Pairwise Relations Oriented Discriminative Regression

Chao Zhang, Huaxiong Li, Yuhua Qian, Chunlin Chen, and Yang Gao

***Abstract*—Linear Regression (LR) is a popular and effective technique in pattern recognition area, which aims to find a transform matrix between source data and target data (usually label matrix). However, a binary zero-one label matrix may be too strict and inappropriate for regression. Besides, directly projecting source data to target data by one transform matrix may lose some intrinsic data information. To address these issues, this paper proposes a novel Pairwise Relations oriented Discriminative Regression (PRDR) method. In PRDR, the source data is regressed into a latent space instead of label space. To supervise the discriminative projection learning, the pairwise relations in source data space and label space are exploited in the latent space simultaneously. The pairwise label relations are transferred into the latent subspace by solving a distance-distance difference minimization problem, and the intraclass instance relations are also preserved in latent space. These two constraints ensure the pairwise similarity of data points after transformation which is beneficial for classification. By further enlarging the margins between true and false classes, PRDR is extended to a robust version, i.e., R-PRDR. An efficient algorithm is presented to solve the PRDR model. Extensive experiments on several popular image datasets demonstrate the effectiveness and efficiency of the proposed method compared with some state-of-the-art regression approaches.**

***Index Terms*—Latent representation, discriminative regression, pairwise relations, classification.**

I. Introduction

LINEAR regression (LR) is one of the most popular and effective techniques in the fields of machine learning and pattern recognition, and it has been widely used in face recognition [1]–[3], image processing [4]–[6] and classification [7]–[9], visual tracking [10], information retrieval [11], etc. The fundamental objective of LR is to seek an appropriate transform matrix between source data and target data such that the transformed source data can well fit the target data. As a typical LR method, linear regression based classification (LRC) learns a regression vector between a test sample and training samples, and performs classification by regression error [12]. Some variants such as sparse representation based classification (SRC) [13] and collaborative representation based classification (CRC) [14] also adopt the linear regression framework and exploit the characteristic of regression coefficients. However, these methods pay more attention to regression loss and ignore the important label information, which limits their performance in some tasks [15], [16].

To make use of the labels, some researchers directly connect the source data and label information by least squares regression (LSR). In multi-class classification tasks, a binary zero-one label matrix is first defined as regression target and a transform matrix is learned between training data matrix and label matrix. However, this binary label matrix is strict for regression, which may lead to overfitting and degraded performance [17]–[21]. Therefore, various soft label techniques are developed to relax the label matrix [18], [22]–[25]. Xiang and Nie *et al.* introduced the ϵ-dragging technique into LR, and proposed a discriminative least squares regression (DLSR) method for pattern classification [18]. The core idea of DLSR is to add an auxiliary vector on the binary label vectors which enlarges the distances between the true and false classes. By imposing a sparsity constraint to explicitly control the margins, DLSR is further extended to margin scalable discriminative LSR (MSDLSR) [17]. Zhang *et al.* proposed a more flexible model, i.e., retargeted least squares regression (ReLSR) [22]. ReLSR adaptively learns a target label matrix, in which the margin between correct class and false classes of each sample is forced to be large. These soft label strategies are widely used in other researches [26]–[29]. These methods mainly pursuit the large margins between different classes (i.e., to enhance the interclass separability) by various techniques to improve the discrimination of projection. However, they cannot guarantee the label consistency of samples from the same class and the intraclass compactness is destroyed due to the dynamic of regression target [30].

To ensure the intraclass compactness in regression, various regularization terms are utilized including nuclear norm [30], [31], $l_{2,1}$ norm [32] and other constraints [33], [34]. In [30], the authors imposed a classwise low-rank constraint on the latent features to enhance the similarity of data representation, and proposed a group low-rank representation-based discriminant linear regression (GLRRDLR) method. In [32], an interclass sparsity based discriminative LSR (ICS_DLSR) method is proposed. ICS_DLSR uses a row-sparsity regularized term to preserve the data similarity and an error term to relax the label matrix. Wang *et al.* extended ReLSR to groupwise ReLSR (GReLSR) by re-

This work was partially supported by the National Key Research and Development Program of China (Nos. 2018YFB1402600, 2016YFD0702100), and the National Natural Science Foundation of China (Nos. 71671086, 61672332, 61876079).

C. Zhang, H. Li and C. Chen are with the Department of Control and Systems Engineering, Nanjing University, Nanjing 210093, China (e-mail: chzhang@smail.nju.edu.cn, {huaxiongli, clchen}@nju.edu.cn).

Y. Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China, and also the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China (e-mail: jinchengqyh@126.com).

Y. Gao is with the Department of Computer Science and Technology, National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail:gaoy@nju.edu.cn).

stricting the regression target [33]. These approaches mainly impose classwise constraints on transformed data, which can improve the discrimination of regression to a certain extent but lead to cumbersome and time-consuming optimization. Due to the merits of manifold learning on local structure preservation [35], [36], the graph regularization provides another effective way to improve the data intraclass compactness [37]–[39]. Yang *et al.* incorporated a new Laplacian matrix into regression model to capture the local structure of data [37]. Shi *et al.* combined the graph embedding and sparse regression into a unified model [36]. Xue *et al.* constructed a label based graph to regularize the projection which pulls the samples from the same class to be close [40]. In [41], the authors performed non-negative sparse graph learning and linear regression simultaneously, and applied it on semi-supervised classification. These researches demonstrate that the use of graph based regularization is beneficial to capture the intrinsic data structure and preserve the local relationships.

Despite different constraints and regularizations, what these methods mentioned above have in common is that they all seek one transform matrix from original data space to target label space, and we refer to them as one-step transform based LR methods. The one-step operation may lose some underlying information or structures of data [42], [43]. Some researchers proposed two-step transform based methods, in which a latent space is generated and bridges the data space and label space [43]–[46]. In [43], the authors conducted LR in regularized linear discriminant analysis (LDA) space, and proposed low-rank ridge regression (LRRR) with Frobenius norm and sparse low-rank regression (SLRR) with $l_{2,1}$ norm. Fang *et al.* proposed a robust latent subspace learning (RLSL) method [44]. RLSL jointly learns a middle transition space and regresses the latent features to label matrix. The data reconstruction mechanism like PCA is integrated into RLSL to regularize the latent features. Zhen *et al.* learned a latent space between training data space and label space with low-rank regularization [46]. Although these methods flexibly learn the data representation in latent space, they still try to learn a linear transform matrix from latent space to label space for supervised learning, thus, the problem of rigid regression target still exists. Besides, these methods usually learn multiple projection matrices in a unified model, which makes the optimization complicated and time-consuming.

To this end, in this paper, we propose a novel pairwise relations oriented discriminative regression (PRDR) for multi-class classification. PRDR can be categorized into two-step transform based methods, which learns a discriminative transform matrix leveraging a latent subspace. Differently, PRDR only learns one projection matrix which makes it efficient in real application. To avoid a strict regression target, the proposed method adopts the label relations rather than original label matrix to supervise the projection learning. The label relations are explored in latent space by solving a pairwise distance-distance difference minimization problem which enhances the intraclass compactness and interclass separability. Besides, the instance relations are also explored, and an intraclass similarity graph is constructed from the training data and embedded into the framework. PRDR is proved to constrain the pairwise cosine distance of latent representation. By further enlarging the margins between different classes, PRDR is extended to a more discriminative version. The main contributions of this paper are summarized as follows:

- We propose a novel pairwise relations oriented discriminative regression (PRDR) with application to image classification. The training data is regressed into the latent space rather than label space to avoid the strict regression target problem.
- To make use of label information, a distance-distance difference minimization constraint is used to preserve the pairwise label relations of data. Moreover, an intraclass similarity graph is incorporated into PRDR to preserve the pairwise instance relations.
- PRDR is proved to constrain the pairwise distances of latent representation. By enlarging the margins between true and false classes, PRDR is further extended to R-PRDR method.
- An efficient algorithm based on alternating direction method of multipliers (ADMM) is presented to solve the proposed model with both theoretical and empirical analysis. The experimental results demonstrate the effectiveness and efficiency of our proposed methods.

The remainder of this paper is organized as follows. Section II briefly overviews the related methods. Section III introduces our proposed methods in detail. Section IV reports the experimental results and analysis. Section V concludes the paper.

II. Related Works

For convenience, we first present some notations used in this paper. Matrices and vectors are written in boldface uppercase and boldface lowercase, respectively. For matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$, $\mathbf{M}_{i,j}$ denotes its i-th row and j-th column element, $\mathbf{M}_{i,:}$ and $\mathbf{M}_{:,j}$ represent the i-th row vector and j-th column vector, respectively. The Frobenius norm, nuclear norm and $l_{2,1}$ norm of matrix $\mathbf{M}$ are defined as: $\|\mathbf{M}\|_F = \sqrt{\sum_i^p \sum_{j=1}^q \mathbf{M}_{i,j}^2}$, $\|\mathbf{M}\|_* = \sum_i \delta_i(\mathbf{M})$, where $\delta_i(\mathbf{M})$ is the i-th singular value of $\mathbf{M}$, and $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^p \sqrt{\sum_{j=1}^q \mathbf{M}_{i,j}^2}$ respectively. $Tr(\cdot)$ is the trace function and $\mathbf{I}$ is an identity matrix.

Given a training matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{m\times n}$ composed of n instances and its corresponding binary label matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathbb{R}^{c\times n}$, where c is the number of classes. For label vector $\mathbf{y}_i$, its k-th entry is 1 and all the others are 0 if instance $\mathbf{x}_i$ belongs to the k-th class. The basic model of LR can be expressed as follows:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \lambda\|\mathbf{W}\|_F^2, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{c\times m}$ is the to-be-learned transform matrix and $\lambda > 0$ is a balance parameter. Based on model (1),

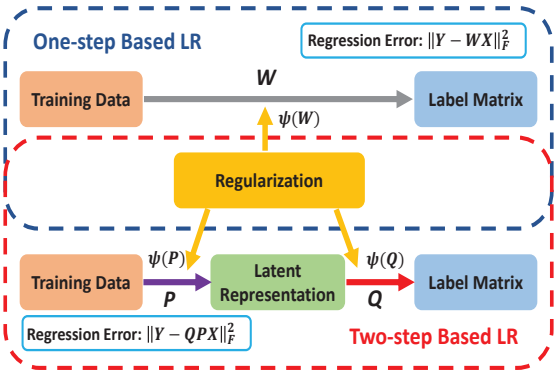


Fig. 1. The overall frameworks of one-step and two-step transform based LR models.

many robust variants are proposed and developed [18], [22], [23], [32], [33], [44]. These regression based models can be generally categorized into two types, i.e., one-step and two-step transform based methods, depending on whether a latent space between training data space and label space is used. Fig. 1 shows the overall frameworks of the two kinds of models.

A. One-step Transform Based LR

One-step transform based LR methods directly link the data space with label space. By relaxing the label matrix, the regression model can be described as

$$\min_{\mathbf{W}} \|\vartheta(\mathbf{Y}) - \mathbf{W}\mathbf{X}\|_F^2 + \lambda\psi(\mathbf{W}), \tag{2}$$

where ϑ is a relaxation function, and $\psi(\mathbf{W})$ is the regularization term. The commonly used $\psi(\mathbf{W})$ are $\|\mathbf{W}\|_F^2$, $\|\mathbf{W}\|_*$ and $\|\mathbf{W}\|_{2,1}$ [42], [43]. ϵ-dragging technique is widely used to relax $\mathbf{Y}$, i.e., $\vartheta(\mathbf{Y}) = \mathbf{Y} + \mathbf{M} \odot \mathbf{B}$, where $\mathbf{M}$ is a nonnegative matrix, $\odot$ is the element-wise production, and $\mathbf{B}$ is defined as

$$\mathbf{B}_{i,j} = \begin{cases} +1, & \text{if } \mathbf{Y}_{i,j} = 1, \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

It can be observed that the target value of true class in relaxed label matrix is above 1, while the values of false classes are negative. Thus, ϵ-dragging enlarges the margins between true and false classes to learn a discriminative $\mathbf{W}$. Its drawback is also obvious that the relaxed label vectors of two samples from the same class are different due to the dynamic of $\mathbf{M}$, and the intraclass similarity cannot be ensured. To address this problem, model (2) is extended as follows:

$$\min_{\mathbf{W}} \|\vartheta(\mathbf{Y}) - \mathbf{W}\mathbf{X}\|_F^2 + \lambda_1\psi(\mathbf{W}) + \lambda_2\phi(\mathbf{W}\mathbf{X}), \tag{4}$$

where $\phi(\mathbf{W}\mathbf{X})$ is a regularization term. Graph regularization is widely used in $\phi(\mathbf{W}\mathbf{X})$ to preserve the local structure of data [47], [48], in which a similarity graph is embedded into projection learning. ICS_DLSR adopts classwise $l_{2,1}$ norm constraint as $\phi(\mathbf{W}\mathbf{X}) = \sum_{i=1}^{c} \|\mathbf{W}\mathbf{X}_i\|_{2,1}$, where $\mathbf{X}_i$ denotes the training data of the i-th class [32]. GLRRDLR uses low-rank constraint, i.e., $\phi(\mathbf{W}\mathbf{X}) = \sum_{i=1}^{c} \|\mathbf{W}\mathbf{X}_i\|_*$ [30]. These classwise constraints can improve the intraclass compactness, which is beneficial to classification. However, they usually lead to c subproblems in optimization, which are time-consuming when the number of classes is large.

B. Two-step Transform Based LR

Instead of directly mapping the training data into label space, two-step transform based methods learn a latent space as a bridge, which is more flexible compared with one-step transform based methods. The general framework of two-step based methods can be formulated as

$$\min_{\mathbf{P},\mathbf{Q}} \|\vartheta(\mathbf{Y}) - \mathbf{Q}(\mathbf{P}\mathbf{X})\|_F^2 + \lambda_1\psi(\mathbf{P},\mathbf{Q}) + \lambda_2\phi(\mathbf{P}\mathbf{X}), \tag{5}$$

where two transform matrices $\mathbf{P} \in \mathbb{R}^{d\times m}$ and $\mathbf{Q} \in \mathbb{R}^{c\times d}$ are learned, and d is the dimension of latent data representation $\mathbf{P}\mathbf{X}$. LRRR restricts the dimension d to force the whole transform matrix $\mathbf{Q}\mathbf{P}$ to be low-rank [43]. In RLSL [44], the regularizations of $\mathbf{P}$ and $\mathbf{Q}$ are the square of Frobenius norm, i.e., $\psi(\cdot) = \|\cdot\|_F^2$, and data reconstruction property is used as the constraints on latent features $\mathbf{P}\mathbf{X}$. In [46], the authors use nuclear norm and Frobenius norm to constrain the projection matrix $\mathbf{Q}$ simultaneously. With different constraints, two-step based methods are generally more flexible to learn discriminative projections compared with one-step based methods which can only project original samples into a c-dimensional subspace. However, two-step based methods are usually time-consuming since they need to learn multiple projection matrices. Besides, the strict regression target problem still exists in these methods.

III. The Proposed Method

In this section, we introduce our proposed PRDR method in detail, and present the optimization algorithm as well as the convergence and computational complexity analysis. Finally, we will extend PRDR to R-PRDR method.

A. Formulation

As analyzed before, two-step based methods are more flexible to exploit the intrinsic information of data and learn a discriminant projection matrix by using a latent space. Thus, in this paper, instead of directly regressing the training data into label space, we use latent data representation as the regression target. Denote the latent representation as $\mathbf{V} \in \mathbb{R}^{d\times n}$, and the preliminary PRDR model can be described as follows:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{V} - \mathbf{W}\mathbf{X}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2. \tag{6}$$

This is a standard least squares regression problem which can be efficiently solved with a closed-form solution. However, the latent representation $\mathbf{V}$ is unknown. From the view of classification, $\mathbf{V}$ should have optimal intraclass similarity as well as interclass separability [49], [50]. To achieve the goals, in some existing works [44]–[46], another transform

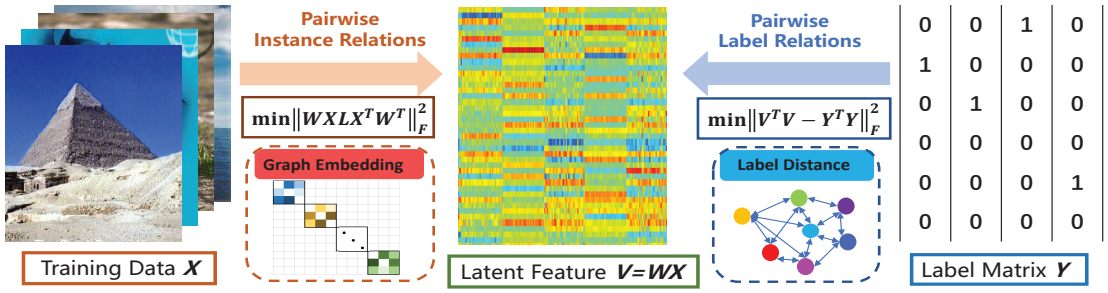


Fig. 2. The overall framework of PRDR. PRDR learns the discriminative latent representation by simultaneously considering the pairwise instance relations and label relations. The pairwise label distance is transferred into latent space for supervised learning and the intraclass compactness graph is embedded into projection learning for structure preserving.

matrix for $\mathbf{V}$ is employed into (6) which attempts to approximate label matrix $\mathbf{Y}$ by transformed latent representation, as presented in (5). However, such strategy for supervised learning introduces rigid regression target problem and more unknown variables, making the optimization complicated and inefficient.

To tackle these problems, we consider the pairwise label relations in this paper to guide the latent representation learning, instead of using label matrix as regression target. In label matrix $\mathbf{Y} \in \{0,1\}^{c\times n}$, the samples from the same class share a same label while others not. In other words, in label space, the label distance between two samples from same class is 0, and that is above 0 between two samples from different classes. To transfer the pairwise label relations into latent space, we can minimize the following pairwise distance-distance difference problem

$$\begin{aligned}
&\sum_{i,j=1}^{n}(h_{i,j}^{V}-h_{i,j}^{Y})^2 \\
&=\sum_{i,j=1}^{n}(\|\mathbf{V}_{:,i}-\mathbf{V}_{:,j}\|^2-\|\mathbf{Y}_{:,i}-\mathbf{Y}_{:,j}\|^2)^2 \\
&=\sum_{i,j=1}^{n}(\|\mathbf{V}_{:,i}\|^2+\|\mathbf{V}_{:,j}\|^2-2\mathbf{V}_{:,i}^T\mathbf{V}_{:,j}-\|\mathbf{Y}_{:,i}\|^2 \\
&\quad-\|\mathbf{Y}_{:,j}\|^2+2\mathbf{Y}_{:,i}^T\mathbf{Y}_{:,j})^2,
\end{aligned} \tag{7}$$

where $h_{i,j}^{V}$ is the squared Euclidean distance between sample $\mathbf{X}_{:,i}$ and $\mathbf{X}_{:,j}$ in latent feature space, and $h_{i,j}^{Y}$ is that in label space. It is worth noting that $\|\mathbf{Y}_{:,i}\|=1$ because $\mathbf{Y}_{:,i}$ is a one-hot vector. By adding a normalization constraint $\|\mathbf{V}_{:,i}\| = 1(i = 1,...,n)$, Eq. (7) can be equivalently rewritten as

$$\sum_{i,j=1}^{n}(2\mathbf{V}_{:,i}^T\mathbf{V}_{:,j}-2\mathbf{Y}_{:,i}^T\mathbf{Y}_{:,j})^2=4\|\mathbf{V}^T\mathbf{V}-\mathbf{Y}^T\mathbf{Y}\|_F^2. \tag{8}$$

Eq. (8) uses the pairwise label relations from label space to guide the latent representation learning without other auxiliary variables. By minimizing the above distance-distance difference problem, the label similarity is preserved in latent space. With the above objective, we can get the following model

$$\begin{aligned}
&\min_{\mathbf{W},\mathbf{V}}\ \frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{V}-\mathbf{Y}^T\mathbf{Y}\|_F^2 \\
&\quad\text{s.t. } \{\|\mathbf{V}_{:,i}\|\}_{i=1}^{n}=1.
\end{aligned} \tag{9}$$

Although problem (9) considers the pairwise label similarity in latent subspace for supervised learning, the instance pairwise similarity is ignored, which is important for locality structure preserving [51]. To further improve the intraclass compactness, a graph based regularization term is introduced as follows:

$$\sum_{i,j=1,i\neq j}^{n}\|\mathbf{W}\mathbf{X}_{:,i}-\mathbf{W}\mathbf{X}_{:,j}\|^2\mathbf{S}_{i,j}=2Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T), \tag{10}$$

where $\mathbf{S}=(\mathbf{S}_{i,j})_{n\times n}$ is the intraclass similarity graph of training data $\mathbf{X}$, and $\mathbf{L}=\mathbf{D}-\mathbf{S}$ is the Laplacian matrix. $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{i,i}=\sum_j\mathbf{S}_{i,j}$. The similarity matrix $\mathbf{S}$ is defined by Gaussian kernel function as follows:

$$\mathbf{S}_{i,j}=\begin{cases}\exp(-\|\mathbf{X}_{:,i}-\mathbf{X}_{:,j}\|^2/\delta^2) & \text{if } \mathbf{Y}_{:,i}=\mathbf{Y}_{:,j},\\ 0, & \text{otherwise,}\end{cases} \tag{11}$$

where δ is a bandwidth parameter, which is set as the average distance between all pairs of training samples. Eq. (10) uses the instance relations from training data space. The effect of minimizing Eq. (10) is locality preserving, i.e., the close samples in original training data space are enforced to be close as well after transformation, which can further improve the feature intraclass compactness. Combining (10) and (9), and the final objective function of PRDR is as follows:

$$\begin{aligned}
&\min_{\mathbf{W},\mathbf{V}}\frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{V}-\mathbf{Y}^T\mathbf{Y}\|_F^2 \\
&\qquad+\frac{\lambda_3}{2}Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) \\
&\text{s.t. } \{\|\mathbf{V}_{:,i}\|\}_{i=1}^{n}=1,
\end{aligned} \tag{12}$$

where λ_1, λ_2, and λ_3 are balance parameters. Fig. 2 illustrates the overall framework of our proposed method. The PRDR model jointly learns the transform matrix $\mathbf{W}$ and latent representation $\mathbf{V}$ by preserving the label and instance relations in latent feature space, which is characterized by

label distance and an intraclass compactness graph. Different with other methods, PRDR does not learn the projection from $\mathbf{V}$ to $\mathbf{Y}$. It focuses on the label and instance relations to guide regression learning, which utilizes the information from two channels (i.e., label space and training data space). The dimensionality of subspace can be arbitrary in PRDR, allowing it to learn latent representation with more flexibility. Besides, PRDR only contains two unknown variables which makes it efficient to be optimized.

B. Solution to PRDR

To directly solve the PRDR model (12) is difficult, since the overall model is nonconvex. We provide an iterative algorithm based on ADMM framework which is an effective tool for constrained optimization problems [52]–[54].

To make the variables in (12) separable, we first introduce an auxiliary variable $\mathbf{U}$ and rewrite it as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{V},\mathbf{U}} \frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{U}-\mathbf{Y}^T\mathbf{Y}\|_F^2 \\ +\frac{\lambda_3}{2}Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) \\ \text{s.t. } \{\|\mathbf{V}_{:,i}\|\}_{i=1}^n=1, \mathbf{V}=\mathbf{U}.\end{aligned} \tag{13}$$

The augmented Lagrangian function of problem (13) is

$$\begin{aligned}\mathcal{L}_\mu =&\frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{U}-\mathbf{Y}^T\mathbf{Y}\|_F^2 \\ &+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_3}{2}Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) \\ &+Tr\left(\mathbf{Z}^T(\mathbf{V}-\mathbf{U})\right)+\frac{\mu}{2}\|\mathbf{V}-\mathbf{U}\|_F^2,\end{aligned} \tag{14}$$

where $\mathbf{Z}$ is Lagrange multiplier and $\mu>0$ is a penalty factor.

Step 1 (*Update* $\mathbf{W}$): Fix other variables and update $\mathbf{W}$ by solving the following problem:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_3}{2}Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T). \tag{15}$$

This is a smooth and convex problem. By setting its derivative w.r.t $\mathbf{W}$ to zero, i.e.,

$$-(\mathbf{V}-\mathbf{W}\mathbf{X})\mathbf{X}^T+\lambda_1\mathbf{W}+\lambda_3\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T=0. \tag{16}$$

We can get its closed-form solution as follows:

$$\tilde{\mathbf{W}}=\mathbf{V}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T+\lambda_3\mathbf{X}\mathbf{L}\mathbf{X}^T+\lambda_1\mathbf{I})^{-1}. \tag{17}$$

Step 2 (*Update* $\mathbf{V}$): Fix other variables and update $\mathbf{V}$ by solving the following problem:

$$\begin{aligned}\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{U}-\mathbf{Y}^T\mathbf{Y}\|_F^2 \\ +\frac{\mu}{2}\|\mathbf{V}-\mathbf{U}+\mathbf{Z}/\mu\|_F^2.\end{aligned} \tag{18}$$

With the same optimization strategy of $\mathbf{W}$, we can obtain its solution as follows:

$$\mathbf{V}=\left((1+\mu)\mathbf{I}+\lambda_2\mathbf{U}\mathbf{U}^T\right)^{-1}\mathbf{G}, \tag{19}$$

Algorithm 1 algorithm for solving PRDR

Input: Training matrix $\mathbf{X}\in\mathbb{R}^{m\times n}$, label matrix $\mathbf{Y}\in\mathbb{R}^{c\times n}$, parameters λ_1, λ_2, λ_3 and d.
Output: Transform matrix $\mathbf{W}$.

1: Initialization: $\mathbf{W}$ and $\mathbf{V}$ with random values, $\mathbf{U}=\mathbf{V}$, $\mathbf{Z}=\mathbf{0}, \mu_{\max}=10^5, \mu=1, \rho=1.1$.
2: Compute similarity graph $\mathbf{S}$ by rule (11) and its Laplacian matrix $\mathbf{L}$.
3: Compute $\mathbf{H}=\mathbf{X}^T(\mathbf{X}\mathbf{X}^T+\lambda_3\mathbf{X}\mathbf{L}\mathbf{X}^T+\lambda_1\mathbf{I})^{-1}$.
4: **while** not converged **do**
5: Update $\mathbf{W}$ by $\mathbf{W}=\mathbf{V}\mathbf{H}$;
6: Update $\mathbf{V}$ by $\mathbf{V}=Norm(\tilde{\mathbf{V}})$, where $\tilde{\mathbf{V}}=((1+\mu)\mathbf{I}+\lambda_2\mathbf{U}\mathbf{U}^T)^{-1}(\mathbf{W}\mathbf{X}+\lambda_2\mathbf{U}\mathbf{Y}^T\mathbf{Y}+\mu\mathbf{U}-\mathbf{Z})$ and $Norm(\cdot)$ is the column normalization operator;
7: Update $\mathbf{U}$ by $\mathbf{U}=(\mu\mathbf{I}+\lambda_2\mathbf{V}\mathbf{V}^T)^{-1}(\lambda_2\mathbf{V}\mathbf{Y}^T\mathbf{Y}+\mu\mathbf{V}+\mathbf{Z})$;
8: Update $\mathbf{Z}$ and μ by $\mathbf{Z}=\mathbf{Z}+\mu(\mathbf{V}-\mathbf{U})$ and $\mu=\min(\rho\mu,\mu_{\max})$;
9: **end while**
10: **return** $\mathbf{W}$.

where $\mathbf{G}=\mathbf{W}\mathbf{X}+\lambda_2\mathbf{U}\mathbf{Y}^T\mathbf{Y}+\mu\mathbf{U}-\mathbf{Z}$. Due to the column normalization constraint $\{\|\mathbf{V}_{:,i}\|\}_{i=1}^n=1$, the optimal $\tilde{\mathbf{V}}$ is

$$\tilde{\mathbf{V}}=[\tilde{\mathbf{V}}_{:,1},\tilde{\mathbf{V}}_{:,2},...,\tilde{\mathbf{V}}_{:,n}], \tag{20}$$

where $\tilde{\mathbf{V}}_{:,i}=\mathbf{V}_{:,i}/\sqrt{\sum_{k=1}^d\mathbf{V}_{k,i}^2}$.

Step 3 (*Update* $\mathbf{U}$): Fix other variables and update $\mathbf{U}$ by solving the following problem:

$$\min_{\mathbf{U}} \frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{U}-\mathbf{Y}^T\mathbf{Y}\|_F^2+\frac{\mu}{2}\|\mathbf{V}-\mathbf{U}+\mathbf{Z}/\mu\|_F^2. \tag{21}$$

This is also a smooth and convex optimization problem, which can be efficiently solved by a closed-form solution as below:

$$\tilde{\mathbf{U}}=(\mu\mathbf{I}+\lambda_2\mathbf{V}\mathbf{V}^T)^{-1}(\lambda_2\mathbf{V}\mathbf{Y}^T\mathbf{Y}+\mu\mathbf{V}+\mathbf{Z}). \tag{22}$$

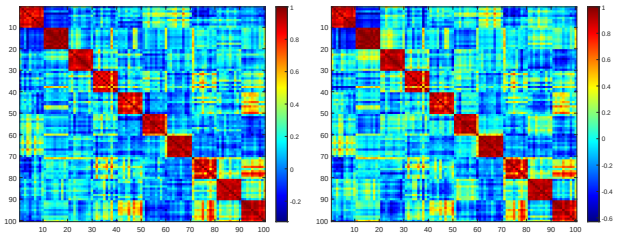
Step 4 (*Update* $\mathbf{Z}$ *and* μ):

$$\begin{aligned}\tilde{\mathbf{Z}}&=\mathbf{Z}+\mu(\mathbf{V}-\mathbf{U}),\\ \tilde{\mu}&=\min(\rho\mu,\mu_{\max}),\end{aligned} \tag{23}$$

where $\rho>1$ and $\mu_{\max}$ are constants. By performing step 1-5 iteratively, the objective function can be gradually minimized until convergence or reaching the maximum number of iterations. It is noticed that, in step 1, $\mathbf{X}^T(\mathbf{X}\mathbf{X}^T+\lambda_3\mathbf{X}\mathbf{L}\mathbf{X}^T+\lambda_1\mathbf{I})^{-1}$ is fixed in iterations, thus we can compute and store it in advance for faster speed. The algorithm for solving the PRDR model is summarized in Algorithm 1. Once the optimal transform matrix $\mathbf{W}$ is learned by Algorithm 1, we can directly use $\mathbf{W}$ to obtain the features $\mathbf{W}\mathbf{X}$. Then, the nearest neighbor classifier is used for classification.

C. Complexity and Convergence Analysis

Apart from classification accuracy, computational complexity is also an important issue for evaluating an algorithm [55]. From Algorithm 1, the time cost contains two



(a) (b)

Fig. 3. Visualization of matrix $\mathbf{V}^T\mathbf{V}$ of (a) PRDR and (b) R-PRDR.

parts: the computation outside the loop and iterative steps inside the loop. For similarity graph $\mathbf{S}$, the computational complexity is $O(n^2)$. For matrix $\mathbf{H}$, its main time cost is the inverse operation which takes $O(m^3)$. In the loop, matrix $\mathbf{H}$ is only calculated once outside the loop, and the computation for $\mathbf{W}$ equals multiplication operation which is very simple and can be ignored. For $\mathbf{V}$ and $\mathbf{U}$, the main computational costs are also the inverse operations with $O(d^3)$ complexity. The computations of Lagrange multiplier $\mathbf{Z}$ and penalty factor μ are also very simple, and thus their computational costs can be ignored. Thus, we can conclude that the total computational complexity of PRDR is about $O(n^2 + m^3 + \tau d^3)$ if there are τ iterations.

As presented in previous section, the classical ADMM framework is adopted to solve the proposed model. The overall problem (12) w.r.t all unknown variables is non-convex, and it is difficult to theoretically prove the strong convergence property of Algorithm 1. In this section, we present a proof of its weak convergence to a local minimum. It is worth noting that Karush-Kuhn-Tucker (KKT) conditions are the necessary conditions for a constrained local optimal solution, and any converging point must be a KKT point [56]. The following theorem guarantees a weak convergence property of the proposed optimization algorithm.

Theorem 1. *Let $\{\theta^t\}_{t=1}^{\infty}$ be a solution sequence generated by Algorithm 1 with $\theta^t = (\mathbf{W}^t, \mathbf{V}^t, \mathbf{U}^t, \mathbf{Z}^t)$. Suppose the solution sequence $\{\theta^t\}_{t=1}^{\infty}$ is bounded and $\lim_{t\to\infty}(\theta^{t+1} - \theta^t) = 0$, then every limit point of $\{\theta^t\}_{t=1}^{\infty}$ satisfies the KKT conditions. Whenever $\{\theta^t\}_{t=1}^{\infty}$ converges, it converges to a KKT point.*

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

D. Discussion and Extension of PRDR

Instead of regressing the training data to label matrix, PRDR leverages a latent subspace to facilitate the regression learning by considering the pairwise instance relations in original feature space and label space simultaneously. The pairwise label distance is used as supervised information and transferred into latent space to guide the regression. In specific, for training sample pair $(\mathbf{x}_i, \mathbf{x}_j)$, it minimizes the following problem with $\{\|\mathbf{V}_{:,i}\|\}_{i=1}^n = 1$,

$$\begin{aligned}
&(\mathbf{V}_{:,i}^T\mathbf{V}_{:,j} - \mathbf{Y}_{:,i}^T\mathbf{Y}_{:,j})^2 \\
&= \left(\frac{\mathbf{V}_{:,i}^T\mathbf{V}_{:,j}}{\|\mathbf{V}_{:,i}\| \cdot \|\mathbf{V}_{:,j}\|} - \frac{\mathbf{Y}_{:,i}^T\mathbf{Y}_{:,j}}{\|\mathbf{Y}_{:,i}\| \cdot \|\mathbf{Y}_{:,j}\|}\right)^2 \\
&= \left(\cos(\theta_{ij}^V) - \cos(\theta_{ij}^Y)\right)^2 \\
&= \left(\cos(\theta_{ij}^V) - 1\right)^2 I(\mathbf{Y}_{:,i} = \mathbf{Y}_{:,j}) \\
&\quad + \left(\cos(\theta_{ij}^V) - 0\right)^2 I(\mathbf{Y}_{:,i} \neq \mathbf{Y}_{:,j}),
\end{aligned} \tag{24}$$

where $\mathbf{V}_{:,i}$ and $\mathbf{V}_{:,j}$ are the latent features of $\mathbf{x}_i$ and $\mathbf{x}_j$, $\|\mathbf{Y}_{:,i}\| = 1$, and $I(\cdot)$ is an indicator operator that $I(g) = 1$ if condition g holds and otherwise 0. Under the observation of $\mathbf{Y}_{:,i}^T\mathbf{Y}_{:,j} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class and otherwise 0, the $\|\mathbf{V}^T\mathbf{V} - \mathbf{Y}^T\mathbf{Y}\|_F^2$ term in PRDR model (9) reveals the pairwise cosine distance of latent feature $\mathbf{V}$ and label matrix $\mathbf{Y}$ is in one-to-one correspondence, which preserves the semantic similarity of training data and enforces the learned regression target $\mathbf{V}$ to be semantically discriminative. In other words, PRDR directly enforces the training samples from the same class to be close as much as possible (i.e., enforce the pairwise cosine distance to be 1) after projection, which improves the intraclass compactness in regression.

As mentioned previously, some one-step and two-step based regression methods directly regress the training data $\mathbf{X}$ or latent features $\mathbf{V}$ to the rigid binary label matrix $\mathbf{Y}$ (i.e., minimize $\|\mathbf{WX} - \mathbf{Y}\|_F^2$ or $\|\mathbf{WV} - \mathbf{Y}\|_F^2$), which may be not suitable as regression target and lead to overfitting. In standard least squares regression, the Euclidean distance of regression targets between two samples from different classes is a definite constant, i.e., $\sqrt{2}$, no matter the data dimensionality. Such rigid Euclidean distance constraint may harm the discriminative learning capacity [32]. DLSR and ReLSR adopt soft label technique to relax the label matrix, in which the regression target is adaptively learned. Although the margins between true and false classes are enlarged, however, the distances of samples from the same class may be also enlarged after projection in these soft label based methods [32]. That is said, the semantic correlation of samples from the same class is weakened after relaxation, and the discriminative power of transformation matrix will certainly be compromised. Besides, these methods regress each sample to the pre-defined label vectors separately, which may ignore the data relationships and easily lead to overfitting [43]. Different with these existing methods, on the one hand, PRDR regresses the training data to a latent subspace $\mathbf{V}$ rather than label space, which discards the rigid Euclidean distance constraint. The dimensionality of $\mathbf{V}$ can be arbitrary and $\mathbf{V}$ may be more flexible for regression. On the other hand, PRDR considers the pairwise relations among all training samples, which are ignored by previous regression methods, to improve the intraclass compactness and interclass separability. Although some methods like ICS_DLSR and GLRRDLR also consider the intraclass compactness, they generally adopt classwise constraints on

TABLE I
GENERAL STATISTICS OF SIX DATASETS USED IN EXPERIMENTS.

Datasets	# Classes	# Instances	# Features
PIE	68	11554	1024
LFW	86	1251	1024
USPS	10	9298	256
COIL100	100	7200	1024
Caltech101	101	8731	4096
AwA	50	30733	4096

Fig. 4. Some typical images from (a) PIE, (b) LFW, (c) USPS, (d) COIL100, and (e) Caltech101 datasets (The images of AwA are unavailable because of copyright restrictions).

the features of each class. The classwise constraints may have a limited effect on reducing the distances between samples from the same class, while the regularization term in PRDR directly minimizes the pairwise intraclass distances, which has a direct effect on improving the discrimination of regression. Furthermore, the regularization term in PRDR p-reserves the local data structure, which is helpful to alleviate the overfitting problem [23].

It is known that the maximum of $\cos(\theta_{ij})$ could be 1 and the minimum could be -1. The smaller the cosine similarity value is, the larger the angle or difference between $\mathbf{V}_{:,i}$ and $\mathbf{V}_{:,j}$ is. In PRDR, the cosine similarity value is forced to be 1 for two samples from the same class and 0 from different classes. For optimal interclass separability, the cosine value (i.e., $\mathbf{V}_{:,i}^T\mathbf{V}_{:,j}$) is expected to be -1 for two samples from different classes, which pulls the two samples away from each other as much as possible. Therefore, the PRDR model can be extended to a more discriminative version by further enlarging the distances between different classes, i.e.,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{V}} \frac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{X}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_3}{2}Tr(\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) \\ + \frac{\lambda_2}{2}\|\mathbf{V}^T\mathbf{V}-\mathcal{S}(\mathbf{Y}^T\mathbf{Y})\|_F^2 \\ \text{s.t. } \{\|\mathbf{V}_{:,i}\|\}_{i=1}^n = 1,\end{aligned} \tag{25}$$

where $\mathcal{S}(\cdot)$ is an operator that sets element 0 to -1. We denote the extended model (25) as R-PRDR. It can be

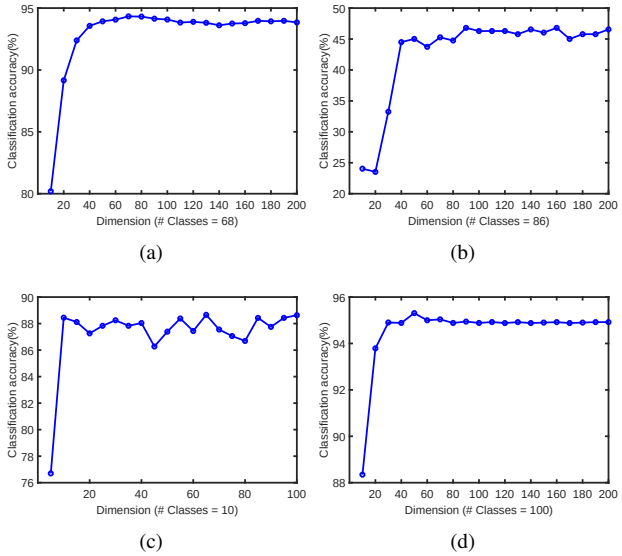


Fig. 5. Classification accuracies (%) versus the dimensionality of latent space on (a) PIE, (b) LFW, (c) USPS and (d) COIL100 datasets. For the four datasets, we randomly select 20, 10, 20, and 20 images per subject for training and the rest for testing, respectively.

observed that the only difference between PRDR and R-PRDR is the matrix $\mathbf{Y}^T\mathbf{Y}$. Thus R-PRDR can be solved by Algorithm 1 as well. In specific, we only need to replace the $\mathbf{Y}^T\mathbf{Y}$ by $\mathcal{S}(\mathbf{Y}^T\mathbf{Y})$ in optimization. The computational complexity and convergence property of R-PRDR are also the same as PRDR. The matrix $\mathbf{V}^T\mathbf{V}$ in PRDR and R-PRDR characterizes the pairwise cosine distances of all training samples. To more clearly illustrate the effect of proposed methods, the matrix $\mathbf{V}^T\mathbf{V}$ of PRDR and R-PRDR are visualized in Fig. 3. Obviously, the cosine distances between samples from the same class are small (i.e., the diagonal blocks) and the distances between samples from different classes are large, which indicates the effectiveness of PRDR and R-PRDR to improve the intraclass similarity and interclass separability.

IV. EXPERIMENTS

In this section, we conduct experiments on several popular datasets, including PIE[1], LFW[2], USPS[3], COIL100[4], Caltech101[5] and AwA[6], to validate the effectiveness of our proposed PRDR and R-PRDR. Some state-of-the-art related methods for classification are used for fair comparison, including LRC [12], CRC [14], LRRR [43], SLRR [43], DLSR [18], ReLSR [22], ICS_DLSR [32], RDR [47], RSL-DA [57], RLSL [44] and GLRRDLR [30]. For LRC and CRC, the label of a test sample is determined by minimum classwise regression error. For other methods and PRDR,

[1]https://www.ri.cmu.edu/project/pie-database/
[2]http://vis-www.cs.umass.edu/lfw/
[3]http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html
[4]https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php
[5]http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[6]https://cvml.ist.ac.at/AwA/

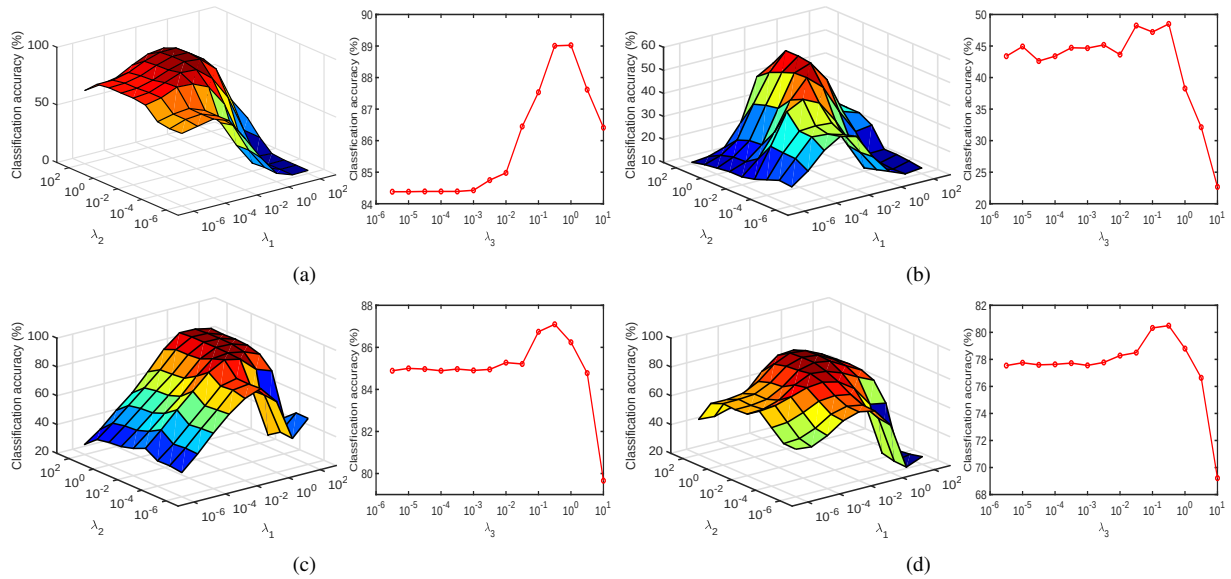


Fig. 6. Classification accuracies (%) of PRDR versus λ_1, λ_2, and λ_3 on (a) PIE, (b) LFW, (c) COIL100 and (d) Caltech101 datasets. For each dataset, the number of training samples per subject is 10.

TABLE II
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON PIE DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR METHOD RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	10	15	20	25
LRC	75.63±0.88	85.01±0.66	90.21±0.37	92.54±0.16
CRC	85.98±0.64	90.68±0.81	92.95±0.28	94.20±0.24
LRRR	85.63±0.75	89.71±0.33	92.61±0.80	93.89±0.36
SLRR	88.03±0.62	91.54±0.51	93.55±0.32	94.27±0.23
DLSR	87.23±0.69	91.81±0.62	93.76±0.44	94.61±0.29
ReLSR	87.48±0.48	91.79±0.53	93.74±0.30	94.98±0.28
RLSL	88.17±0.57	91.95±0.56	93.71±0.26	94.85±0.17
RDR	86.85±0.62	91.28±0.33	93.77±0.30	94.99±0.38
RSLDA	78.82±1.09	85.03±0.70	89.22±0.45	91.75±0.19
ICSR	88.78±0.51	92.12±0.43	94.14±0.31	95.05±0.23
GLRR	88.29±0.60	92.46±0.58	94.36±0.18	95.70±0.27
PRDR	89.31±0.67	92.45±0.42	94.29±0.23	95.59±0.23
R-PRDR	**89.53**±0.61	**92.74**±0.44	**94.89**±0.24	**96.09**±0.21

TABLE III
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON LFW DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR METHOD RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	7	8	9	10
LRC	35.83±1.74	37.44±1.26	39.56±1.89	40.27±2.37
CRC	36.66±1.94	37.87±1.53	38.83±1.59	39.80±2.35
LRRR	38.58±1.59	39.33±1.25	40.62±2.25	41.56±1.67
SLRR	31.37±1.02	32.54±1.16	33.51±1.12	33.80±2.00
DLSR	35.12±2.75	36.34±1.47	38.72±2.20	39.80±2.27
ReLSR	36.50±1.65	38.17±1.42	40.19±2.16	41.74±2.15
RLSL	41.51±2.72	42.38±1.42	44.47±1.88	45.19±1.56
RDR	36.41±1.90	37.56±1.64	38.88±1.98	40.87±1.98
RSLDA	30.21±2.25	31.48±1.57	32.76±2.45	33.24±2.14
ICSR	42.13±1.45	42.91±2.07	45.12±2.08	45.93±1.74
GLRR	43.62±2.63	44.51±1.14	46.32±1.55	48.21±2.31
PRDR	43.70±2.08	44.55±2.19	46.27±1.69	48.13±1.73
R-PRDR	**43.87**±2.74	**44.82**±1.68	**46.87**±1.92	**48.44**±2.22

a transform matrix $\mathbf{W}$ is learned and the nearest neighbor is used for classification on transformed data $\mathbf{WX}$ [16], [30], [32]. In these methods, DLSR, ReLSR, ICS_DLSR and GLRRDLR are one-step based methods which directly regress the training data to the label space. LRRR, SLRR, RDR, RLSL, RSLDA and PRDR can be regarded as two-step based methods that leverage a middle latent subspace for classification. For each experiment on all datasets, these methods are repeated 10 times with random training and test data partitions. We report the mean accuracies with standard deviations for comparison. All experiments are implemented on MATLAB R2017b with Win10 system, Inter Core i7-8550 CPU and 8GB RAM. TABLE I lists the main information of the six datasets and Fig. 4 shows some example samples. The MATLAB code for proposed method is available at https://github.com/ChZhang96/PRDR.

A. Parameter Analysis

1) The Dimensionality of Latent Space: In PRDR, the dimensionality d of latent space $\mathbf{V}$ is a hyperparameter and difficult to determine, since it can range from zero to infinity. A small d is not enough to preserve the discriminative information, while a large d will increase the computational and storage costs. According to [43], [44], the dimension d can be set around c, where c is the number of classes. In our experiments. we observe that PRDR can achieve satisfactory performance where the value of d approximately equals to c. Fig. 5 shows the classification accuracies of PRDR versus d on PIE, LFW, USPS and COIL100 datasets. We can see that the changes of accuracy are not obvious when $d > c$ and the peak is achieved if d is around c. Thus, in our experiments, the dimensionality d of PRDR and R-PRDR is fixed as c for all datasets.

TABLE IV
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON USPS DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR METHOD RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	10	30	50	70
LRC	80.76±2.00	87.79±0.64	90.13±0.45	90.93±0.41
CRC	81.77±1.36	87.37±0.42	89.18±0.54	89.93±0.30
LRRR	78.67±1.46	86.73±0.70	88.75±0.62	89.28± 0.43
SLRR	81.34±2.92	88.67±0.59	89.88±0.36	90.39±0.69
DLSR	81.07±2.12	86.66±1.19	88.78±0.55	90.40±0.67
ReLSR	84.28±1.53	88.80±0.84	89.97±0.40	91.15±0.50
RLSL	84.78±1.12	86.91±1.39	88.48±0.68	89.05±0.47
RDR	84.21±1.49	88.93±0.92	89.45±0.68	90.33±0.39
RSLDA	78.00±2.26	83.61±0.79	86.51±0.49	87.76±0.65
ICSR	86.10±1.23	89.58±0.64	90.23±0.31	91.06±0.57
GLRR	86.02±1.15	90.41±0.68	91.05±0.42	91.95±0.39
PRDR	86.11±1.23	90.42±0.64	90.98±0.31	91.79±0.47
R-PRDR	**87.43**±1.13	**91.03**±0.57	**91.87**±0.32	**92.48**±0.53

TABLE V
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON COIL100 DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR METHOD RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	10	15	20	25
LRC	81.03±0.94	88.58±0.34	90.41±0.47	92.89±0.19
CRC	74.56±0.40	79.83±0.54	82.97±0.45	84.76±0.50
LRRR	80.35±0.41	86.27±0.58	89.05±0.36	91.88±0.57
SLRR	81.61±0.95	87.33±0.67	90.57±0.32	92.96±0.47
DLSR	82.68±0.61	87.81±0.49	91.20±0.75	93.27±0.51
ReLSR	85.94±0.47	90.45±0.52	93.44±0.45	94.95±0.44
RLSL	87.39±0.49	90.88±0.56	93.34±0.44	94.62±0.44
RDR	84.78±0.55	89.12±0.58	93.13±0.39	94.56±0.27
RSLDA	84.73±0.37	89.59±0.59	93.23±0.45	93.82±0.47
ICSR	87.02±0.87	92.36±0.55	93.97±0.49	94.84±0.38
GLRR	87.27±0.50	91.38±0.68	94.08±0.47	95.56±0.41
PRDR	88.75±0.49	92.48±0.53	94.51±0.34	95.64±0.39
R-PRDR	**88.92**±0.44	**92.61**±0.60	**94.98**±0.37	**96.02**±0.46

2) Regularization Parameters Sensitivity Analysis: In PRDR, there are three regularization parameters, i.e., λ_1, λ_2 and λ_3, which influence the performance of algorithm. To analyze the parameter sensitivity of PRDR, we first define a candidate set $\{10^{-6}$,10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, $10^{2}\}$ for λ_1 and λ_2, and $\{10^{-6}$,10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, $10^{1}\}$ for λ_3. Then PRDR is performed on PIE, LFW, COIL100 and Caltech101 datasets with different combinations of the three parameters. Fig. 6 shows the changes of classification accuracy versus the three parameters on four datasets. It can be observed that all three parameters influence the classification performance. Compared with λ_1, the proposed method is less sensitive to parameter λ_2. We can further find that the proposed method can achieve a satisfactory classification result when λ_1 and λ_2 are located in $[10^{-3}, 1]$ and $[10^{-2}, 10^{1}]$, respectively. The performance degrades when λ_3 is extremely small or large, which indicates that the locality relationships between data points is beneficial to preserve intraclass similarity and improve the classification accuracy. When λ_3 is tuned from 0.1 to 1, our method can always achieve the best performance. In our experiments, we first find an optimal λ_3 due to the robustness by fixing other parameters in candidate range. Then, by fixing the value of λ_3, the optimal λ_1 and λ_2 are searched with a fixed step in their own candidate range.

B. Classification Performance

Six popular datasets listed in TABLE I are used to evaluate the performance of our proposed methods.

1) CMU PIE Face Dataset: PIE dataset contains total 41,368 images of 68 subjects, collected under various facial poses, illumination conditions and expressions. In this experiment, all methods are compared on a subset of PIE which contains 11,554 samples of 68 classes with 5 poses. All images are resized to 32×32 pixels and reshaped to 1024 dimensional vectors [44], [57], [58]. We randomly select 10, 15, 20 and 25 samples for training and the rest for testing. The mean classification accuracies with standard deviations are listed in TABLE II. We can see that DLSR and ReLSR outperform LRRR, which indicates that the ϵ-dragging and dynamic regression target learning techniques are effective to improve the performance. ICSR also obtains comparable results which relaxes the label matrix and considers the classwise data similarity in projection learning. PRDR utilizes the local relationship information from training data space and distance information from label space simultaneously, and it outperforms ReLSR, RLSL and ICSR. GLRR combines the dynamic regression target technique and classwise low-rank constraint, and it can achieve competitive and similar performance with PRDR. R-PRDR extends PRDR by further enlarging the distances between true and false classes, and it achieves the best classification accuracy in all competing methods.

2) LFW Face Dataset: LFW is a challenging large-scale wild dataset for unconstrained face recognition whose images are collected from the web. In this experiment, we use a subset of LFW which contains 1251 samples of 86 subjects to evaluate these different methods. Each class has 11-20 images and all images are resized to 32×32 pixels [32]. We randomly select 7, 8, 9 and 10 images per subject as training set and the rest as test set. The classification rates of different methods are listed in TABLE III. Under all training protocols, R-PRDR outperforms other methods. Due to the big challenge of LFW, the performance of DLSR, ReLSR, RDR and RSLDA is not desirable. RLSL and GLRR obtain comparable results with PRDR which are more robust than DLSR and ReLSR. Under the four training protocols, the average accuracy improvements of R-PRDR over RLSL and ReLSR are 2.61% and 6.92%, respectively. These experimental results demonstrate the superiority of the proposed method on challenging datasets.

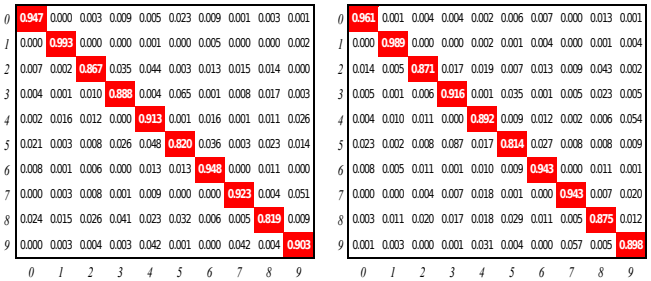
3) USPS Dataset: USPS is a handwritten digits dataset containing 9298 images of 10 classes, i.e., 0-9. All images are 16×16 grayscale pixels [18]. In our experiments, 10, 30, 50 and 70 images per subject are randomly selected for training and the rest for testing. The mean classification rates of PRDR, R-PRDR and baselines are listed in TABLE IV. It is obvious that our PRDR and R-PRDR achieve the best

TABLE VI
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON CALTECH101 DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	10	15	20	25
LRC	75.84±0.59	78.17±0.37	78.89±0.37	79.22±0.25
CRC	76.61±0.51	79.11±0.38	79.97±0.29	80.63±0.36
LRRR	74.62±0.74	76.67±0.44	78.38±0.28	79.34±0.26
SLRR	74.71±0.30	78.46±0.10	80.58±0.31	82.46±0.53
DLSR	78.65±0.39	80.29±0.43	81.28±0.47	81.64±0.55
ReLSR	79.48±0.53	81.26±0.49	82.17±0.31	82.47±0.55
RLSL	79.17±0.48	81.39±0.40	82.54±0.36	83.06±0.35
RDR	80.12±0.41	81.32±0.56	82.14±0.37	82.46±0.42
RSLDA	74.36±0.46	76.30±0.66	77.43±0.42	78.35±0.28
ICSR	79.53±0.45	82.07±0.29	82.96±0.24	83.19±0.44
GLRR	80.42±0.49	81.88±0.33	82.71±0.40	83.05±0.42
PRDR	80.75±0.42	82.29±0.36	82.63±0.24	83.85±0.46
R-PRDR	**80.98**±0.51	**82.64**±0.33	**83.09**±0.31	**83.97**±0.39

TABLE VII
MEAN CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON AWA DATASET. NOTE: ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	10	30	50	70
LRC	38.90±0.89	50.23±0.60	57.55±0.44	58.63±0.67
CRC	35.90±0.76	47.25±0.37	54.09±0.31	57.71±0.42
LRRR	42.94±1.01	50.14±0.53	52.58±0.51	53.52±0.42
SLRR	32.32±1.24	41.65±0.53	50.68±0.04	54.80±0.35
DLSR	45.18±1.09	52.96±0.48	55.43±0.59	56.85±0.36
ReLSR	45.61±1.12	53.73±0.69	56.52±0.40	57.89±0.31
RLSL	46.45±1.02	53.88±0.38	56.77±0.48	58.26±0.41
RDR	46.16±1.03	53.78±0.32	56.06±0.23	57.00±0.40
RSLDA	42.89±0.68	49.31±0.39	52.70±0.36	54.51±0.43
ICSR	44.06±1.02	52.86±0.37	56.15±0.45	57.75±0.42
GLRR	47.36±0.93	55.71±0.22	58.78±0.27	60.26±0.35
PRDR	48.13±0.89	56.24±0.38	59.01±0.39	61.08±0.31
R-PRDR	**48.59**±0.85	**56.59**±0.34	**59.58**±0.30	**61.33**±0.28



(a) PRDR confusion matrix (b) R-PRDR confusion matrix

Fig. 7. The confusion matrix of PRDR and R-PRDR on USPS dataset (50 samples per subject are selected for training).

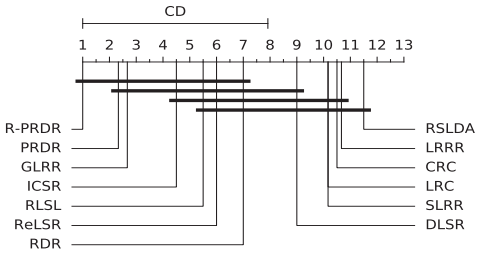
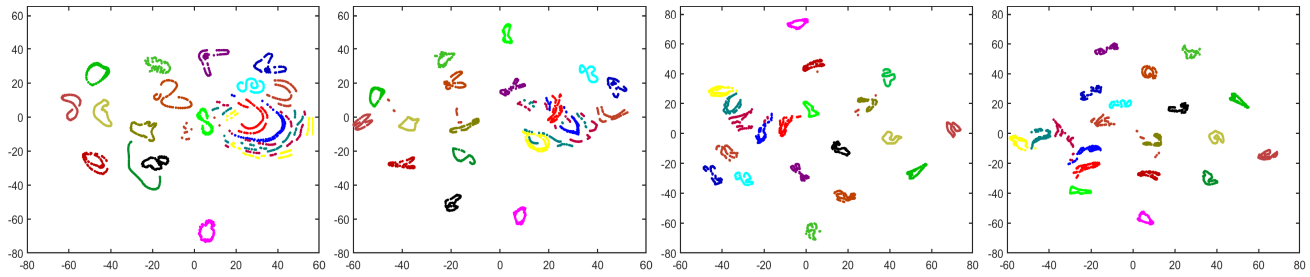


Fig. 10. CD diagram of different methods with significance level $\alpha = 0.1$.

classification performance among all competing algorithms. Fig. 7 shows the confusion matrix of PRDR and R-PRDR on USPS dataset with 50 images per subject used for training, in which the classification accuracy on each class locates along the diagonal. Specifically, for the digit "1", PRDR can achieve over 99% recognition accuracy. The classification accuracy of R-PRDR on digit "1" is slightly lower than that of PRDR, however, R-PRDR obtains 5.6% accuracy improvement on digit "8" than PRDR.

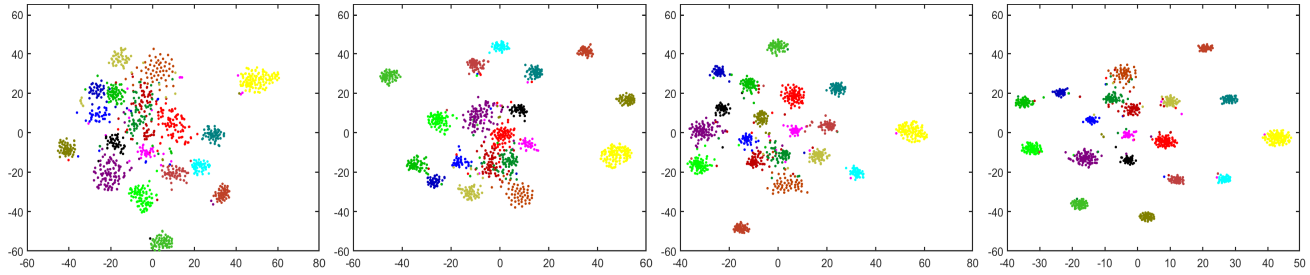
4) COIL100 Dataset: COIL100 is an object image dataset containing 7200 images of 100 objects. The images are captured at pose intervals of 5 degrees. All images are resized to 32×32 pixels [26], [44]. For each class, we randomly select 10, 15, 20 and 25 images for training and the rest for testing. TABLE V shows the classification results of different methods on COIL100 dataset. From TABLE V, we can observe that our PRDR and R-PRDR are superior to other baseline approaches. To better illustrate the effect the PRDR, the high-dimensional original data and low-dimensional latent features learned by PRDR are visualized in Fig. 8 using t-SNE [59]. Total 1440 samples from the first 20 classes, including training samples and test samples, are visualized. Fig. 8(a) visualizes the original data. It can be obviously observed that the samples of some classes are scattered messily. Fig. 8(b), (c) and (d) visualize the PRDR features when 10, 30 and 50 images per subject are used for training, respectively. When the training scale per subject is 50, our PRDR can achieve 98.55% classification accuracy. Obviously, the intraclass compactness and interclass separability are greatly enhanced in the learned features of PRDR. From Fig. 8(b) to (d), the projection gets more discriminative with the increase of training samples. In particular, samples from the same class are obviously clustered together, which demonstrates that the proposed method can greatly enhance the intraclass similarity and pull data points to their own subspace.

5) Caltech101 Dataset: Caltech101 is a widely used image dataset for object recognition which contains over 9000 images from 101 objects categories and a background category. Following [44], the 4096-dimensional DeCAF6 deep features of 8791 images in 101 classes are used for experiments. The DeCAF6 deep features are available at https://sites.google.com/site/crossdataset/home/files. To improve the computational efficiency, PCA is used as preprocessing step to preserve the 98% energy of data. 10, 15, 20 and 25 samples per subject are randomly chosen for training and the remaining for testing. We report the mean classification accuracies and standard deviations for all algorithms in TABLE VI. In general, PRDR and R-PRDR perform the best among all competing approaches, which indicates that our proposed methods can flexibly leverage deep features. Total 1312 original samples from 20 classes and corresponding learned features are visualized in Fig. 9 by t-SNE. For each class, 20 samples are randomly



(a) Original data (b) PRDR features (No. Train=10) (c) PRDR features (No. Train=30) (d) PRDR features (No. Train=50)

Fig. 8. t-SNE visualization of original samples and learned features of PRDR on COIL100 dataset. The 1440 samples from the first 20 classes are visualized when 10, 30 and 50 images per subject are used for training, respectively. Both training samples and testing samples are visualized.



(a) Original data (b) RSLDA features (c) ReLSR features (d) PRDR features

Fig. 9. t-SNE visualization of (a) original data and (b) RSLDA, (c) ReLSR and (d) PRDR features on Caltech101 dataset. Total 1312 samples from 20 classes are selected for visualization and the training data size is 20 per subject.

selected to learn the discriminative projection matrix and both training samples and test samples are visualized. The features of RSLDA and ReLSR, which focus on enlarging the distances between different classes, are taken for comparison. Obviously, PRDR features have larger interclass scatter and smaller intraclass scatter, which indicates that PRDR can learn a more discriminant regression model.

6) AwA Dataset: AwA dataset contains over 30,000 images of 50 animals classes. Similar to the tests on Caltech101 dataset, the DeCAF6 deep features of AwA are adopted for experiments. Total 30,733 instances of 50 classes in 4096 dimensions are used and then PCA is applied to save 98% energy. We randomly select 10, 30, 50 and 70 instances per subject for training and the rest for testing. The mean classification accuracies and standard deviations of different methods are listed in TABLE VII. PRDR and R-PRDR outperform other methods. All the experimental results on PIE, LFW, USPS, COIL100, Caltech101 and AwA datasets demonstrate that the proposed PRDR and R-PRDR can outperform other regression methods for image classification with discriminative projection learning.

C. Statistical Significance

The Friedman test with a post-hoc test [60] is widely used to compare different algorithms on multiple datasets. In Friedman test, the null hypothesis states that all the algorithms are equivalent, and thus their average ranks should be equal. If the null hypothesis is rejected, we can proceed with a post-hoc Nemenyi test to find out those algorithms that significantly differ. If the average ranks of two algorithms differ by at least the critical difference (CD), the performance of the two algorithms is significantly different [60].

TABLE VIII
THE TIME COMPLEXITY OF DIFFERENT METHODS FOR TRAINING. ICSR AND GLRR DENOTE ICS_DLSR AND GLRRDLR RESPECTIVELY.

Method	Time Complexity	Method	Time Complexity
LRRR	$O(m^3+d^3)$	RLSL	$O(\tau(m^3+m^2n+d^3))$
SLRR	$O(\tau(m^3+d^3))$	RSLDA	$O(\tau(m^3+m^2n+d^3))$
DLSR	$O(\tau mn+m^2n)$	RDR	$O(\tau(m^3+mn^2+d^3))$
ReLSR	$O(\tau mn+m^2n)$	GLRR	$O(\tau(m^3+n^2))$
ICSR	$O(\tau(m^3+n^2))$	PRDR	$O(\tau d^3+m^3+n^2)$

In Section IV-B, total 13 algorithms are evaluated on six datasets with different train/test splits. For each algorithm, the average classification accuracy on each dataset is regarded as its final performance on this dataset. Fig. 10 shows the CD diagram of the 13 methods, where the average rank of each method locates along the axis. The methods in a group linked by a thick line are not significantly different. We can see that LRC, CRC, LRRR, SLRR and RSLDA have high ranks and they are significantly different with proposed PRDR and R-PRDR. R-PRDR achieves the lowest (best) rank, and PRDR ranks the second in all methods.

D. Time Cost

In Section III-C, we analyzed the computational complexity of the proposed algorithm. In this section, we conduct experiments on the PIE, LFW, USPS, COIL100, Caltech101 and AwA datasets to illustrate the efficiency of PRDR. In

TABLE IX
AVERAGE TRAINING TIME (SECOND) OF DIFFERENT METHODS ON PIE, LFW, USPS, COIL100, CALTECH101 AND AWA DATASETS WITH DIFFERENT NUMBERS OF TRAINING SAMPLES PER SUBJECT. NOTE: "#TR." MEANS THAT # SAMPLES PER SUBJECT ARE USED FOR TRAINING AND THE RESTING FOR TESTING, ICSR AND GLRR DENOTE THE ICS_DLSR AND GLRRDLR METHOD RESPECTIVELY.

Method	PIE			LFW			USPS			COIL100			Caltech101			AwA		
	10Tr.	20Tr.	25Tr.	7Tr.	8Tr.	10Tr.	10Tr.	30Tr.	50Tr.	10Tr.	20Tr.	25Tr.	10Tr.	20Tr.	25Tr.	10Tr.	30Tr.	50Tr.
LRRR	0.771	0.780	0.791	0.691	0.699	0.705	0.032	0.039	0.043	0.651	0.678	0.691	0.324	0.358	0.374	0.475	0.507	0.557
SLRR	7.987	8.554	8.885	7.789	7.816	8.017	0.292	0.328	0.398	7.282	7.429	7.633	3.684	3.805	3.924	4.708	5.176	5.394
DLSR	0.461	1.057	1.174	0.422	0.505	0.685	0.015	0.044	0.052	0.893	1.446	1.678	0.681	0.968	1.179	0.275	0.952	1.305
ReLSR	0.171	0.321	0.406	0.181	0.203	0.251	0.005	0.011	0.016	0.354	0.704	0.914	0.328	0.671	0.918	0.089	0.262	0.494
RLSL	12.68	15.52	16.89	13.24	13.60	14.81	0.486	0.548	0.582	13.43	18.71	21.34	8.684	10.87	12.37	9.032	13.91	16.73
RDR	4.453	5.994	7.396	3.642	3.855	4.256	0.167	0.285	0.539	5.172	11.34	17.44	2.629	6.420	11.62	2.309	5.853	10.58
RSLDA	3.105	4.920	5.831	2.994	3.221	3.724	0.075	0.121	0.163	4.098	7.029	9.032	2.617	4.226	5.812	2.180	4.564	7.051
ICSR	1.240	1.465	1.605	1.273	1.299	1.359	0.058	0.062	0.075	1.486	1.906	2.438	1.008	1.312	1.594	0.982	1.296	1.557
GLRR	2.581	4.412	5.276	2.910	3.231	3.777	0.089	0.134	0.199	5.404	9.363	10.67	4.480	6.126	7.979	1.693	3.632	5.722
PRDR	0.332	0.619	0.809	0.284	0.328	0.420	0.012	0.026	0.043	0.546	1.178	1.781	0.512	1.201	1.763	0.173	0.572	1.278
R-PRDR	0.306	0.610	0.818	0.288	0.339	0.427	0.012	0.022	0.048	0.557	1.196	1.799	0.492	1.137	1.709	0.183	0.586	1.219

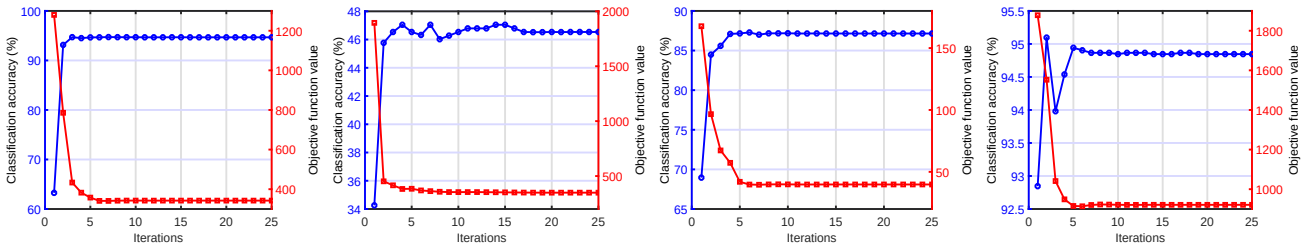


Fig. 11. Convergence curve and classification accuracy (%) versus iterations on (a) PIE, (b) LFW, (c) USPS and (d) COIL100 datasets.

those competing methods, LRRR, SLRR, DLSR, ReLSR, RLSL, RDR, RSLDA, ICSR, GLRR, PRDR and R-PRDR all learn a transform matrix and use nearest neighbor for classification. Thus, we compare the training time of these methods for learning the transform matrix. For a fair comparison, the dimension of learned transform matrix keeps the same for these algorithms (i.e., $d = c$). Since LRC and CRC do not learn a transform matrix, they are not included in comparisons. The experiments are repeated 20 times and the average training time of different methods are reported. Since the training time is relevant to the training data scale, we set different numbers of training samples for each dataset and compare the execution time of these methods. TABLE VIII lists the time complexity of different methods. TABLE IX exhibits the average training time (second) of different methods on six datasets under different training protocols.

We can see that ReLSR is the fastest algorithm in these methods which has linear computational complexity w.r.t training data size n and the computations in iterations are simple [22]. RLSL, a two-step transform based method, consumes the longest training time in all algorithms. The main reason is that RLSL needs to conduct SVD operation and solve a Sylvester equation problem in each iteration [44]. The $l_{2,1}$ norm and classwise constraints in SLRR, RDR, RSLDA, ICSR and GLRR make these algorithms inefficient as well. For LRRR, it can outperform PRDR on some datasets like COIL100 when the training size is large. It is because LRRR is equivalent to perform ridge regression in regularized LDA subspace and can be directly solved without iterations [43]. The training time costs of PRDR and R-PRDR are very close. In general, PRDR significantly outperforms most other regression methods on training time cost, which demonstrates the efficiency of proposed method. Although LRRR and ReLSR can obtain faster speed than PRDR, PRDR and R-PRDR achieve better classification performance than these methods.

E. Convergence Analysis

In Section III-C, we present a weak proof of convergence of the proposed optimization algorithm. Here we demonstrate its good convergence property by experimental examples. We show the convergence curves and classification accuracies of PRDR versus the number of iterations on PIE, LFW, USPS and COIL100 datasets in Fig. 11, where the red line is the convergence curve and the blue one is the classification accuracy curve. We can see that the objective function value fast decreases to a stable value, usually after 5 iterations. The classification accuracy goes up quickly in the first several iterations and changes only a litter bit after 10 iterations. The results in Fig. 11 prove that the proposed optimization algorithm is effective and efficient for solving PRDR model.

V. CONCLUSION

In this paper, we propose a novel pairwise relations oriented discriminative regression (PRDR) method with applications to image classification. In PRDR, the training data

is transformed into a latent space rather than label space to avoid the strict regression target problem. The pairwise label relations are exploited and preserved in latent space to guide the projection learning in a supervised manner. Besides, a graph based regularization term is introduced into PRDR to preserve the pairwise instance relations. The pairwise label relations, instance relations and projection learning are seamlessly integrated into a unified model. PRDR is proved to constrain the pairwise cosine distances between samples. By further enlarging the margins between true and false classes, PRDR is extended to a more discriminative version, i.e., R-PRDR. An iterative optimization algorithm is proposed to solve the proposed model. Extensive experiments on PIE, LFW, USPS, COIL100, Caltech101 and AwA datasets demonstrate that our proposed methods can achieve higher classification accuracy and lower training time costs than some state-of-the-art regression methods.

APPENDIX

Proof of Theorem 1. For optimization problem (14), its KKT conditions are as follows (note that the normalization constraint of $\mathbf{V}$ does not involve in the Lagrange multipliers, thus we do not proof the KKT condition for it):

$$\mathbf{V} = \mathbf{U}, \tag{26}$$

$$\partial\mathcal{J}/\partial\mathbf{W} = \mathbf{W}\mathbf{X}\mathbf{X}^T - \mathbf{V}\mathbf{X}^T + \lambda_1\mathbf{W} + \lambda_3\mathbf{W}\mathbf{P} = 0, \tag{27}$$

$$\partial\mathcal{J}/\partial\mathbf{V} = \mathbf{V} - \mathbf{W}\mathbf{X} + \lambda_2\mathbf{U}(\mathbf{U}^T\mathbf{V} - \mathbf{Y}^T\mathbf{Y}) + \mathbf{Z} = 0, \tag{28}$$

$$\partial\mathcal{J}/\partial\mathbf{U} = \lambda_2\mathbf{V}\mathbf{V}^T\mathbf{U} - \lambda_2\mathbf{Y}^T\mathbf{Y} - \mathbf{Z} = 0. \tag{29}$$

For Lagrange multiplier $\mathbf{Z}$, we update it by

$$\mathbf{Z}^+ \leftarrow \mathbf{Z} + \mu(\mathbf{V} - \mathbf{U}), \tag{30}$$

where $\mathbf{Z}^+$ is the next point of $\mathbf{Z}$ in the solution sequence $\{\theta^t\}_{t=1}^{\infty}$. If sequences of variables $\{\mathbf{Z}^j\}_{j=1}^{\infty}$ converge to a stationary point, i.e., $\mathbf{Z}^+ - \mathbf{Z} \to 0$, then $\mathbf{V} - \mathbf{U} \to 0$. Therefore, the first KKT condition (26) is obtained.

The second condition (27) obviously holds since the optimal $\mathbf{W}^+$ is derived from it.

For the third condition (28), we can obtain the following equation from the solution of $\mathbf{V}$:

$$\begin{aligned}\mathbf{V}^+ - \mathbf{V} = &\left((1+\mu)\mathbf{I} + \lambda_2\mathbf{U}\mathbf{U}^T\right)^{-1} \cdot \\ &(\mathbf{W}\mathbf{X} + \lambda_2\mathbf{U}\mathbf{Y}^T\mathbf{Y} + \mu\mathbf{U} - \mathbf{Z}) - \mathbf{V}.\end{aligned} \tag{31}$$

Then we can obtain

$$\begin{aligned}&\left((1+\mu)\mathbf{I} + \lambda_2\mathbf{U}\mathbf{U}^T\right)(\mathbf{V}^+ - \mathbf{V}) \\ =&\mathbf{W}\mathbf{X} + \lambda_2\mathbf{U}\mathbf{Y}^T\mathbf{Y} + \mu\mathbf{U} - \mathbf{Z} - (1+\mu)\mathbf{V} - \lambda_2\mathbf{U}\mathbf{U}^T\mathbf{V} \\ =&-\left(\mathbf{V} - \mathbf{W}\mathbf{X} + \lambda_2\mathbf{U}(\mathbf{U}^T\mathbf{V} - \mathbf{Y}^T\mathbf{Y}) + \mathbf{Z}\right) \\ &+ \mu(\mathbf{V} - \mathbf{U}).\end{aligned} \tag{32}$$

Based on the equation $\mathbf{V} - \mathbf{U} = 0$, it can be inferred that $\mathbf{V} - \mathbf{W}\mathbf{X} + \lambda_2\mathbf{U}(\mathbf{U}^T\mathbf{V} - \mathbf{Y}^T\mathbf{Y}) + \mathbf{Z} \to 0$, when $\mathbf{V}^+ - \mathbf{V} \to 0$. The third KKT condition is proved.

For the last condition (29), we can obtain that

$$\mathbf{U}^+ - \mathbf{U} = (\mu\mathbf{I} + \lambda_2\mathbf{V}\mathbf{V}^T)^{-1}(\lambda_2\mathbf{V}\mathbf{Y}^T\mathbf{Y} + \mu\mathbf{V} + \mathbf{Z}) - \mathbf{U}, \tag{33}$$

which can be equivalently rewritten as

$$\begin{aligned}&(\mu\mathbf{I} + \lambda_2\mathbf{V}\mathbf{V}^T)(\mathbf{U}^+ - \mathbf{U}) \\ &= \lambda_2\mathbf{V}\mathbf{Y}^T\mathbf{Y} + \mu\mathbf{V} + \mathbf{Z} - \mu\mathbf{U} - \lambda_2\mathbf{V}\mathbf{V}^T\mathbf{U} \\ &= -(\lambda_2\mathbf{V}\mathbf{V}^T\mathbf{U} - \lambda_2\mathbf{Y}^T\mathbf{Y} - \mathbf{Z}) + \mu(\mathbf{V} - \mathbf{U}).\end{aligned} \tag{34}$$

Similar to the procedure of proving the third condition, we can obtain that $\lambda_2\mathbf{V}\mathbf{V}^T\mathbf{U} - \lambda_2\mathbf{Y}^T\mathbf{Y} - \mathbf{Z} \to 0$, when $\mathbf{U}^+ - \mathbf{U} \to 0$. The fourth KKT condition is proved.

Since the solution sequence $\{\theta^t\}_{t=1}^{\infty}$ is assumed to be bounded and satisfy the condition of $\lim_{t\to\infty}(\theta^{t+1} - \theta^t) = 0$, every limit point of $\{\theta^t\}_{t=1}^{\infty}$ satisfies the above four KKT conditions. Thus, we complete the proof.

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] J. Lu, Y. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, 2013.

[2] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, 2018.

[3] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, "A two-phase test sample sparse representation method for use with face recognition," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 9, pp. 1255–1262, 2011.

[4] X. Fu, Z. Zha, F. Wu, X. Ding, and J. W. Paisley, "JPEG artifacts reduction via deep convolutional sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2501–2510.

[5] M. Yin, D. Zeng, J. Gao, Z. Wu, and S. Xie, "Robust multinomial logistic regression based on RPCA," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 6, pp. 1144–1154, 2018.

[6] J. Liu, P. C. Cosman, and B. D. Rao, "Robust linear regression via ℓ_0 regularization," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 698–713, 2018.

[7] X. Shu, J. Tang, G.-J. Qi, Z. Li, Y.-G. Jiang, and S. Yan, "Image classification with tailored fine-grained dictionaries," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 28, no. 2, pp. 454–467, 2018.

[8] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 281–288.

[9] X. Fang, N. Han, G. Zhou, S. Teng, Y. Xu, and S. Xie, "Dynamic double classifiers approximation for cross-domain recognition," *IEEE Trans. Cybern.*, 2020, doi:10.1109/TCYB.2020.3004398.

[10] K. Chen and W. Tao, "Learning linear regression via single-convolutional layer for visual object tracking," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 86–97, 2019.

[11] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, 2018.

[12] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, 2010.

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

[14] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2012, pp. 471–478.

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

[16] C. Zhang, H. Li, C. Chen, and X. Zhou, "Nonnegative representation based discriminant projection for face recognition," *Int. J. Mach. Learn. Cybern.*, 2020, doi:10.1007/s13042-020-01199-z.

[17] L. Wang, X. Zhang, and C. Pan, "MSDLSR: margin scalable discriminative least squares regression for multicategory classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2711–2717, 2016.

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

[19] X. Li, Y. Wang, Z. Zhang, R. Hong, Z. Li, and M. Wang, "RMoR-Aion: Robust multioutput regression by simultaneously alleviating input and output noises," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2020.2984635.

[20] Z. Zhang, W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, and S. Yan, "Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3798–3814, 2018.

[21] Z. Li, Z. Zhang, J. Qin, Z. Zhang, and L. Shao, "Discriminative fisher embedding dictionary learning algorithm for object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 786–800, 2020.

[22] X. Zhang, L. Wang, S. Xiang, and C. Liu, "Retargeted least squares regression algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2206–2213, 2015.

[23] X. Fang, Y. Xu, X. Li, Z. Lai, W. K. Wong, and B. Fang, "Regularized label relaxation linear regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1006–1018, 2018.

[24] Z. Zhang, L. Shao, Y. Xu, L. Liu, and J. Yang, "Marginal representation learning with graph structure self-adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4645–4659, 2018.

[25] N. Han, J. Wu, X. Fang, W. K. Wong, Y. Xu, J. Yang, and X. Li, "Double relaxed regression for image classification," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 30, no. 2, pp. 307–319, 2020.

[26] J. Wen, Z. Zhong, Z. Zhang, L. Fei, Z. Lai, and R. Chen, "Adaptive locality preserving regression," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 30, no. 1, pp. 75–88, 2020.

[27] M. Yang, C. Deng, and F. Nie, "Adaptive-weighting discriminative regression for multi-view classification," *Pattern Recognit.*, vol. 88, pp. 236–245, 2019.

[28] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, and G. Xie, "Discriminative elastic-net regularized linear regression," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1466–1481, 2017.

[29] N. Han, J. Wu, X. Fang, S. Xie, S. Zhan, K. Xie, and X. Li, "Latent elastic-net transfer learning," *IEEE Trans. Image Process.*, vol. 29, pp. 2820–2833, 2020.

[30] S. Zhan, J. Wu, N. Han, J. Wen, and X. Fang, "Group low-rank representation-based discriminant linear regression," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 30, no. 3, pp. 760–770, 2020.

[31] W. Yang, Y. Shi, Y. Gao, L. Wang, and M. Yang, "Incomplete-data oriented multiview dimension reduction via sparse low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6276–6291, 2018.

[32] J. Wen, Y. Xu, Z. Li, Z. Ma, and Y. Xu, "Inter-class sparsity based discriminative least square regression," *Neural Networks*, vol. 102, pp. 36–47, 2018.

[33] L. Wang and C. Pan, "Groupwise retargeted least-squares regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1352–1358, 2018.

[34] D. Zeng, M. Yin, S. Xie, and Z. Wu, "Robust regression with nonconvex schatten p-norm minimization," in *Int. Conf. Neural Inf. Process.*, vol. 11302, 2018, pp. 498–508.

[35] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2472–2484, 2018.

[36] X. Shi, Z. Guo, Z. Lai, Y. Yang, Z. Bao, and D. Zhang, "A framework of joint graph embedding and sparse regression for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1341–1355, 2015.

[37] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, "Local and global regressive mapping for manifold learning with out-of-sample extrapolation," in *Proc. AAAI Conf. Artif. Intell.*, 2010.

[38] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability." *Artif. Intell.*, vol. 273, pp. 37–55, 2019.

[39] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, 2016.

[40] H. Xue, S. Chen, and Q. Yang, "Discriminatively regularized least-squares classification," *Pattern Recognit.*, vol. 42, no. 1, pp. 93–104, 2009.

[41] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760–2771, 2015.

[42] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.

[43] X. Cai, C. H. Q. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1124–1132.

[44] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, "Robust latent subspace learning for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2502–2515, 2018.

[45] X. Zhen, M. Yu, F. Zheng, I. B. Nachum, M. Bhaduri, D. T. Laidley, and S. Li, "Multitarget sparse latent regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1575–1586, 2018.

[46] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 497–504, 2018.

[47] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2472–2484, 2018.

[48] Y. Zhang, W. Li, H. Li, R. Tao, and Q. Du, "Discriminative marginalized least-squares regression for hyperspectral image classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 5, pp. 3148–3161, 2020.

[49] H. Li, L. Zhang, B. Huang, and X. Zhou, "Cost-sensitive dual-bidirectional linear discriminant analysis," *Inf. Sci.*, vol. 510, pp. 283–303, 2020.

[50] N. Han, J. Wu, X. Fang, J. Wen, S. Zhan, S. Xie, and X. Li, "Transferable linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2020.2966746.

[51] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong, "Feature selection based on neighborhood discrimination index," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2986–2999, 2018.

[52] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of admm for multi-block convex minimization problems is not necessarily convergent," *Math. Program.*, vol. 155, no. 1-2, pp. 57–79, 2016.

[53] C. Chen, R. H. Chan, S. Ma, and J. Yang, "Inertial proximal ADMM for linearly constrained separable convex optimization," *SIAM J. Imaging Sci.*, vol. 8, no. 4, pp. 2239–2267, 2015.

[54] C. Chen, S. Ma, and J. Yang, "A general inertial proximal point algorithm for mixed variational inequality problem," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2120–2042, 2015.

[55] Z. Ren, D. Dong, H. Li, and C. Chen, "Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2216–2226, 2018.

[56] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

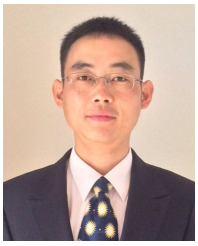
[57] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu, "Robust sparse linear discriminant analysis," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 29, no. 2, pp. 390–403, 2019.

[58] H. Li, L. Zhang, X. Zhou, and B. Huang, "Cost-sensitive sequential three-way decision modeling using a deep neural network," *Int. J. Approx. Reason.*, vol. 85, pp. 68–78, 2017.

[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[60] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

Chao Zhang (S'19) received the B.E. degree in automation from Nanjing University, Nanjing, China in 2018. He is currently pursuing the M.E. degree in the Department of Control and Systems Engineering, Nanjing University, Nanjing, China, and also working as a researcher at the Research Center for Novel Technology of Intelligent Equipment, Nanjing University, Nanjing, China. His current research interests include machine learning, pattern recognition, and computer vision.



Huaxiong Li (M'11) received the M.E. degree in control theory and control engineering from Southeast University, Nanjing, China, in 2006, and Ph.D. degree from Nanjing University, Nanjing, China, in 2009. He is currently an Associate Professor with the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. He was a visiting scholar at the Department of Computer Science, University of Regina, Canada, from 2007 to 2008, and a visiting scholar at the University of Hong Kong, Hong Kong, China, in 2010. His current research interests include machine learning, pattern recognition, and computer vision. He is a Committee Member of JiangSu association of Artificial Intelligence (JSAI) Pattern Recognition Committee, and a Committee Member of China Association of Artificial Intelligence (CAAI) Machine Learning Committee.

Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multi-granulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has authored over 80 articles on these topics in international journals. He served on the Editorial Board of the International Journal of Knowledge-Based Organizations and Artificial Intelligence Research. He has served as the Program Chair or Special Issue Chair of the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and the Conference on Industrial Instrumentation and Control, and a PC Member of many machine learning, data mining, and granular computing conferences.

Chunlin Chen (S'05-M'06) received the B.E. degree in automatic control and Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively. He is currently a Professor and the Chair of the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. He was a visiting scholar at Princeton University, Princeton, USA, from 2012 to 2013. He had visiting positions at the University of New South Wales Canberra at ADFA, Australia, and the City University of Hong Kong, Hong Kong, China.

His recent research interests include machine learning, pattern recognition, intelligent information processing, and quantum control. He is a Co-Chair of Technical Committee on Quantum Cybernetics, IEEE Systems, Man and Cybernetics Society. He is a Committee Member of JiangSu association of Artificial Intelligence (JSAI) Pattern Recognition Committee, and a Committee Member of China Association of Artificial Intelligence (CAAI) Machine Learning Committee.

Yang Gao received the B.S. degree in mechanical engineering from the Dalian University of Technology, Dalian, China, in 1993, the M.S. degree in computer aided design from the Nanjing University of Science and Technology, Nanjing, China in 1996, and the Ph.D. degree in computer science from Nanjing University, Nanjing, in 2000, where he is currently a Professor and the Deputy Director with the Department of Computer Science and Technology.

He is directing the Reasoning and Learning Research Group, Nanjing University. He has published more than 100 articles in top-tired conferences and journals. His research interests include artificial intelligence and machine learning. He serves as a Program Chair and an Area Chair for many international conferences.