{d\times c}$, $\boldsymbol{A} \in \Re^{d\times r}$, $\boldsymbol{B} \in \Re^{r\times c}$, $\boldsymbol{C}_1 \in \Re^{d\times c}$, $\lambda_1 > 0$, $\lambda_2 > 0$, $\mu > 0$.
While not converged do
  While not converged do
    Step 1. Update $\boldsymbol{A}$ by solving problem (16);
    Step 2. Update $\boldsymbol{B}$ by solving problem (18);
    Step 3. Update $\boldsymbol{D}$ by solving problem (20);
    Step 4. Update $\boldsymbol{M}$ by solving problem (22);
  End While
  Step 5. Update the Lagrange multipliers $\boldsymbol{C}_1$ by
    $\boldsymbol{C}_1 = \boldsymbol{C}_1 + \mu(\boldsymbol{D} - \boldsymbol{AB})$.
End While
Output: Projection matrix $\hat{\boldsymbol{D}}$
```

*Updating* $\boldsymbol{D}$: Fix the other variables and update $\boldsymbol{D}$ by solving the following problem.

$$\begin{aligned}\boldsymbol{D}^+ &= \arg\min_{\boldsymbol{D}} \|\boldsymbol{X^T D} - \boldsymbol{S}\|_F^2 + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \\ &\qquad + \langle \boldsymbol{C}_1, \boldsymbol{D} - \boldsymbol{AB}\rangle + \frac{\mu}{2}\|\boldsymbol{D} - \boldsymbol{AB}\|_F^2 \\ &= \arg\min_{\boldsymbol{D}} \|\boldsymbol{X^T D} - \boldsymbol{S}\|_F^2 + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \\ &\qquad + \frac{\mu}{2}\|\boldsymbol{D} - \boldsymbol{AB} + \frac{\boldsymbol{C}_1}{\mu}\|_F^2,\end{aligned} \tag{20}$$

where $\boldsymbol{S} = \boldsymbol{Y} + \boldsymbol{E} \odot \boldsymbol{M}$. By setting the derivative $\frac{\partial \mathcal{L}}{\partial \boldsymbol{D}} = \boldsymbol{0}$, we can infer the optimal solution of $\boldsymbol{D}$ as

$$\boldsymbol{D}^+ = (2\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 \boldsymbol{I} + \mu \boldsymbol{I})^{-1}(2\boldsymbol{XS} + \mu\boldsymbol{AB} - \boldsymbol{C}_1). \tag{21}$$

*Updating* $\boldsymbol{M}$: Fix the other variables and update $\boldsymbol{M}$ by solving the following problem.

$$\boldsymbol{M}^+ = \arg\min_{\boldsymbol{M}} \|\boldsymbol{T} - \boldsymbol{E} \odot \boldsymbol{M}\|_F^2 \quad s.t \quad \boldsymbol{M} \geq 0, \tag{22}$$

where $\boldsymbol{T} = \boldsymbol{X}^T\boldsymbol{D} - \boldsymbol{Y}$. Considering that the squared Frobenius norm of matrix can be optimized element by element, and problem (22) can be divided into $n \times c$ subproblems. For the $i$-th row and $j$-th column entry of $\boldsymbol{M}$, i.e. $\boldsymbol{M}_{ij}$, we have the following subproblem:

$$(\boldsymbol{T}_{ij} - \boldsymbol{E}_{ij}\boldsymbol{M}_{ij})^2 \quad s.t \quad \boldsymbol{M}_{ij} \geq 0. \tag{23}$$

Based on the result from [7], the optimal solution of $\boldsymbol{M}_{ij}$ is

$$\boldsymbol{M}_{ij} = \max(\boldsymbol{E}_{ij}\boldsymbol{T}_{ij}, 0). \tag{24}$$

Therefore, the compact form of the optimal solution of problem (22) is formulated as

$$\boldsymbol{M}^+ = \max(\boldsymbol{E} \odot \boldsymbol{T}, 0). \tag{25}$$

With the block coordinate descend procedures (17), (19), (21) and (25) recursively repeated, the asymptotic point $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{M})$ converges to a minimum point of $\mathcal{L}$ with respect to those variables, which is guaranteed by the theorem as follows.

**Theorem 2.** *Given $\boldsymbol{X}$, $\boldsymbol{C_1}$, and $\boldsymbol{E}$ defined as (9), suppose $\{(\boldsymbol{A}^k, \boldsymbol{B}^k, \boldsymbol{D}^k, \boldsymbol{M}^k)\}$ is a sequence generated recursively via the process (17), (19), (21) and (25), and then every limit point of $\{(\boldsymbol{A}^k, \boldsymbol{B}^k, \boldsymbol{D}^k, \boldsymbol{M}^k)\}$ is a minimum point of the augmented Lagrangian $\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{M}, \boldsymbol{C_1})$.*

***Proof.*** It can be easily verified that the loss function $\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{M}, \boldsymbol{C_1})$ is continuously differentiable with respect to $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{M}$ respectively, and in every subproblems of (16), (18), (20), and (22), the minimum point is uniquely obtained, according to the Proposition 2.7.1 in [36], every limit point of the sequence is a minimum point of $\mathcal{L}$. □

We iteratively optimize all the variables until the convergence condition is satisfied. To more clearly show the main procedures, the detailed algorithm of our optimization process of DENLR is outlined in Algorithm 1.

### *B. Optimization of MENLR*

It is easy to find that optimization of MENLR is very similar to the optimization procedures of DENLR, except for deducing the regression targets matrix $\boldsymbol{R}$. By ignoring the constant terms independent of $\boldsymbol{R}$, minimizing (14) becomes the following optimization problem:

$$\min_{\boldsymbol{R}} \|\boldsymbol{H} - \boldsymbol{R}\|_F^2 \ s.t. \ \boldsymbol{r}_{iy_i} - \max_{j \neq y_i} \boldsymbol{r}_{ij} \geq 1, i = 1, \cdots, n, \tag{26}$$

where $\boldsymbol{H}$=$\boldsymbol{X}^T\boldsymbol{D} \in \Re^{n\times c}$. Because problem (26) is a constrained quadratic programming problem, it can be decomposed into $n$ independent subproblems. Suppose that the $i$-th sample $\boldsymbol{x}_i$ is from the $m$th-class, and then the $i$-th subproblem of (26) is

$$\min_{\boldsymbol{r}_i} \|\boldsymbol{h}_i - \boldsymbol{r}_i\|^2 \ s.t. \ \boldsymbol{r}_{im} - \max_{j \neq m} \boldsymbol{r}_{ij} \geq 1, \tag{27}$$

where $\boldsymbol{r}_i \in \Re^c$ and $\boldsymbol{h}_i \in \Re^c$ are the $i$-th row of $\boldsymbol{R}$ and $\boldsymbol{H}$, respectively. It should be noted that $\|\boldsymbol{h}_i - \boldsymbol{r}_i\|^2 = \sum_{j=1}^c (\boldsymbol{h}_{ij} - \boldsymbol{r}_{ij})^2$. To optimize problem (27), we introduce an auxiliary variable $\boldsymbol{\varphi} \in \Re^c$, and for the $j$-th entry, $\boldsymbol{\varphi}_j = \boldsymbol{r}_{ij} + 1 - \boldsymbol{r}_{im}$, where $\boldsymbol{\varphi}_j \leq 0$ indicates the optimal target, otherwise a unsatisfactory target. Assume that the optimal target for the true class $\boldsymbol{r}_{im}$ can be obtained by a modification of the regression result $\boldsymbol{h}_{im}$, i.e. $\boldsymbol{r}_{im} = \boldsymbol{h}_{im} + \zeta$, where $\zeta$ is a learning parameter. For the false class $\forall j \neq m$, we need $\boldsymbol{r}_{im} - \boldsymbol{r}_{ij} \geq 1$, and then the $j$-th subproblem of (27) is

$$\min_{\boldsymbol{r}_{ij}} (\boldsymbol{h}_{ij} - \boldsymbol{r}_{ij})_2^2 \ s.t. \ \boldsymbol{h}_{im} + \zeta - \boldsymbol{r}_{ij} \geq 1, \forall j \neq m, \tag{28}$$

which is a very simple quadratic programming problem. In this way, the optimal solution is $\boldsymbol{r}_{ij} = \boldsymbol{h}_{ij} + min(\zeta - \boldsymbol{\varphi}_j, 0)$, and the optimal solution of problem (28) is achieved by

$$\boldsymbol{r}_{ij} = \begin{cases} \boldsymbol{h}_{ij} + \zeta, & if \ j = m, \\ \boldsymbol{h}_{ij} + min(\zeta - \boldsymbol{\varphi}_j), & otherwise. \end{cases} \tag{29}$$

By substituting (29) into problem (27), we can obtain the following optimization problem:

$$\arg\min_{\zeta} \phi(\zeta) = \zeta^2 + \sum_{j \neq m} (min(\zeta - \boldsymbol{\varphi}_j))^2, \tag{30}$$

and its first-order derivation $\phi'(\zeta) = 2(\zeta + \sum_{j \neq m} min(\zeta - \boldsymbol{s}_j))$. By setting $\phi'(\zeta) = 0$, we can achieve the optimal value of learning factor $\zeta$ as

$$\zeta = \frac{\sum_{j \neq m} \boldsymbol{\varphi}_j \Pi(\phi'(\boldsymbol{\varphi}_j) > 0)}{1 + \sum_{j \neq m} \boldsymbol{\varphi}_j \Pi(\phi'(\boldsymbol{\varphi}_j) > 0)}, \tag{31}$$

```
Algorithm 2. Solving Problem (27)
Input: $\boldsymbol{r} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_c]^T \in \Re^c$, the true class index $m$.
Initialization: $\forall j, \boldsymbol{\varphi}_j = \boldsymbol{h}_{ij} + 1 - \boldsymbol{h}_{im}$, $\zeta = 0$, $iter = 0$.
for $j \neq m$ do
    if $\psi'(\boldsymbol{\varphi}_j) > 0$ then $\zeta = \zeta + \boldsymbol{\varphi}_j, iter = iter + 1$ end
end
Define $\zeta = \zeta/(1 + iter)$, and then update $\boldsymbol{r}_j$ by Eqn.(29).
Output: Marginalized target vector $\boldsymbol{r}_i$.
```

```
Algorithm 3. Optimization of MENLR by Exact ALM
Require: Feature Matrix $\boldsymbol{X}$; Label Matrix $\boldsymbol{Y}$; Parameters $\lambda_1$, $\lambda_2$.
Initialization: $\boldsymbol{T} = \boldsymbol{Y}$, $\boldsymbol{D} \in \Re^{d\times c}$, $\boldsymbol{A} \in \Re^{d\times r}$, $\boldsymbol{B} \in \Re^{r\times c}$,
$\lambda_1 > 0$, $\lambda_2 > 0$, $\boldsymbol{C}_1 \in \Re^{d\times c}$, $\mu > 0$.
While not converged do
  While not converged do
    Step 1. Update $\boldsymbol{A}$ by using (17);
    Step 2. Update $\boldsymbol{B}$ by using (19);
    Step 3. Update $\boldsymbol{D}$ by using (32);
    Step 4. Update $\boldsymbol{R}$ row-by-row by using Algorithm 2;
  End While
  Step 5. Update the Lagrange multipliers $\boldsymbol{C}_1$ by
          $\boldsymbol{C}_1 = \boldsymbol{C}_1 + \mu(\boldsymbol{D} - \boldsymbol{AB})$.
End While
Output: Projection matrix $\boldsymbol{D}$
```

where $\Pi(\cdot)$ is the indicator operator. The detailed process of learning the optimal solution of the $i$-th row of $\boldsymbol{R}$ is given in Algorithm 2. The optimal solution of $\boldsymbol{D}$ is computed as

$$\boldsymbol{D}^+ = (2\boldsymbol{X}\boldsymbol{X}^T + \lambda_2\boldsymbol{I} + \mu\boldsymbol{I})^{-1}(2\boldsymbol{X}\boldsymbol{R} + \mu\boldsymbol{A}\boldsymbol{B} - \boldsymbol{C}_1). \quad (32)$$

In addition, the optimal solutions of $\boldsymbol{A}$ and $\boldsymbol{B}$ are the same as the optimization of DENLR. The detailed procedures of learning the optimal solutions of MENLR are summarized in Algorithm 3. Similarly, because optimization of $\boldsymbol{R}$ is a convex constrained quadratic programming problem, the following theorem is doubtlessly guaranteed.

**Theorem 3.** *Suppose $\{(\boldsymbol{D}^k, \boldsymbol{R}^k, \boldsymbol{A}^k, \boldsymbol{B}^k)\}$ is a sequence generated recursively via (32), iterative Algorithm 2, (17), and (19), and then every limit point of $\{(\boldsymbol{D}^k, \boldsymbol{R}^k, \boldsymbol{A}^k, \boldsymbol{B}^k)\}$ is a minimum point of the augmented Lagrangian $\mathcal{L}(\boldsymbol{D}, \boldsymbol{R}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C_1})$ of MENLR.*

***Proof***. The proof of Theorem 3 is similar to that of Theorem 2. □

### C. Classification

When the resulting problems of DENLR and MENLR are solved, the compact and discriminant projection matrix $\boldsymbol{D}$ is obtained. Then, we exploit projection matrix $\boldsymbol{D}$ to make linear transformations of both training and test samples. Finally, we employ a simple nearest neighbor (1-NN) classifier to perform multi-category image classificaton. The complete procedures of our classification model are summarized in Algorithm 4.

### D. Algorithm Analysis and Computation Complexity

It is worth noting that our ENLR framework is a generalized but robust extension of the conventional LSR and low-rank linear regression models. The following proposition shows the close relationship between our proposed DENLR and MENLR methods and the LSR and LRLR methods.

```
Algorithm 4. Classification
Input: Training feature set $\boldsymbol{X}$ with label vectors $\boldsymbol{Y}$, test sample
set $\boldsymbol{Z} \subset \Re^{d\times m}$.
Output: Predicted label matrix $\boldsymbol{L}_Z$ for test samples.
  Step 1. Normalize all the samples of both training and test
          samples to unit-norm by using $\boldsymbol{x}_i = \boldsymbol{x}_i/\|\boldsymbol{x}_i\|_2$.
  Step 2. Transform the training sample matrix $\boldsymbol{X}$ to the centering
          matrix by subtracting its mean value.
  Step 3. Exploit Algorithm 1 or Algorithm 3 to obtain an optimal
          projection matrix $\boldsymbol{D}$ is obtained.
  Step 4. Project $\boldsymbol{X}$ and $\boldsymbol{Z}$ onto $\boldsymbol{D}$ by
          $\tilde{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{D}$, $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}^T\boldsymbol{D}$
  Step 5. Predict the label matrix $\boldsymbol{L}_Z$ of test samples $\tilde{\boldsymbol{Z}}$
          by utilizing the nearest neighbor (NN) classifier.
```

***Proposition 3***: *The proposed ENLR framework is a generalized but robust linear regression model, and both of LSR and LRLR are the special cases of the proposed DENLR and MENLR methods.*

***Proof***. For model (13), when $\lambda_1 = 0$, $\lambda_2 = 0$ and $\boldsymbol{M} = \boldsymbol{0}_{n\times c}$, it will degenerate to the conventional LSR model (1). Moreover, if we set $\lambda_1 = 0$ and $\boldsymbol{M} = \boldsymbol{0}_{n\times c}$, it will become the regularized LSR model (2), where the $e_n b^T$ term can be absorbed into the $\boldsymbol{X}^T\boldsymbol{D}$ term. Furthermore, if we set $\lambda_2 = 0$ and $\boldsymbol{M} = \boldsymbol{0}_{n\times c}$, our DENLR model will degenerate to the LRLR model (4). So both of the LSR and LRLR models are the special cases of the proposed DENLR model, which is a general framework of linear regression. Similarly, we can find that the proposed MENLR method (14) is also a generalized version of the LSR and LRLR models.

More importantly, our DENLR and MENLR methods enlarge the margins of different classes by introducing the $\varepsilon$-dragging technique and enforcing the marginalized constraint, respectively. In this way, the regression targets are more reliable to fit the regression tasks such that the proposed methods are more discriminative and robust in comparison with existing linear regression models. Therefore, our methods can be viewed as a generalized discriminative framework of linear regression, and it can also be simply extended to other regression models.

Therefore, our ENLR framework not only intrinsically generalizes the previous LSR and LRLR models, but also extends the existing linear regression model to more robust and discriminative cases by seamlessly incorporating the slack and feasible regression targets. □

The overall computation complexity of our DENLR method mainly depends on the complexity of Algorithm 1. In Algorithm 1, the main computation load is mainly consumed on steps 1-4. The computational complexity of steps 1 and 2 is $\mathcal{O}(dcr)$ where $d$ is the dimensionality of the samples, $c$ is the number of classes, and $r$ is the rank of matrix $\boldsymbol{D}$. Note that calculating $\boldsymbol{D}$ will scale in about $\mathcal{O}(2d^2nc + d)$ due to the matrix inverse calculation, and computing $\boldsymbol{M}$ costs $\mathcal{O}(nc)$. So the total computational complexity of DENLR is $\mathcal{O}(2d^2nc + 2dcr + d)$. Similarly, the only difference between DENLR and MENLR is the calculation of $\boldsymbol{R}$, of which the complexity is $\mathcal{O}(nc)$. Therefore, the runtime complexity of MENLR is also $\mathcal{O}(2d^2nc + 2dcr + d)$ in each iteration.

### *E. Convergence Analysis*

We present a convergence results of the proposed Algorithm 1 and 3. First, it is worth noting that both of algorithms DENLR and MENLR have optimal solutions, and values of the objective functions are bounded. Although it is difficult to obtain a strong convergence property of the proposed algorithms, the empirical results suggest their strong convergence properties. Nevertheless we present a week convergence property of the proposed algorithm.

**Theorem 4.** *For DENLR, denote $(\boldsymbol{D}^k, \boldsymbol{M}^k, \boldsymbol{A}^k, \boldsymbol{B}^k, \boldsymbol{C}_1^k)$ as $\boldsymbol{\Psi}^k$, and suppose $\{\boldsymbol{\Psi}^k\}$ is a sequence generated via the Algorithm 1. Given $\boldsymbol{X}$, and $\boldsymbol{E}$ defined as (9), if the sequence is bounded, and*

$$\lim_{k \to +\infty} \{\boldsymbol{\Psi}^{k+1} - \boldsymbol{\Psi}^k\} = 0, \tag{33}$$

*then every limit point of $\{\boldsymbol{\Psi}^k\}$ is a Karush-Kuhn-Tucker point of the problem (13).*

***Proof.*** The detailed proof of the Theorem 4 is moved to Appendix B for better flow of the paper. □

Similarly, the convergence nature of MENLR is also easily demonstrated by the following theorem.

**Theorem 5.** *For MENLR, denote $(\boldsymbol{D}^k, \boldsymbol{R}^k, \boldsymbol{A}^k, \boldsymbol{B}^k, \boldsymbol{C}_1^k)$ as $\boldsymbol{\Phi}^k$, and suppose $\{\boldsymbol{\Phi}^k\}$ is a sequence generated via the Algorithm 3. Given $\boldsymbol{X}$, if the sequence is bounded, and*

$$\lim_{k \to +\infty} \{\boldsymbol{\Phi}^{k+1} - \boldsymbol{\Phi}^k\} = 0, \tag{34}$$

*then every limit point of $\{\boldsymbol{\Phi}^k\}$ is a Karush-Kuhn-Tucker point of the problem (14).*

***Proof.*** The proof of Theorem 5 is similar to Theorem 4. □

Although each exact minimum of the augmented Lagrangian of the Algorithms 1 and 3 guarantees a sound convergence property, it is impractical to obtain an exact solution in each iteration. The inner loop of BCD embedded in the main loop of ALM is time consuming. It is very common to boost up the computation time of ALM via inexact solution of the subproblems. Precision in each iteration is favored but not indispensable. In many cases the convergence of a recessive method could be preserved within a mild loss of precision in subproblems. Hence in this paper we speed up the Algorithm 1 and 3 by quitting the inner loop of BCD after one iteration. As a result the convergence issue may be questioned, but we empirically show in Section V-E that the convergence of the resulting inexact ALM is well preserved.

## V. Experimental Evaluation

In this section, we evaluate the performance of our proposed methods on publicly available databases, and compare them with currently popular linear regression methods for different classification tasks, i.e. face recognition, objection recognition and scene categories recognition. All the experiments are run 10 times with random data splits of training and test samples, and then the average classification results are reported on different datasets. We test our proposed methods on six public databases for three main tasks. Specifically, the performance of face recognition task is evaluated on four face databases: Extended YaleB [37], CMU PIE [38], AR [39], LFW [40]. Object recognition and scene recognition are performed on COIL-100 [41] and fifteen scene categories databases [42], respectively. It is worth pointing out that these databases are commonly used in multi-category image recognition and the existing methods have achieved decent results. Thus, challenging recognition results are convincing enough to verify the superiority of our methods, and a number of related state-of-the-art classification methods are enumerated as follows.

1) *SRC* [15]: It is to learn sparse representation based regression model with the $l_1$-norm regularization. Both reconstruction error and sparse codes are employed for classification.
2) *LLC* [19]: It is to learn a locality constrained regression model for large scale image classification. Locality-constrained codes are used for classification.
3) *CRC* [18]: It is to learn a linear regression model by using all the training samples with the $l_2$-norm regularization. Both reconstruction error and collaborative representation codes are used for classification.
4) *LRC* [17]: It is to learn a linear regression model by using each class of training samples with the $l_2$-norm regularization. Similar to CRC, the reconstruction error and learned representation codes are used for classification.
5) *LRLR* [23]: It is to learn a low-rank regression model by introducing the low-rank (nuclear norm) regularization. The learned projection matrix is used for classification.
6) *LRRR* [23]: It is to learn a low-rank ridge regression model by adding a Frobenius norm regularization on linear regression loss. Similar to LRLR, the learned projection matrix is used for classification.
7) *SLRR* [23]: It is to learn a sparse low-rank regression for feature section by imposing sparsity constraint on the low-rank regression loss. The low-rank projection matrix and selected features are used for classification.
8) *RPCA* [22], [27]: It is to learn clean images by decomposing a data matrix into low-rank term and sparse noise term, and then SRC on the clean data is used for classification.
9) *LatLRR* [26]: It is to learn salient features from the original dataset, and then linear regression model (2) is used to learn projection matrix. Subsequently, the linear transformation of the salient features obtained by using the learned projection matrix is employed for classification.
10) *LRSI* [27]: It is to learn a low-rank structured incoherence dictionary with shared features, and then the SRC method on the learned dictionary is used for classification.
11) *CBDS* [28]: It is to learn the data representation of training samples, test samples and dictionary with class-wise block-diagonal structure by imposing the low-rank regularization, and then the learned representation is used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2651396, IEEE Transactions on Image Processing

for classification.

12) *DLSR* [7]: It is to learn a discriminative LSR model by enlarging the distance between different classes in regression targets. The learned projection matrix are used for classification.
13) *SVM* [43]: It is to utilize support vector machine with Gaussian kernel on raw image features for classification. We use the LibSVM software [43]. Note that there exists an important regularization parameter $C$ in SVM. A cross validation approach is utilized to select it from the range of $[0.001, 0.01, 1.0, 10.0, 100.0]$. Actually, SVM is also a popular derivation of LSR model.
14) *DENLR*: The proposed model is to learn a compact and discriminative regression model by imposing the elastic-net regularization and enlarging the margin of the regression targets. The objective function is Eqn. (13). To justify the effectiveness of the $\epsilon$-dragging technique, we remove the $\epsilon$-dragging term, i.e. $\boldsymbol{E} \odot \boldsymbol{M}$, from Eqn.(13), and denote it as ***ENLR**** in the experiments.
15) *MENLR*: The proposed model is to learn a marginalized regression model by embedding the marginalized constraint of the regression targets into the elastic-net regularized framework, which is presented in Eqn. (14).

For fair comparison, we directly use the Matlab codes from the corresponding authors with the optimal parameter settings, or directly cite the experimental results from their original papers. Specifically, to guarantee the same experimental settings between all the compared methods and our methods on each benchmark, we re-implemented all the methods using optimal parameters via tenfold cross validation, and the training and test samples were randomly selected from each dataset. Moreover, the experimental settings on scene recognition is the same as that of the LC-KSVD [42] method, and we directly cite some experimental results from the original paper. For the compared methods that are not included in [42], we rerun them following the same experimental settings in [42]. Therefore, all the methods presented in our paper are performed for each dataset on the same testbed such that our experimental results are convincing and reliable.

### A. Face Recognition Evaluation

In this section, we evaluate the performances of our method for face recognition on four face databases.

***The Extended YaleB Database:*** The extended YaleB database contains 2414 front face images of 38 individuals and each subject has around 64 near frontal images under different illuminations. We randomly select 15, 20, 25, 30 images per subject for training, and the rest for testing. For all the compared methods, we exploit the suggested parameters in their papers for classification. The number of neighbors of LLC algorithm is set to fifteen, which is suggested as the best parameter for this dataset. Each image in this database for our experiments has been simply resized to 32×32 pixels. The classification accuracies of different methods on this database are summarized in Table I. Note that the mean classification accuracy and corresponding standard deviation (acc±std) are reported, and the bold numbers suggest the best classification accuracies. From Table I, it is clear to see that our method can consistently achieve the best classification accuracies with varying number of training samples. Moreover, we can see that if we remove the relax term of the regression target matrix, the performance of ENLR* is obviously better than other LSR methods, such as LRC, SRC, LRLR, LRRR and SLRR. This also reflects the fact that the elastic-net regularization term can lead to a more compact projection matrix such that higher classification accuracies can be achieved. Moreover, DENLR and MENLR can achieve the best classification accuracies in comparison with all the compared algorithms.

TABLE I: Classification accuracies (mean ± std %) of different methods with different numbers of training samples on the Extended YaleB database. The bold numbers are the best classification accuracy.

| Alg. | 15 | 20 | 25 | 30 |
|---|---|---|---|---|
| LLC | 88.63±0.31 | 91.52±0.48 | 94.20±0.58 | 95.21±0.35 |
| LRC | 89.47±1.16 | 92.05±0.99 | 93.50±0.67 | 94.62±0.66 |
| CRC | 91.39±1.35 | 94.26±1.27 | 95.91±0.90 | 97.04±0.72 |
| SRC | 91.72±0.48 | 93.71±0.69 | 95.56±0.36 | 96.37±0.45 |
| LRLR | 82.05±0.98 | 83.81±1.53 | 85.03±1.00 | 85.29±1.00 |
| LRRR | 82.37±1.24 | 83.65±0.78 | 85.46±0.93 | 86.01±0.94 |
| SLRR | 82.32±1.03 | 84.25±0.70 | 85.16±1.12 | 85.84±1.20 |
| DLSR | 92.37±0.73 | 94.78±0.71 | 95.84±0.42 | 96.97±0.43 |
| RPCA | 90.52±0.44 | 93.52±0.61 | 95.41±0.36 | 96.68±0.46 |
| LatLRR | 90.92±1.32 | 92.92±0.92 | 93.81±0.78 | 95.13±0.83 |
| LRSI | 92.71±0.58 | 94.26±0.33 | 96.16±0.55 | 96.98±0.45 |
| CBDS | 93.13±1.39 | 95.89±1.07 | 96.46±0.83 | 97.44±0.74 |
| SVM | 89.35±1.24 | 92.74±0.87 | 95.07±0.57 | 96.20±0.46 |
| **ENLR*** | 92.18±0.89 | 94.28±0.62 | 95.70±0.61 | 96.80±0.48 |
| **DENLR** | 94.34±1.05 | 96.66±0.56 | **97.70±0.57** | 98.51±0.45 |
| **MENLR** | **94.76±0.62** | **97.27±0.62** | 97.68±0.65 | **98.74±0.48** |

TABLE II: $p$-values between the proposed DENLR and MENLR methods and other methods on the Extended YaleB database.

| Alg. | DENLR | | MENLR | |
|---|---|---|---|---|
| | 15 | 25 | 15 | 25 |
| LLC | $2.09\times10^{-10}$ | $2.95\times10^{-12}$ | $3.48\times10^{-12}$ | $1.50\times10^{-11}$ |
| LRC | $4.35\times10^{-9}$ | $2.05\times10^{-13}$ | $5.41\times10^{-11}$ | $1.09\times10^{-12}$ |
| CRC | $3.55\times10^{-5}$ | $6.53\times10^{-6}$ | $1.46\times10^{-6}$ | $2.15\times10^{-5}$ |
| SRC | $3.55\times10^{-5}$ | $2.09\times10^{-12}$ | $3.83\times10^{-10}$ | $1.08\times10^{-10}$ |
| LRLR | $4.95\times10^{-16}$ | $2.09\times10^{-19}$ | $6.61\times10^{-18}$ | $4.93\times10^{-19}$ |
| LRRR | $4.55\times10^{-17}$ | $1.54\times10^{-19}$ | $8.67\times10^{-20}$ | $4.33\times10^{-19}$ |
| SLRR | $4.64\times10^{-16}$ | $1.09\times10^{-17}$ | $4.51\times10^{-18}$ | $2.03\times10^{-17}$ |
| DLSR | $2.42\times10^{-4}$ | $1.33\times10^{-9}$ | $6.22\times10^{-7}$ | $2.78\times10^{-8}$ |
| RPCA | $6.56\times10^{-12}$ | $2.50\times10^{-7}$ | $2.09\times10^{-10}$ | $7.64\times10^{-7}$ |
| LatLRR | $4.76\times10^{-6}$ | $1.75\times10^{-11}$ | $1.33\times10^{-7}$ | $6.01\times10^{-11}$ |
| LRSI | $4.47\times10^{-4}$ | $2.51\times10^{-7}$ | $1.60\times10^{-6}$ | $1.86\times10^{-6}$ |
| CBDS | 0.0412 | $6.87\times10^{-4}$ | $3.3\times10^{-3}$ | $2.3\times10^{-3}$ |
| SVM | $1.38\times10^{-8}$ | $5.02\times10^{-11}$ | $3.16\times10^{-10}$ | $4.39\times10^{-10}$ |

In addition, we conducted a statistical significance test for the results summarized in Table I to judge the significant improvements of the developed models in comparison with the state-of-the-art regression methods. The significance level, i.e. $p$-value, is typically set to 0.05, which means that if the significance evaluation is lower than this level, the performance difference between the evaluated methods is statistically significant. The $p$-values between the proposed DENLR and MENLR methods and the compared methods are shown in Table II, when the number of training samples for each subject

TABLE III: Classification accuracies (mean±std %) of different methods with different numbers of training samples on the CMU PIE database.

| Alg. | 15 | 20 | 25 | 30 |
|---|---|---|---|---|
| LLC | 84.62±0.57 | 90.90±0.25 | 93.27±0.36 | 94.46±0.41 |
| LRC | 85.61±0.62 | 90.17± 0.52 | 92.65±0.38 | 94.01±0.22 |
| CRC | 89.76±0.59 | 92.42±0.29 | 93.80±0.29 | 94.61±0.12 |
| SRC | 88.97±0.66 | 91.14±0.39 | 92.62±0.38 | 93.71±0.18 |
| LRLR | 83.70±0.57 | 85.73±0.58 | 86.80±0.45 | 87.62±0.48 |
| LRRR | 83.88±0.69 | 85.78±0.61 | 86.79±0.58 | 87.59±0.47 |
| SLRR | 83.69±0.64 | 85.85±0.50 | 86.77±0.61 | 87.58±0.47 |
| DLSR | 90.73±0.50 | 92.53±0.45 | 93.68±0.29 | 94.47±0.29 |
| RPCA | 84.26±0.41 | 88.24±0.32 | 91.06±0.12 | 92.26±0.22 |
| LatLRR | 84.26±0.41 | 88.24±0.32 | 91.06±0.12 | 92.26±0.22 |
| LRSI | 87.56±0.58 | 90.60±0.36 | 93.25±0.61 | 94.52±0.54 |
| CBDS | 88.58±0.65 | 91.50±0.42 | 93.41±0.46 | 94.53±0.37 |
| SVM | 86.66±0.75 | 90.70±0.63 | 92.66±0.53 | 93.06±0.35 |
| **ENLR*** | 90.47±0.53 | 92.82±0.45 | 93.94±0.45 | 94.67±0.26 |
| **DENLR** | 92.25±0.49 | 94.06±0.41 | 95.61±0.31 | 95.85±0.09 |
| **MENLR** | **93.21±0.44** | **94.88±0.29** | **95.74±0.23** | **96.18±0.15** |

TABLE IV: Classification accuracies (mean±std %) of different methods with different numbers of training samples on the AR database.

| Alg. | 8 | 11 | 14 | 17 |
|---|---|---|---|---|
| LLC | 54.26±1.27 | 60.87±0.91 | 66.88±1.03 | 71.58±1.32 |
| LRC | 63.87±1.42 | 76.87±1.13 | 85.20±1.00 | 90.88±0.97 |
| CRC | 86.53±1.07 | 91.66±0.77 | 94.06±0.77 | 95.74±0.76 |
| SRC | 84.08±0.98 | 89.45±0.74 | 92.20±1.19 | 95.14±0.67 |
| LRLR | 76.75±1.37 | 88.93±0.86 | 93.02±0.63 | 94.92±0.68 |
| LRRR | 91.40±0.71 | 93.82±0.70 | 95.42±0.48 | 96.47±0.70 |
| SLRR | 90.02±0.76 | 93.70±0.55 | 95.15±0.70 | 96.04±0.49 |
| DLSR | 89.56±0.68 | 93.65±0.67 | 94.36±0.62 | 95.18±0.46 |
| RPCA | 77.32±1.43 | 84.39±1.33 | 88.82±0.90 | 92.62±0.77 |
| LatLRR | 88.42±0.76 | 92.13±1.06 | 95.96±0.70 | 97.13±0.80 |
| LRSI | 78.78±1.02 | 85.93±1.01 | 89.92±0.76 | 93.17±0.97 |
| CBDS | 88.65±0.73 | 92.92±0.69 | 95.17±0.60 | 96.63±0.63 |
| SVM | 75.74±1.60 | 86.19±1.02 | 91.99±0.70 | 95.08±0.91 |
| **ENLR*** | 90.42±0.87 | 93.80±0.83 | 95.41±0.68 | 96.31±0.56 |
| **DENLR** | 91.94±0.80 | **95.69±0.70** | **97.30±0.62** | 98.21±0.54 |
| **MENLR** | **92.61±0.64** | 95.63±0.75 | 97.16±0.59 | **98.56±0.61** |

TABLE V: Classification accuracies (mean±std %) of different methods with different numbers of training samples on the LFW database.

| Alg. | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| LLC | 27.42±1.42 | 29.50±1.59 | 31.06±1.25 | 31.90±0.80 |
| LRC | 29.88±1.58 | 33.13±1.76 | 35.42±1.79 | 37.23±1.86 |
| CRC | 29.54±1.16 | 31.72±1.22 | 32.86±1.36 | 33.81±1.32 |
| SRC | 29.03±1.57 | 32.21±1.53 | 33.36±2.00 | 36.21±2.54 |
| LRLR | 29.88±1.02 | 30.18±0.74 | 34.45±1.63 | 35.16±2.17 |
| LRRR | 30.58±1.39 | 32.83±1.74 | 34.80±1.33 | 36.48±1.77 |
| SLRR | 30.72±1.23 | 33.02±1.53 | 35.32±1.41 | 36.40±1.69 |
| DLSR | 31.22±0.83 | 33.81±1.53 | 35.87±1.60 | 37.02±1.58 |
| RPCA | 29.82±1.59 | 32.52±1.36 | 34.45±1.63 | 36.27±1.43 |
| LatLRR | 29.96±1.06 | 33.22±1.85 | 35.30±1.90 | 37.12±1.65 |
| LRSI | 29.51±1.91 | 32.16±1.34 | 34.62±1.49 | 36.61±1.65 |
| CBDS | 31.13±1.44 | 32.83±1.46 | 34.30±1.52 | 36.30±1.82 |
| SVM | 29.66±1.64 | 32.36±1.70 | 35.46±1.42 | 36.73±1.45 |
| **ENLR*** | 30.66±1.01 | 33.28±2.13 | 35.22±1.67 | 36.41±1.87 |
| **DENLR** | 32.69±1.26 | 36.04±1.43 | 38.32±1.51 | 40.09±1.80 |
| **MENLR** | **34.97±1.35** | **37.13±1.37** | **39.79±1.29** | **41.26±1.65** |

is set to 15 and 25. We can see that the performance differences between our methods and all the compared methods are statistically significant, which also improves the effectiveness of our methods.

***The CMU PIE Database:*** The CMU PIE face database contains 41,368 face images from 68 subjects as a whole. The images under five near frontal poses (C05, C07, C09, C27 and C29) are used in our experiment. We randomly select 15, 20, 25, 30 images from each subject as training samples and the remaining images as test samples. The classification rates using different methods are summarized in Table III. We can see that our methods DENLR and MENLR always outperform the compared methods in different cases, and the performance of ENLR* in most cases is better than or competitive with all the compared methods.

***The AR Database:*** The AR face database contains about 4,000 color face images of 126 subject, which consist of the frontal faces with different facial expressions, illuminations and disguises. In this experiment, we select a subset including 2600 images from 50 female and 50 male subjects. We randomly select 8, 11, 14, 17 images for each subject as training samples and the rest of images as test samples. Following the implementation in [44], each image is projected onto a 540-dimensional feature vector with a randomly generated matrix from a zero-mean normal distribution. The experimental results obtained by using different classification methods are shown in Table IV. Apparently, our methods in most cases achieve the best classification results, which also verify that the proposed regression models outperform all the other regression models under different training conditions.

***The LFW Database:*** The Labeled Faces in the Wild (LFW) face database is designed for the study of unconstrained identity verification and face recognition. It contains more than 13,000 face images from 1680 subject pictured under the unconstrained conditions. In this experiment, we use a subset including 1251 images from 86 peoples and each subject has only 10-20 images [45]. Each image was manually cropped and resized to 32 × 32 pixels. We randomly choose 5, 6, 7, 8 images of each subject as training samples, and the remaining face images are exploited as test samples. Because the LFW database is a very difficult database for image classification, the accuracies obtained by utilizing different classification methods are comparatively not high, but the highest classification accuracies are still established by using our methods, which again certify the effectiveness of the proposed methods.

Overall, the proposed ENLR methods outperform all the compared regression methods on the four face image databases, which demonstrates that our methods can effectively solve the face recognition problem.

### *B. Object Recognition Evaluation*

To verify the assumption that our methods are feasible to solve object recognition task, we evaluate the performances of our methods on Columbia Object Image Library (COIL-100) database [41], which contains various views of 100 objects with different lighting conditions. In our experiments, the images are converted to gray-scale images with the 32 × 32 pixels, and then the robustness is evaluated on alternative viewpoints. We randomly select 15, 20, 25, 30 images per

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2651396, IEEE Transactions on Image Processing

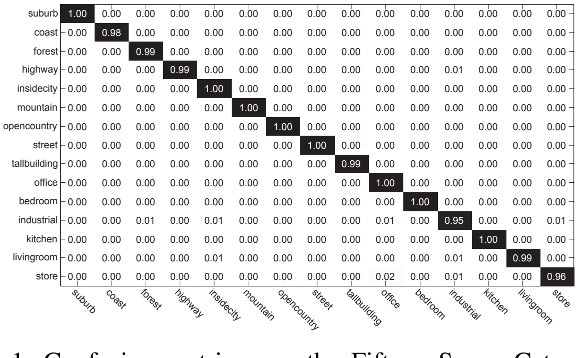
TABLE VI: Classification accuracies (mean±std %) of different methods with different numbers of training samples on the COIL-100 database.

| Alg. | 15 | 20 | 25 | 30 |
|---|---|---|---|---|
| LLC | 86.93±0.49 | 90.25±0.46 | 92.50±0.50 | 93.84±0.37 |
| LRC | 85.33±0.66 | 88.79±0.75 | 91.09±0.55 | 92.63±0.42 |
| CRC | 81.36±0.42 | 84.33±0.59 | 86.33±0.52 | 87.72±0.51 |
| SRC | 86.10±0.83 | 89.47±0.45 | 91.99±0.45 | 93.91±0.58 |
| LRLR | 70.59±0.64 | 72.79±0.82 | 74.47±0.70 | 76.00±0.76 |
| LRRR | 70.61±0.69 | 73.22±0.85 | 74.64±0.57 | 75.80±0.57 |
| SLRR | 71.85±0.59 | 73.81±0.70 | 73.69±0.53 | 76.47±0.51 |
| DLSR | 88.07±0.50 | 90.19±0.39 | 92.09±0.46 | 93.24±0.29 |
| RPCA | 88.31±0.87 | 91.72±0.31 | 93.53±0.35 | 95.28±0.34 |
| LatLRR | 85.30±0.40 | 88.43±0.43 | 90.72±0.44 | 92.47±0.45 |
| LRSI | 87.87±0.39 | 91.56±0.47 | 93.74±0.51 | 95.22±0.43 |
| CBDS | 77.04±0.80 | 77.84±0.66 | 79.55±0.60 | 81.32±0.75 |
| SVM | 84.89±0.62 | 88.10±0.47 | 90.80±0.65 | 92.44±0.42 |
| **ENLR*** | 88.40±0.36 | 91.28±0.39 | 93.37±0.39 | 94.66±0.27 |
| **DENLR** | 91.92±0.40 | 94.36±0.41 | 95.80±0.43 | 96.87±0.37 |
| **MENLR** | **92.75±0.51** | **94.88±0.48** | **96.34± 0.41** | **97.36±0.32** |

object to construct the training set, and the test set contains the rest of the images. The experimental results of different methods are summarized in Table VI. We can see that our methods always outperform all the other methods. Specifically, when the number of the training samples is 15, more than three percentages of classification rates are improved in comparison with the rest of methods. Accordingly, our methods have great potential in solving objective recognition task, which also reflects their effectiveness for multi-category recognition.

TABLE VII: Classification accuracies (mean±std %) of different methods on the fifteen scene database.

| Alg. | Accuracy | | Alg. | Accuracy |
|---|---|---|---|---|
| LLC | 79.4 | | SVM | 93.6 |
| LLC* | 89.2 | | LRSI | 92.4 |
| LRC | 91.9 | | LatLRR | 91.5 |
| CRC | 92.3 | | CBDS | 95.7 |
| LRLR | 94.4 | | DLSR | 95.9 |
| LRRR | 87.2 | | SRC | 91.8 |
| SLRR | 89.5 | | Lazebnik [42] | 81.4 |
| RPCA | 92.1 | | Lian [46] | 86.4 |
| LC_KSVD1 [44] | 90.4 | | Xie [47] | 83.27±0.83 |
| LC_KSVD2 [44] | 92.9 | | **ENLR*** | 97.1±0.24 |
| LRRC [29] | 90.1 | | **DENLR** | 98.7±0.17 |
| SLRRC [29] | 91.3 | | **MENLR** | **98.8±0.22** |



| | suburb | coast | forest | highway | insidecity | mountain | opencountry | street | tallbuilding | office | bedroom | industrial | kitchen | livingroom | store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| suburb | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| coast | 0.00 | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| forest | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| highway | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| insidecity | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| mountain | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| opencountry | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| street | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| tallbuilding | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| office | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| bedroom | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| industrial | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.95 | 0.00 | 0.00 | 0.01 |
| kitchen | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| livingroom | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.99 | 0.00 |
| store | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.00 | 0.96 |

Fig. 1: Confusion matrices on the Fifteen Scene Categories database.

### C. Scene Recognition Evaluation

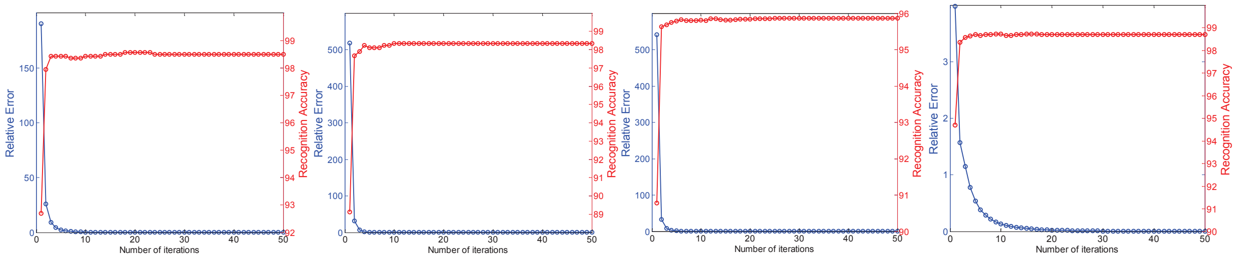
We evaluate the performance of our methods for scene recognition task by utilizing the Fifteen Scene Categories database [42]. It contains 4485 pictures falling into 15 categories, such as living rooms and kitchens. The data features of this database are provided in [44], which can be publicly available at http://www.umiacs.umd.edu/zhuolin/projectlcksvd.html. The following steps are processed to obtain the features. First, we compute a spatial pyramid feature with a four-level spatial pyramid [42] on the SIFT-descriptor codebook with size of 200, and then the final spatial pyramid features are reduced to 3,000 by PCA based feature dimension reduction. Following the common experimental settings [42] [44], we randomly select 100 images per category as training data, and use the remaining samples for testing. The number of neighborhoods of LLC$^*$ and LLC are respectively set to 30 and 70. The comparison results are summarized in Table VII. Our methods again establish the highest classification results and consistently outperform the performances of all the compared methods. Specifically, the classification accuracy of our method is better than the second best competitor about three percent. Furthermore, the confusion matrix of our DENLR method on the this database has been shown in Fig. 2. From confusion matrix of Fig. 2, we can see that each category classification accuracy is presented along the diagonal elements. It is notable that the classification accuracies for all the categories are close to 100%, and the worst performance is still very impressive with 95%, which also reflect the effectiveness of our DENLR method.

### D. Experiment Analysis

The average classification rates on six databases demonstrate the robustness and effectiveness of the proposed regression framework. Based on the experimental results on these databases, the following observations are achieved.

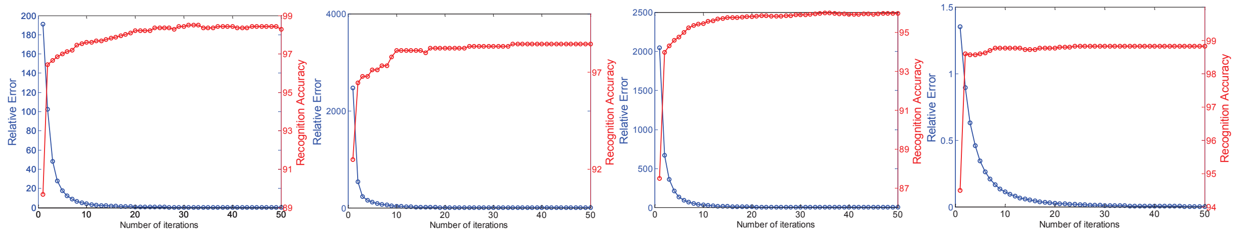
(1) DENLR and MENLR simultaneously consider the elastic-net property of the projection matrix and discriminative structure of the regression targets. As a result, it outperforms other regression methods, which only hold part properties. Our experimental results verify our previous key point that the proposed ENLR framework is better than the compared regression methods including representation based methods, linear regression and low-rank regression models.

(2) The proposed DENLR and MENLR methods are greatly superior to other regression models such as DLSR, LRLR, L-RRR, LRC and SLRR, because it takes the elastic-net property into consideration. The elastic-net property not only can better estimate the underlying distribution and structure of samples but also can enhance the generalization capabilities of DENLR and MENLR such that the learned projection matrix is more robust and discriminative. Specifically, low-rank regularization can capture the underlying subspace structure and correlation information of classes, while the Frobenius norm regularization avoids over-fitting of the proposed models. Integrating both terms as an elastic-net regularization of singular values is reasonable, and this also indicates that a compact and



(a) Extended YaleB (#25) (b) AR (#17) (c) CUM PIE (#25) (d) Fifteen scene (#100)

Fig. 2: Convergence curves of the relative error and classification accuracies versus the number of iterations for DENLR on (a) Extended YaleB, (b) AR, (c) CMU PIE and (d) Fifteen scene categories databases.



(a) Extended YaleB (#25) (b) AR (#17) (c) CUM PIE (#25) (d) Fifteen scene (#100)

Fig. 3: Convergence curves of the relative error and classification accuracies versus the number of iterations for MENLR on (a) Extended YaleB, (b) AR, (c) CMU PIE and (d) Fifteen scene categories databases.

discriminant projection matrix is significant and beneficial. The experimental results of ENLR* also demonstrate the above standpoint that in most cases ENLR* can achieve better classification results in comparison with the conventional LSR methods, such as LRC, CRC, LRLR, LRRR, SLRRR, and even recently proposed DLSR method.

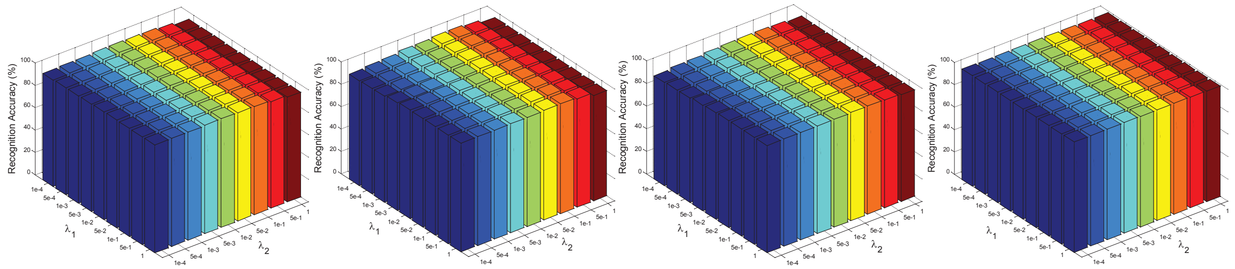
(3) Instead of employing the binary regression targets in conventional LSR methods, enlarging the margins of different classes in regression targets makes the regression task be further favorable such that accuracies of the proposed method are greatly improved. This is the another main reason that DENLR and MENLR outperform conventional LSR models. In addition, SVM is difficult to find the best decision function when the margins of different classes are close, while our method can obtain the optimal margins under the slack but discriminative target matrix. Thus, the performances of our methods are better than the compared low-rank and linear regression models. Due to slackening the strict binary matrix to the relaxed regression targets, there is no doubt that the performances of our robust ENLR methods are greatly improved, and DENLR and MENLR achieve the highest classification results in comparison with state-of-the-art linear regression methods.

(4) In addition, the experimental results of DENLR and MENLR are better than ENLR*, which further demonstrates that discriminative regression targets are beneficial to regression tasks. Moreover, we can see that MENLR in the most cases is better or comparable to DENLR, which indicates that providing more flexibility of regression targets is helpful to enhance the performances of linear regression models.

### *E. Convergence condition and parameters sensitivity*

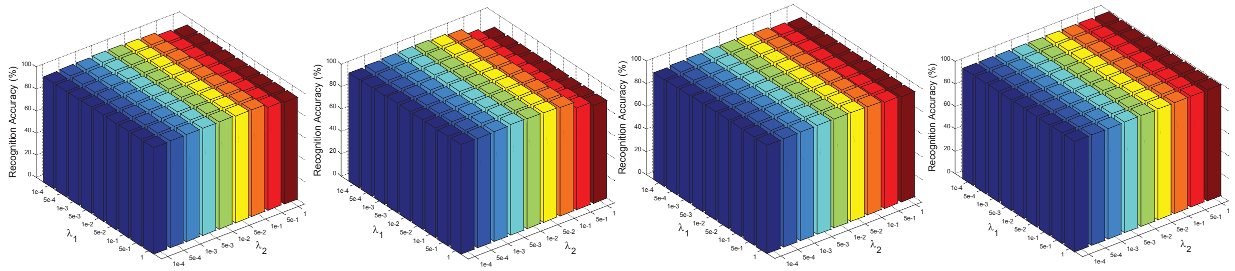
In this subsection, the convergence condition of the proposed method is analyzed and the influences of parameters $\lambda_1, \lambda_2$ are studied.

The overall convergence properties of our methods have been generally proved in theorems 4 and 5, which show that under mild conditions the iteration sequence of objective formulations of DENLR and MENLR can converge to the stationary point satisfying the KKT conditions, respectively. However, too much iterations can not fully meet the needs of practical applications. To this end, in our experiments we consider that the main concern of our regression model is to learn a compact and discriminative projection matrix $\boldsymbol{D}$ to make multi-category image classification. So, we directly take $\|\boldsymbol{D}^{k+1} - \boldsymbol{D}^k\|_F^2 \leq 10^{-5}$ as the convergence condition of algorithms, where $\boldsymbol{D}^k$ is the value of $\boldsymbol{D}$ for the $k$-th iteration. To confirm the efficient convergence of our methods, we implement the proposed DENLR and MENLR methods on four different datasets, i.e. the extended YaleB, AR, CMU PIE and fifteen scene categories databases. Figs. 2 and 3 show the convergence curves of DENLR and MENLR from the perspective of the relative error and classification accuracies versus the number of iterations on different databases, in which $\#Tr$ denotes the number of training samples selected from each subject. The results shown in figs. 2 and 3 demonstrate that the proposed optimization algorithms are effective and converge efficiently. Furthermore, empirical evidences show algorithms 1 and 3 converge within a small number of iterations and usually no more than 50 iterations, and the classification results become stable after 35 iterations.



(a) Extended YaleB (#25) (b) AR (#17) (c) CUM PIE (#25) (d) Fifteen scene (#100)

Fig. 4: Variations of DENLR classification (%) versus the parameters $\lambda_1$ and $\lambda_2$ on the (a)Extended YaleB, (b) AR, (c) CUM PIE and fifteen scene categories databases.



(a) Extended YaleB (#25) (b) AR (#17) (c) CUM PIE (#25) (d) Fifteen scene (#100)

Fig. 5: Variations of MENLR classification (%) versus the parameters $\lambda_1$ and $\lambda_2$ on the (a)Extended YaleB, (b) AR, (c) CUM PIE and fifteen scene categories databases.

In order to further investigate the properties of the proposed method, the classification performances versus the different values of regularization parameters, $\lambda_1$ and $\lambda_2$, are explicitly explored. To clearly show the results, we perform experiments on four databases, i.e. the extended YaleB, AR, CMU PIE and fifteen scene categories databases, to verify the parameters sensitivity. Specifically, we tune the value of both parameters from $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. Figs. 4 and 5 respectively show the classification results of DENLR and MENLR over variations of parameters. From figs. 4 and 5, we can observe that the performances of our DENLR and MENLR models are not very sensitive to the settings of $\lambda_1$ and $\lambda_2$. Apparently, when the parameters are not very large, the classification accuracies of our methods are not severely influenced. This also demonstrates that both regularization terms are indispensable for superior performances, and the best classification accuracies are achieved when both parameters are nonzero. Overall, the proposed regression models are not sensitive to the parameters provided they're in a reasonable range.

### *F. Efficiency Comparison*

To manifest the efficiency of the proposed methods, the runtime comparisons of our DENLR and MENLR methods with other compared methods are presented in this section. All algorithms were reimplemented using Matlab 2013a under Window 7 on a PC with a 3.3-GHZ CPU and 8-GB memory. We conduct experiments on the extended YaleB dataset to evaluate the computational time of different methods. For simplicity, we randomly choose 25 images from each subject as training samples, and the remaining images are utilized as the test samples. The computational time comparisons of different methods are summarized in Table VIII. We can see that most of methods have the training and test time, but the representation based methods, such as LLC, LRC, CRC and SRC, have only test time because they are designed to learn specific representations of test samples, and then directly use the representations to make classification. From Table VIII, we can see that our DENLR method is the fastest algorithm in comparison with all the other methods. Therefore, the efficiency of the proposed methods is demonstrated.

TABLE VIII: Run time comparisons of different methods (s).

| Alg. | Train | Test | | Alg. | Time | Test |
|---|---|---|---|---|---|---|
| LLC | — | 42.83 | | LatLRR | 128.80 | 0.63 |
| LRC | — | 59.50 | | LRSI | 9.29 | 44.41 |
| CRC | — | 43.39 | | CBDS | 153.90 | 1.71 |
| SRC | — | 899.55 | | DLSR | 4.56 | 0.36 |
| LRLR | 3.74 | 0.13 | | SVM | 0.12 | 4.24 |
| LRRR | 2.58 | 0.14 | | LC_KSVD | 64.50 | 0.84 |
| SLRR | 20.44 | 0.16 | | DENLR | 1.19 | 0.26 |
| RPCA | 89.43 | 0.61 | | MENLR | 2.91 | 0.24 |

## VI. Conclusion

In this paper, we developed a novel regression framework (ENLR) based on the elastic-net regularization of singular values for multi-category image classification. By introducing the elastic-net regularization scheme to capture the underlying structures of different classes, a more compact and discriminative projection matrix can be learned. Moreover, two robust elastic-net regularized linear regression methods were also introduced to demonstrate the effectiveness of the ENLR framework. Unlike conventional linear regression models which use the binary regression targets, our discriminative ENLR model relaxes the regression targets into a slack formulation, and the margins between different classes are enlarged to construct a more feasible regression scheme. Experimental results on public databases for different tasks demonstrated the superior performance of our DENSR and MENLR methods against the state-of-the-art image classification methods. We believe that the proposed method is not limited to classification tasks, and can be used for other general problems. For future work, we also plan to extend the proposed method to large-scale image analysis and understanding tasks.

## APPENDIX A
## PROOF OF THEOREM 1

Here we give a simple proof the Theorem 1. Suppose that $\boldsymbol{D} = \boldsymbol{A}\boldsymbol{B}$, $\boldsymbol{A} \in \Re^{d\times r}$ and $\boldsymbol{B} \in \Re^{r\times c}$ where $r$ is the rank of $\boldsymbol{D}$. Denote the singular value decomposition of $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary, and $\boldsymbol{\Sigma}$ is diagonal with non-negative entries. Then, $\boldsymbol{\Sigma} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{V} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{B}\boldsymbol{V}$, and then we have

$$\begin{aligned}\|\boldsymbol{D}\|_* = tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{B}\boldsymbol{V}) \leq \|\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{A}\|_F\|\boldsymbol{B}\boldsymbol{V}\|_F \\ = \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F.\end{aligned} \tag{35}$$

Thus, the first equality is obtained. For the second inequality '$\leq$' holds just according to the well-known inequality of arithmetic and geometric means (AM-GM inequality).

Notably, let $\boldsymbol{A} = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}$ and $\boldsymbol{B} = \sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^T$, and then $\boldsymbol{D} = \boldsymbol{A}\boldsymbol{B}$. It is easy to know that $\|\boldsymbol{D}\|_* = \|\boldsymbol{A}\boldsymbol{B}\|_* = \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}\|_* = tr(\boldsymbol{\Sigma})$. Furthermore,

$$\begin{aligned}\|\boldsymbol{D}\|_* = tr(\boldsymbol{\Sigma}) &= \sqrt{tr(\boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{U}^{\boldsymbol{T}})}\sqrt{tr(\sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V}\sqrt{\boldsymbol{\Sigma}})} \\ &= \sqrt{\|\boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\|_F^2}\sqrt{\|\sqrt{\boldsymbol{\Sigma}}V^T\|_F^2} = \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F.\end{aligned} \tag{36}$$

and then, the first '=' in Eqn. (11) holds such that the minimization of $\|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F$ is $\|\boldsymbol{D}\|_*$. On the other hand,

$$\begin{aligned}\|\boldsymbol{D}\|_* = tr(\boldsymbol{\Sigma}) &= \frac{1}{2}(tr(\boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{U}^{\boldsymbol{T}}) + tr(\sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V}\sqrt{\boldsymbol{\Sigma}})) \\ &= \frac{1}{2}(\|\boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\|_F^2 + \|\sqrt{\boldsymbol{\Sigma}}V^T\|_F^2) = \frac{1}{2}(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2).\end{aligned} \tag{37}$$

and then, the second '=' in Eqn. (35) holds. Therefore, under the constraint $\boldsymbol{D} = \boldsymbol{A}\boldsymbol{B}$, the minimization of $\frac{1}{2}(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2$ is $\|\boldsymbol{D}\|_*$. In this way, the above conclusion is proved.

## APPENDIX B
## PROOF OF THEOREM 4

Denote the loss function of the problem (13) as $\Psi(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{M})$. Karush-Kuhn-Tucker points of the problem (13) are those points which satisfy the conditions as follows:

$$\begin{aligned}\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B} &= \boldsymbol{0}, \\ \frac{\partial\Psi}{\partial\boldsymbol{A}} &= \boldsymbol{A}(\lambda_1\boldsymbol{I} + \mu\boldsymbol{B}\boldsymbol{B}^T) - (\boldsymbol{C}_1 + \mu\boldsymbol{D})\boldsymbol{B}^T = \boldsymbol{0}, \\ \frac{\partial\Psi}{\partial\boldsymbol{B}} &= (\lambda_1\boldsymbol{I} + \mu\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{B} - \boldsymbol{A}^T(\boldsymbol{C}_1 + \mu\boldsymbol{D}) = \boldsymbol{0}, \\ \frac{\partial\Psi}{\partial\boldsymbol{D}} &= (2\boldsymbol{X}\boldsymbol{X}^T + \lambda_2\boldsymbol{I} + \mu\boldsymbol{I})\boldsymbol{D} - 2\boldsymbol{X}\boldsymbol{S} \\ &\quad -\mu\boldsymbol{A}\boldsymbol{B} + \boldsymbol{C}_{\boldsymbol{1}} = \boldsymbol{0}, \\ \frac{\partial\Psi}{\partial\boldsymbol{M}} &= \boldsymbol{R}\odot\boldsymbol{E} - \boldsymbol{M} = \boldsymbol{0}.\end{aligned} \tag{38}$$

where $\boldsymbol{S} = \boldsymbol{Y} + \boldsymbol{E}\odot\boldsymbol{M}$ and $\boldsymbol{T} = \boldsymbol{X}^T\boldsymbol{D} - \boldsymbol{Y}$. We can obtain the Lagrange multipliers $\boldsymbol{C}_1$ from Algorithm 1 as

$$\boldsymbol{C}_1^k = \boldsymbol{C}_1^{k-1} + \mu(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B}), \tag{39}$$

where $\boldsymbol{C}_1^k$ is the $k$-th iteration of $\boldsymbol{C}_1$ in a sequence $\{\boldsymbol{C}_1^k\}_{k=1}^\infty$. If the sequences of multipliers $\{\boldsymbol{C}_1^k\}_{k=1}^\infty$ can converge to a stationary point, i.e. $(\boldsymbol{C}_1^k - \boldsymbol{C}_1^{k-1}) \to 0$, the following approximation results are obtained: $(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B}) \to 0$. So the first condition in Eqn. (38) is obtained.

For the second condition of the KKT conditions, the following equation can be obtained by using the optimization result of $\boldsymbol{A}$ in Algorithm 1 such that

$$\boldsymbol{A}^k - \boldsymbol{A}^{k-1} = (\boldsymbol{C}_1 + \mu\boldsymbol{D})\boldsymbol{B}^T(\lambda_1\boldsymbol{I} + \mu\boldsymbol{B}\boldsymbol{B}^T)^{-1} - \boldsymbol{A}, \tag{40}$$

which is equivalent to

$$\begin{aligned}(\boldsymbol{A}^k - \boldsymbol{A}^{k-1})(\lambda_1\boldsymbol{I} + \mu\boldsymbol{B}\boldsymbol{B}^T) = \boldsymbol{C}_1\boldsymbol{B}^T + \mu\boldsymbol{D}\boldsymbol{B}^T \\ -\lambda_1\boldsymbol{A} - \mu\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^T,\end{aligned} \tag{41}$$

where $\boldsymbol{A}^{k-1} = \boldsymbol{A}$ here. Based on the first condition $\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$, we can infer that $(\boldsymbol{C}_1\boldsymbol{B}^T - \lambda_1\boldsymbol{A}) \to \boldsymbol{0}$, if $(\boldsymbol{A}^k - \boldsymbol{A}^{k-1}) \to \boldsymbol{0}$. So the second condition is obtained.

Similar to the procedure of verifying the second condition, the third condition also can be obtained by utilizing the optimization result of $\boldsymbol{B}$ in Algorithm 1 such that

$$\begin{aligned}(\lambda_1\boldsymbol{I} + \mu\boldsymbol{A}^T\boldsymbol{A})(\boldsymbol{B}^k - \boldsymbol{B}^{k-1}) = (\boldsymbol{A}^T\boldsymbol{C}_1 - \lambda_1\boldsymbol{B}) \\ +\mu\boldsymbol{A}^T(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B}),\end{aligned} \tag{42}$$

where $\boldsymbol{B}^{k-1} = \boldsymbol{B}$ here. Similarly, we can infer that $(\boldsymbol{A}^T\boldsymbol{C}_1 - \lambda_1\boldsymbol{B}) \to \boldsymbol{0}$, when $(\boldsymbol{B}^k - \boldsymbol{B}^{k-1}) \to \boldsymbol{0}$. So we get the third condition.

Based on the optimization result of $\boldsymbol{D}$ in Algorithm 1, We also can get the following equation

$$\begin{aligned}(2\boldsymbol{X}\boldsymbol{X}^T + \lambda_2\boldsymbol{I} + \mu\boldsymbol{I})(\boldsymbol{D}^k - \boldsymbol{D}^{k-1}) = (\mu\boldsymbol{A}\boldsymbol{B} - \mu\boldsymbol{D}) \\ +(2\boldsymbol{X}\boldsymbol{S} + -\boldsymbol{C}_1 - 2\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{D} - \lambda_2\boldsymbol{D}).\end{aligned} \tag{43}$$

Based on the previous conditions, $\boldsymbol{A}\boldsymbol{B} - \boldsymbol{D}$ is approximate to zero such that the forth condition is satisfied based on the condition, i.e. $(2\boldsymbol{X}\boldsymbol{S} - \boldsymbol{C}_1 - 2\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{D} - \lambda_2\boldsymbol{I}) \to \boldsymbol{0}$, when $(\boldsymbol{D}^k - \boldsymbol{D}^{k-1}) \to \boldsymbol{0}$. Thus, the forth condition is achieved.

For the last condition, if we do not consider the constraint $\boldsymbol{M} \geq 0$, the optimization problem (22) can be rewritten as

$$f = \|\boldsymbol{T} - \boldsymbol{E}\odot\boldsymbol{M}\|_F^2, \tag{44}$$

where $\boldsymbol{T} = \boldsymbol{X}^T\boldsymbol{D} - \boldsymbol{Y}$. Similarly, the problem (44) can be divided into $n \times c$ subproblems. If we take the derivative of each subproblem and set it to zero, the final optimal solution is

$$\boldsymbol{M} = \boldsymbol{T} \odot \boldsymbol{E}, \tag{45}$$

where $\boldsymbol{E} \odot \boldsymbol{E} = \mathbf{1}_{d\times c}$ based on the definition of Eqn. (9). Like before operations, the following equation is satisfied

$$\boldsymbol{M}^k - \boldsymbol{M}^{k-1} = \boldsymbol{T} \odot \boldsymbol{E} - \boldsymbol{M}, \tag{46}$$

where $\boldsymbol{M}^{k-1} = \boldsymbol{M}$ here. So, if $(\boldsymbol{M}^k - \boldsymbol{M}^{k-1}) \to \mathbf{0}$, then $(\boldsymbol{T} \odot \boldsymbol{E} - \boldsymbol{M}) \to \mathbf{0}$. Furthermore, with the nonnegative constraint of $\boldsymbol{M} \geq \mathbf{0}$, we directly threshold the values of $\boldsymbol{M}$, which does not influence the convergence process.

It is easy to see that the value of our objective function has the minimum bound. Thus, the value sequence $\{\Psi^k\}_{k=1}^{\infty}$ of the objective function (13) is bounded, and $\{(\boldsymbol{A}^k)^T\boldsymbol{A}^k\}_{k=1}^{\infty}$ and $\{\boldsymbol{B}^k(\boldsymbol{B}^k)^T\}_{k=1}^{\infty}$ in Eqn. (42) and (41) are bounded. As a result, $\lim\limits_{k\to\infty}\{\Psi^{k+1} - \Psi^k\} = \mathbf{0}$ can deduce that both sides of equations (39), (41), (42), (43) and (46) are approximate to zero when $k \to \infty$. Therefore, the value sequence $\{\Psi^k\}_{k=1}^{\infty}$ of the objective function (13) can gradually satisfies the KKT conditions and the optimization algorithm, Algorithm 1, can converge to a local optimal solution. This is the end of proof.

## References

[1] L. Zhang, H. Shum and L. Shao, "Discriminative Semantic Subspace Analysis for Relevance Feedback", *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1275-1287, 2016.

[2] D. Huang, R. Cabral, F. D. l. Torre, "Robust Regression", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 363-375, 2016.

[3] S. Li, Y. Fu, "Learning Robust and Discriminative Subspace With Low-Rank Constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160-2173, 2016.

[4] M. Yu, L. Shao, X. Zhen and X. He, "Local Feature Discriminant Projection", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 38, no. 9, pp. 1908-1914, 2016.

[5] L. Shao, L. Liu and M. Yu, "Kernelized Multiview Projection for Robust Action Recognition," *Int. J. Comput. Visi.*, vol. 118, no. 2, pp. 115-129, Jun. 2016.

[6] Z. Zhang, Y. Xu, J. Yang, X. Li, D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE Access*, vol. 3, pp. 490-530, 2015.

[7] S. Xiang, F. Nie, G. Meng, et al." Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738-1754, 2012.

[8] S. Wold, H. Ruhe, H. Wold, and W. Dunn,"The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverse," *J. Sci. Stat. Comput.*, vol. 5, no. 3, pp. 735-743, 1984.

[9] T. Strutz, "Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond," Wiesbaden, Germany: Vieweg, 2010.

[10] Y. Li, A. Ngom, "Nonnegative Least-Squares Methods for the Classification of High-Dimensional Biological Data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2013, vol. 10, no. 2, pp. 447-456, 2013.

[11] X. Zhang, L. Wang, S. Xiang, C. Liu, "Retargeted Least Squares Regression Algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2206 - 2213, 2014.

[12] F. D. l. Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Analy. Mach. Intell.*, vol. 34, no. 6, pp. 1041-1055, 2012.

[13] X. Wen, L. Shao, W. Fang and Y. Xue, A Rapid Learning Algorithm for Vehicle Classification, Information Sciences, vol. 295, pp. 395C406, Feb. 2015.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, vol. 58, no. 1, pp. 267-288, 1996.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, 2009.

[16] Z. Zhang, L. Wang, Q. Zhu, Z. Liu, Y. Chen, "Noise modeling and representation based classification methods for face recognition," *Neurocomputing*, vol. 148, pp. 420-429, 2015.

[17] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106C2112, 2010.

[18] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" *in Proceedings of IEEE International Conference on Computer Vision*, pp. 471-478, 2011.

[19] J. Wang, J. Yang, K. Yu, et al. "Locality-constrained linear coding for image classification, " *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010.

[20] G. Xie, X. Zhang, S. Yan, C.-L. Liu, "Hybrid CNN and Dictionary-Based Models for Scene Recognition and Domain Adaptation,"*IEEE Trans. Circuits Syst. Video Technol.*, 2017, doi: 10.1109/TCSVT.2015.2511543.

[21] S. Li and Y. Fu, "Learning Balanced and Unbalanced Graphs via Low-Rank Coding," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1274-1287, 2015.

[22] E. Candès, X. Li,Y. Ma, et al. "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11, 2011.

[23] X. Cai, C. Ding, F. Nie, H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions,"*in proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1124-1132, 2013.

[24] G. Liu, Z. Lin, S. Yan, et al. "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171-184, 2013.

[25] S. Li and Yun Fu, "Robust Subspace Discovery through Supervised Low-Rank Constraints," *SIAM International Conference on Data Mining*, pp. 163-171, 2014.

[26] G. Liu, S. Yan. "Latent low-rank representation for subspace segmentation and feature extraction," *in proceeding of IEEE International Conference on Computer Vision*, pp. 1615-1622, 2011.

[27] C. Wei, C. Chen, Y. Wang, "Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3294-3307, 2014.

[28] Y. Li, J. Liu, H. Lu, et al. "Learning Robust Face Representation With Classwise Block-Diagonal Structure," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 12, pp. 2051-2062, 2014.

[29] Y. Zhang, Z. Jiang, L. Davis, "Learning structured low-rank representations for image classification," *in proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 676-683, 2013.

[30] E. Kim, M. Lee, S. Oh, "Elastic-Net Regularization of Singular Values for Robust Subspace Learning," *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 915-923, 2015.

[31] R. Cabral, F. D. l. Torre, J. Costeira, et al. "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," *in proceedings of IEEE International Conference on Computer Vision*, pp. 2488-2495, 2013.

[32] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *in proceedings of Advances in Neural Information Processing Systems*, pp. 41-18, 2007.

[33] H. Zou, T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal. Statist. Soc B.*, vol. 67, no. 2, pp. 301-320, 2005.

[34] R. Mazumder, T. Hastie, R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Machin. Learn. Research*, vol. 11, pp. 2287-2322, 2010.

[35] Z. Lin, M. Chen, Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices", arXiv preprint arXiv:1009.5055, 2010.

[36] D. P. Bertsekas, "Nonlinear Programming," Athena Scientific, Belmont, MA, 1999.

[37] A. Georghiades, P. Belhumeur, D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machin. Intell.*, vol. 23, no. 6, pp. 643-660, 2001.

[38] T. Sim, S. Baker, M. Bsat, "The CMU pose, illumination, and expression (PIE) database," *in proceedings of the fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-51, 2002.

[39] A. M. Martinez and R. Benavente, "The AR Face Database," *CVC Technical Report*, no. 24, June 1998.

[40] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[41] S. A. Nene, S. K. Nayar and H. Murase, “Columbia Object Image Library (COIL-100),” Technical Report CUCS-006-96, 1996.

[42] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.

[43] C. Chang, C. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Sys. Technol.*, vol. 2, no. 3, pp. 27, 2011.

[44] Z. Jiang, Z. Lin, L. Davis, “Label consistent K-SVD: Learning a discriminative dictionary for recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651-2664, 2013.

[45] S. J. Wang, J. Yang, M. F. Sun, et al. “Sparse tensor discriminant color space for face verification”, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, pp. 876-888, 2012.

[46] X. Lian, z. Li, B. Lu, et al. “Max-margin dictionary learning for multiclass image categorization,” *in proceedings of ECCV*, pp. 157-170, 2010.

[47] G. Xie, X. Zhang, C.-L. Liu, “Efficient Feature Coding Based on Auto-encoder Network for Image Classification,” *in proceedings of Asian Conference on Computer Vision*, pp. 628-642, 2014.

**Zheng Zhang** received the B.S degree from Henan University of Science and Technology and M.S degree from Shenzhen Graduate School, Harbin Institute of Technology (HIT) in 2012 and 2014, respectively. Currently, he is pursuing the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, machine learning and computer vision.

**Zhihui Lai** received the B.S degree in Mathematics from South China Normal University, M.S. degree from Jinan University, and the Ph.D. degree in Pattern Recognition and Intelligence System from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University since 2010. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He serves as an Associate Editor on International Journal of Machine Learning and Cybernetics. For more information, the readers are referred to the website http://www.scholat.com/laizhihui.

**Yong Xu** (M’06-SM’15)was born in Sichuan, China, in 1972. He received his B.S. degree and M.S. degree at Air Force Institute of Meteorology (China) in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern recognition and Intelligence System at the Nanjing University of Science and Technology (NUST) in 2005. Now, he works at Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning and video analysis.

**Ling Shao** (M’09-SM’10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with the University of Sheffield and a senior scientist (2005- 2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of the IEEE.

**Jian Wu** received the B.S. degree in mathematics from Liaoning Normal University, Dalian, China, in 2010, and the M.S. degree in mathematics from Gannan Normal University, Ganzhou, China, in 2014. He is currently pursuing the Ph.D. degree in computer science with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His research interests focus on medical biometrics, pattern recognition, and image processing.

**Guosen Xie** is currently an Assistant Professor at the Department of Information Engineering at Henan University of Science and Technology. He received the M.S. degree in computational Mathematics from Wuhan University, Wuhan, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. He received the Best Student Paper Awards from MMM’16. His research interests include machine learning, deep learning, and their applications to object recognition and DNA sequence analysis.