Robust and Sparse Linear Discriminant Analysis via an Alternating Direction Method of Multipliers

Chun-Na Li, Yuan-Hai Shao, Wotao Yin, and Ming-Zeng Liu

***Abstract*— In this paper, we propose a robust linear discriminant analysis (RLDA) through Bhattacharyya error bound optimization. RLDA considers a nonconvex problem with the L_1-norm operation that makes it less sensitive to outliers and noise than the L_2-norm linear discriminant analysis (LDA). In addition, we extend our RLDA to a sparse model (RSLDA). Both RLDA and RSLDA can extract unbounded numbers of features and avoid the small sample size (SSS) problem, and an alternating direction method of multipliers (ADMM) is used to cope with the nonconvexity in the proposed formulations. Compared with the traditional LDA, our RLDA and RSLDA are more robust to outliers and noise, and RSLDA can obtain sparse discriminant directions. These findings are supported by experiments on artificial data sets as well as human face databases.**

***Index Terms*— Alternating direction method of multipliers (ADMMs), dimensionality reduction (DR), linear discriminant analysis (LDA), robust linear discriminant analysis (RLDA), sparse feature extraction.**

I. Introduction

DIMENSIONALITY reduction (DR) plays an important role in pattern recognition and has been studied extensively. Several DR methods are widely used, such as principal component analysis (PCA) [1], canonical correlation analysis (CCA) [2], [3], locality preserving projections (LPPs) [4], neighborhood preserving embedding (NPE) [5], sparsity preserving projections (SPP) [6], [7], and linear discriminant analysis (LDA) [8], [9]. Among them, LDA is a powerful tool for feature extraction and has been extensively studied, including multimodal DR [10], audiovisual speech recognition [11], and tensor extension on image representation [12]–[14]. LDA tries to find the optimal projection direction by maximizing the between-class variance while simultaneously minimizing the within-class variance in the projected space.

However, when dealing with polluted data, minimizing the sum of squared errors makes LDA easily affected by outliers and noise. This may lead the projection vectors to drift from the desired directions. Moreover, the classical L_2-norm-based LDA may suffer from the small sample size (SSS) problem since it is difficult to evaluate the scatter matrix accurately in this situation. To address these problems, various approaches are investigated. Replacing the sample means and covariance matrices with their robust counterparts was considered in [15]–[17]. Lanckrie *et al.* [18] reported a robust performance of LDA with two classes by considering a probability-based minimax optimization technique. Kim *et al.* [19] achieved robustness by considering a convex uncertainty LDA model. Local Fisher discriminant analysis (LFDA) [20], [21] that incorporates the spirit of LPP [4] into the classical LDA was studied to alleviate the effect of outliers by considering neighborhood information. As an improvement of LFDA, Zhang and Chow [22] not only considered the neighborhood information but also utilized the density region computed from each class. Okwonu and Othman [23] further presented a robust LDA by robustifying the classical sample means and covariance.

Recently, the implementation of the L_1-norm-related techniques has been regarded as an effective way to reduce the impact of outliers, which has been proven to be more robust than the L_2-norm-related techniques [24]–[26]. In fact, compared to the L_2-norm, the L_1-norm ensures that samples with large errors do not dominate. Li *et al.* [27] put forward a rotationally invariant L_1-norm (R_1-norm)-based LDA (DCL_1), which was the robust version of the weighted maximum margin criterion (WMMC) [28]. However, it was found that the L_1-norm is more efficient than the R_1-norm in suppressing outliers [29]. Therefore, an L_1-norm-based LDA (LDA-L1) was proposed in [30]–[32], and an iterative algorithm was presented at the same time. It was reported in [30]–[32] that the proposed LDA-L1 was more robust than the classical LDA and DCL_1. Inspired by the idea of LDA-L1, Li *et al.* [33] put forward a matrix version of LDA-L1. Recently, Zheng *et al.* [34] devised a different form of the L_1-norm discriminant analysis (L1-LDA) criterion via Bayes error bound optimization.

Manuscript received October 19, 2017; revised March 28, 2018, September 11, 2018 and January 6, 2019; accepted April 5, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61703370, Grant 61866010, Grant 11871183, and Grant 61603338, in part by the Natural Science Foundation of Zhejiang Province under Grant LQ17F030003 and Grant LY18G010018, and in part by the Natural Science Foundation of Hainan Province under Grant 118QN181. *(Corresponding author: Yuan-Hai Shao.)*

C.-N. Li is with the School of Management, Hainan University, Haikou 570228, China, and also with the Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China (e-mail: na1013na@163.com).

Y.-H. Shao is with the School of Management, Hainan University, Haikou 570228, China (e-mail: shaoyuanhai21@163.com).

W. Yin is with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095-1555 USA (e-mail: wotaoyin@math.ucla.edu).

M.-Z. Liu is with the School of Mathematics and Physics Science, Dalian University of Technology at Panjin, Dalian 124221, China (e-mail: mzliu@dlut.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2910991

Nevertheless, since LDA-L1 and L1-LDA are ratio formulations, they are solved by introducing different surrogate functions in [30]–[32] and [34], which may affect the optimality of the solutions. Moreover, for L1-LDA in [34], its Bayes error bound theoretical derivation was established under the rigorous condition that each class has an equal number of samples [34], which does not always hold in the real world.

In this paper, we study robust DR and propose a robust LDA criterion (RLDA). RLDA incorporates the same idea as LDA such that the projected center of each class is far from the projected center of the whole set of training data and simultaneously close to the projected data of its own class but takes the L_1-norm between-class scatter. Different from L1-LDA proposed in [34], RLDA can be realized through the Bhattacharyya error bound optimization without the requirement that each class has an equal number of samples. Since the proposed RLDA involves the L_1-norm operation, we here employ a simple but efficient alternating direction method of multipliers (ADMM) [35], [36]. ADMM works well for convex problems, and its convergence was established in [37]–[39]. For nonconvex problems, its convergence can only be ensured under some restrictive conditions [40], [41]. However, ADMM still has good performance on many nonconvex problems, for example, matrix completion [42], tensor factorization [43], and compressive sensing [44]. Because the optimization problem in our model is nonconvex, we actually cannot strictly prove the convergence of the algorithm. We present the optimality conditions and hence the reasonability of the stopping criterion of the proposed algorithm in Section II of the Supplementary Material. We also note that both LDA-L1 and L1-LDA lack sparseness, whereas sparseness has become increasingly more important for LDA [45]–[49]. Therefore, we also extend our RLDA to its sparse version and call it RSLDA.

Specifically, the proposed RLDA and RSLDA have the following several characteristics.

1) Due to the utilization of the L_1-norm, RLDA and RSLDA are more robust to outliers and noise than the classical LDA and can avoid the SSS problem.
2) The number of extracted features obtained by RLDA and RSLDA is not constrained to $c-1$, while LDA can extract at most $c-1$ features, where c is the number of classes.
3) Two simple but efficient ADMM algorithms are constructed for the proposed nonconvex RLDA and RSLDA.
4) The experimental results from two artificial data sets and three human face databases demonstrate the robustness of RLDA and RSLDA, while also showing the sparseness of RSLDA.

We now define the notation used in this paper. All vectors are column vectors. Given a training data set $T = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_m, l_m)\}$, where $\mathbf{x}_t \in \mathbb{R}^n$ is the input and $l_t \in \{1, \ldots, c\}$ is the corresponding label, $t = 1, \ldots, m$, we organize the m inputs by a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{n\times m}$. Assume that the ith class contains m_i samples. Then, $\sum_{i=1}^{c} m_i = m$. Denote $\overline{\mathbf{x}}_i$ as the mean of samples in the ith class and $\overline{\mathbf{x}}$ as the center of the whole set of samples, that is, $\overline{\mathbf{x}}_i = (1/m_i)\sum_{j=1}^{m_i} \mathbf{x}_{ij}$ and $\overline{\mathbf{x}} = (1/m)\sum_{l=1}^{m} \mathbf{x}_l$, where $\mathbf{x}_{ij}$ is the jth element in the ith class. For $\mathbf{w} \in \mathbb{R}^n$ and a sample $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{w}^T\mathbf{x}$ maps each $\mathbf{x}$ into a 1-D vector. Generally, if $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_d) \in \mathbb{R}^{n\times d}$ with $d \le n$, then $\mathbf{W}^T\mathbf{x}$ maps each $\mathbf{x} \in \mathbb{R}^n$ into a d-dimensional space.

This paper is organized as follows. Section II gives a brief summary of LDA. Section III proposes our RLDA in the primal space. Section IV extends RLDA to its sparse version RSLDA. Section V discusses the relationship between the proposed methods and their related ones, as well as an analysis of computational cost. Section VI experimentally compares RLDA and RSLDA to their related methods. Concluding remarks are given in Section VII.

II. Linear Discriminant Analysis

LDA is a well-known method for DR and classification. LDA provides low-dimensional projections of the data onto the most discriminative directions, which is useful for data interpretation. LDA is also a favored tool for supervised classification in many applications due to its simplicity and predictive accuracy. LDA can be seen as arising from Fisher's discriminant problem. Define

$$\mathbf{S}_b = \frac{1}{m}\sum_{i=1}^{c} m_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T \tag{1}$$

$$\mathbf{S}_w = \frac{1}{m}\sum_{i=1}^{c}\sum_{j=1}^{m_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T. \tag{2}$$

Fisher's discriminant problem involves seeking discriminant vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$, $d \le c-1$ by successively solving the problem

$$\max_{\mathbf{w}_h} \frac{{\mathbf{w}_h}^T\mathbf{S}_b\mathbf{w}_h}{\mathbf{w}_h^T\mathbf{S}_w\mathbf{w}_h} \tag{3}$$

such that $\mathbf{w}_h^T\mathbf{S}_w\mathbf{w}_l = 0$, $1 \le l < h \le d$. In addition to the above $\mathbf{S}_w$-orthogonal constraints, the constraint $\mathbf{w}_h^T\mathbf{w}_l = 0$ is also used as a powerful replacement [50]. The solution to the optimization problem (3) is given by the eigenvectors corresponding to the first largest d nonzero eigenvalues of the generalized problem $\mathbf{S}_b\mathbf{w}_h = \eta\mathbf{S}_w\mathbf{w}_h$, where $\mathbf{S}_w$ is nonsingular, and $\eta \neq 0$. Since the rank of $\mathbf{S}_b$ is at most $c-1$, the number of extracted features is less than or equal to $c-1$. Note that the definition of $\mathbf{S}_b$ is based on the L_2-norm, and therefore, outliers will affect the between-class distance; hence, it is sensitive to outliers. Furthermore, since LDA solves the generalized eigenvalue problem, it may encounter the SSS problem.

III. Robust Linear Discriminant Analysis

A. Problem Formulation

As mentioned before, the classical LDA obtains discriminant vectors by eigendecomposing. However, this requires $\mathbf{S}_w$ to be nonsingular. In addition, the number of reduced dimensions will be restricted because the rank of $\mathbf{S}_b$ is at most $c-1$, and the measure of $\mathbf{S}_b$ by using the L_2-norm distance makes the classical LDA sensitive to outliers. These considerations motivate us to construct a new form of LDA

with a minus operator and to consider the L_1-norm between-class covariance. Therefore, when $d = 1$, our formulation is given as follows:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; -\sum_{i=1}^{c} m_i |\mathbf{w}^T \overline{\mathbf{x}}_i - \mathbf{w}^T \overline{\mathbf{x}}| + \lambda \mathbf{w}^T \mathbf{S}_w \mathbf{w} \tag{4}$$

where λ is a positive tuning parameter.

The geometric interpretation of problem (4) is clear. Optimizing the first term of (4) maximizes the L_1-norm between class scatter, which forces the data points from different classes to be as far as possible from each other while minimizing the second term of (4) makes the within-class scatter as small as possible. In fact, there are already some minus formulations of LDA, such as least-squares LDA [51]. However, eigen-decomposing is still necessary, and the singularity problem remains. It should also be noted that in the case of SSS problems, the rank of $\mathbf{S}_w$ is at most $m \ll n$. Note that the dimension of the space spanned by m points is at most $m-1$.

Remark: We can show that under some assumptions, problem (4) can be derived under the optimality of error bound. Let $\mathbf{x} \in \mathbb{R}^n$ be a sample vector, and P_i and $p_i(\mathbf{x})$ be the prior probability and the probability density function (PDF) of the ith class, respectively, where $i = 1, 2, \ldots, c$. Assume the data samples in each class are independent identically normally distributed and that the PDF of each class is the Gaussian function $p_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Theta})$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Theta}$ are the class mean and covariance matrix, respectively. We further suppose $\boldsymbol{\mu}_i$ and $\boldsymbol{\Theta}$ can be estimated accurately from the sample mean $\overline{\mathbf{x}}_i$ and sample covariance $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, c$. Let the 1-D projected samples be $\{\widetilde{\mathrm{x}}_t = \mathbf{w}^T \mathbf{x}_t\}$, $t = 1, 2, \ldots, m$. Then, $p_i(\widetilde{\mathrm{x}}) = \mathcal{N}(\widetilde{\mathrm{x}}|\widetilde{\overline{\mathrm{x}}}_i, \sigma)$, where σ is the projected standard deviation defined by $\sigma = ((1/m)\sum_{t=1}^{m}(\mathbf{w}^T\mathbf{x}_t - \mathbf{w}^T\overline{\mathbf{x}}_{l_t})^2)^{1/2} = (1/m)^{1/2}\|\mathbf{w}^T\mathbf{X} - \mathbf{w}^T\overline{\mathbf{x}}_{\mathbf{I}}\|_2$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in R^{n\times m}$, $\overline{\mathbf{x}}_{\mathbf{I}} = (\overline{\mathbf{x}}_{l_1}, \ldots, \overline{\mathbf{x}}_{l_m}) \in \mathbb{R}^{n\times m}$, and $\overline{\mathbf{x}}_{l_t}$ is the center of the class that $\mathbf{x}_t$ belongs to.

The error probability minimization is a natural way to obtain DR for classification [9], [52]. Since the Bayes classifier is the best classifier that minimizes the probability of classification error [9], minimizing the Bayes error is always expected to lead to good classification models. In the context of DR, the Bayes error is defined by

$$\epsilon = 1 - \int \max_{i\in\{1,2,\ldots,c\}} \{P_i\, p_i(\widetilde{\mathrm{x}})\} d\widetilde{\mathrm{x}}. \tag{5}$$

Compared to the Bayes error, the Bhattacharyya error [9], [53] that involves the maximization of probabilistic distance measures and probabilistic dependence measures between different classes [52] is a widely used upper bound that provides a close bound to the Bayes error. For multiclass classification, the union Bhattacharyya error in the context of DR is expressed as follows:

$$\epsilon_B = \sum_{1\le i<j\le c} \sqrt{P_i P_j} \int \sqrt{p_i(\widetilde{\mathrm{x}})\, p_j(\widetilde{\mathrm{x}})}\, d\widetilde{\mathrm{x}}. \tag{6}$$

It was shown that the Bhattacharyya error upper bounded the Bayes error by the inequality $\min\{a, b\} \le \sqrt{ab}$ [9]. We can show that our RSLDA error estimation is a tight linear upper bound of the Bhattacharyya bound estimation. The detailed derivation is in Section I of the Supplementary Material.

B. ADMM Algorithm of RLDA for One Discriminant Direction

In this section, we solve problem (4) by using the ADMM technique. To implement the ADMM algorithm, we first transfer problem (4) to its ADMM form as

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{y},\mathbf{z}} \; -\|\mathbf{y}\|_1 + \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} \\ &\text{s.t. } \mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y} = 0 \\ &\qquad \mathbf{w} - \mathbf{z} = 0 \end{aligned} \tag{7}$$

where $\mathbf{X_0} = (m_1(\overline{\mathbf{x}_1} - \overline{\mathbf{x}})), \ldots, m_c(\overline{\mathbf{x}_c} - \overline{\mathbf{x}})) \in \mathbb{R}^{n\times c}$.

Note that the first item of problem (7) is nonconvex and nonsmooth, which will make it hard to solve. Therefore, we here give a smooth approximation of $-\|\mathbf{y}\|_1$ by $-\sum_{i=1}^{c}(y_i^2+\varepsilon)^{(1/2)}$ [54], [55], where $\varepsilon > 0$ is a small enough real number and y_i is the ith element of $\mathbf{y}$. Then, (7) becomes

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{y},\mathbf{z}} \; -\sum_{i=1}^{c} \left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} \\ &\text{s.t. } \mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y} = 0 \\ &\qquad \mathbf{w} - \mathbf{z} = 0. \end{aligned} \tag{8}$$

In the following, we solve problem (8).

The Lagrangian of problem (8) is written as

$$\begin{aligned} &L_{0R}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{v}_1, \mathbf{v}_2) \\ &\quad = \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} - \sum_{i=1}^{c} \left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \mathbf{v}_1^T \left(\mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y}\right) \\ &\qquad + \mathbf{v}_2^T (\mathbf{w} - \mathbf{z}) \end{aligned} \tag{9}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the dual variables.

However, the augmented Lagrangian is usually used to introduce robustness, which can be given by

$$\begin{aligned} &L_{\rho R}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{v}_1, \mathbf{v}_2) \\ &\quad = \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} - \sum_{i=1}^{c} \left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \mathbf{v}_1^T \left(\mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y}\right) \\ &\qquad + \mathbf{v}_2^T (\mathbf{w} - \mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y}\right\|_2^2 + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 \end{aligned} \tag{10}$$

where $\rho > 0$ is the penalty parameter. In real computations, we usually use a scaled version of the above augmented Lagrangian for succinctness, which is formed as

$$\begin{aligned} &L_{\rho R}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{u}_1, \mathbf{u}_2) \\ &\quad = \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} - \sum_{i=1}^{c} \left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \frac{\rho}{2}\left\|\mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y} + \mathbf{u}_1\right\|_2^2 \\ &\qquad + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z} + \mathbf{u}_2\|_2^2 \end{aligned} \tag{11}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the scaled dual variables satisfying $\mathbf{u}_1 = (1/\rho)\mathbf{v}_1$ and $\mathbf{u}_2 = (1/\rho)\mathbf{v}_2$.

Clearly, (10) and (11) are equivalent up to some constant. After the Lagrangian is constructed, the ADMM algorithm of (4) can be given as in Algorithm 1. For the iteration

Algorithm 1 ADMM Algorithm for RLDA

Input: The training input data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, parameters $\lambda, \rho > 0$; small positive tolerances ϵ^{pri1}, ϵ^{pri2}, $\epsilon^{\text{dual}} > 0$; small parameter $\varepsilon > 0$.

Process:

I. Initialization. Set the iteration number $k = 0$ and randomly initialize $\mathbf{w}^0$, $\mathbf{u}_2^0 \in \mathbb{R}^n$, $\mathbf{y}^0$, $\mathbf{u}_1^0 \in \mathbb{R}^c$.

II. **Repeat**

(a) $\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \ \lambda \mathbf{z}^T \mathbf{S}_w \mathbf{z} + \frac{\rho}{2}\|\mathbf{X}_\mathbf{0}^T \mathbf{z} - \mathbf{y}^k + \mathbf{u}_1^k\|_2^2 + \frac{\rho}{2}\|\mathbf{w}^k - \mathbf{z} + \mathbf{u}_2^k\|_2^2$;

(b) $\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} -\sum_{i=1}^{c}(y_i^2+\varepsilon)^{\frac{1}{2}} + \frac{\rho}{2}\|\mathbf{X}_\mathbf{0}^T \mathbf{z}^{k+1} - \mathbf{y} + \mathbf{u}_1^k\|_2^2$;

(c) $\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} \ \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}^{k+1} + \mathbf{u}_2^k\|_2^2$;

(d) $\mathbf{u}_1^{k+1} = \mathbf{u}_1^k + (\mathbf{X}_\mathbf{0}^T \mathbf{z}^{k+1} - \mathbf{y}^{k+1})$;

(e) $\mathbf{u}_2^{k+1} = \mathbf{u}_2^k + (\mathbf{w}^{k+1} - \mathbf{z}^{k+1})$.

Until the stopping criterion is satisfied. Specifically, the iteration stops if the following are satisfied:

$\|\mathbf{X}_\mathbf{0}^T \mathbf{z}^{k+1} - \mathbf{y}^{k+1}\|_2 \le \epsilon^{\text{pri1}}$;

$\|\mathbf{w}^{k+1} - \mathbf{z}^{k+1}\|_2 \le \epsilon^{\text{pri2}}$;

$\|\mathbf{X}_\mathbf{0}(\mathbf{y}^{k+1} - \mathbf{y}^k) + (\mathbf{w}^{k+1} - \mathbf{w}^k)\| \le \epsilon^{\text{dual}}$.

Output: $\mathbf{w}^* = \mathbf{w}^{k+1}$.

termination criterion of Algorithm 1, the primal and dual residuals are required to be as small as possible. The rationality of the stopping criteria is discussed in Section II of the Supplementary Material. We now give the solutions of iterative steps (a)–(c) in Algorithm 1.

For (a) in Algorithm 1, since it is an unconstrained quadratic problem, we can solve it by taking its gradient. Denote $\mathbf{G} = \mathbf{X}_\mathbf{0}\mathbf{X}_\mathbf{0}^T + I + (2\lambda/\rho)\mathbf{S}_w$ and $\mathbf{g}^k = \mathbf{X}_\mathbf{0}(\mathbf{y}^k - \mathbf{u}_1^k) + (\mathbf{w}^k + \mathbf{u}_2^k)$. Then, (a) in Algorithm 1 is equivalent to the following minimization problem:

$$\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \ \frac{\rho}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} - \rho(\mathbf{g}^k)^T\mathbf{z}. \tag{12}$$

It is obvious that $\mathbf{G}$ is positive definite. Therefore, we can give the solution of (12) by $\mathbf{z}^{k+1} = \mathbf{G}^{-1}\mathbf{g}^k$. Note that during the whole iteration, we need only to compute $\mathbf{G}^{-1}$ once because there is no update of $\mathbf{G}$.

For solving (b) in Algorithm 1, it can be seen that the involved problem is separable with respect to its own variable components, and hence, its solution can be obtained elementwisely by direct computation. However, though the involved function is smooth, it is not easy to obtain an analytic solution. Therefore, we here use an iterative form. By setting the gradient of $-\sum_{i=1}^{c}(y_i^2+\varepsilon)^{(1/2)} + (\rho/2)\|\mathbf{X}_\mathbf{0}^T\mathbf{z}^{k+1} - \mathbf{y} + \mathbf{u}_1^k\|_2^2$ to 0, we obtain the iterative form as

$$y_i^{k+1} = \frac{\rho\left(\mathbf{X}_\mathbf{0}^T\mathbf{z}^{k+1} + \mathbf{u}_1^k\right)_i \sqrt{\left(y_i^k\right)^2 + \varepsilon}}{\rho\sqrt{\left(y_i^k\right)^2 + \varepsilon} - 1} \tag{13}$$

where $i = 1, 2, \ldots, c$. It can be computed directly that when $\rho > (1/\varepsilon)(\sqrt{\varepsilon} + (|a|/2)) + ((|a|/\varepsilon\sqrt{\varepsilon}) + (|a|^2/4\varepsilon^2))^{1/2}$, where $a = (\mathbf{X}_\mathbf{0}^T\mathbf{z}^{k+1} + \mathbf{u}_1^k)_i$, the above-mentioned iteration satisfies the Lipschitz condition with a Lipschitz constant of less than 1, and therefore, the iterative form (13) converges to its limit point.

For solving (c) in Algorithm 1, we can easily obtain

$$\mathbf{w}^{k+1} = \mathbf{z}^{k+1} - \mathbf{u}_2^k.$$

Algorithm 2 Recursive RLDA for Multiple Discriminant Directions

Input: The training input data $T = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_m, l_m)\}$, parameters λ, $\rho > 0$; small positive tolerances ϵ^{pri1}, ϵ^{pri2}, ϵ^{dual}; small parameter $\varepsilon > 0$, and desired number $d > 1$ of discriminant vectors.

Process:

I. Initialization. Set the iteration number $k = 0$ and the training set $T_0 = \{\mathbf{x}_{i0} | \mathbf{x}_{i0} = \mathbf{x}_i,\ i = 1, \ldots, m\}$. Let $\mathbf{w}_0 = \mathbf{0} \in \mathbb{R}^n$, where $\mathbf{0}$ is the vector of zeros.

II. For $l = 1, \ldots, d$, do the following iteration:

(a) Compute $T_l = \{\mathbf{x}_{il} | \mathbf{x}_{il} = \mathbf{x}_{i,l-1} - \mathbf{w}_{k-1}\mathbf{w}_{l-1}^T\mathbf{x}_{i,l-1}\ \ i = 1, \ldots, m\}$.

(b) Apply Algorithm 1 to data set X_l to get $\mathbf{w}_l$.

End for

Output: Discriminant vectors $\{\mathbf{w}_l^*\}_{l=1}^d$ for a series of (16).

C. Recursive RLDA for Multiple Discriminant Directions

By implementing Algorithm 1, we can obtain a single discriminant direction to enable the projected samples to be well separated as much as possible. However, for a real data set, one discriminant direction is usually not sufficient, and it is necessary to project the data into a higher-dimensional space rather than a 1-D space. Therefore, to obtain more than one direction, we here employ a greedy search strategy by applying recursive procedure [56]–[58]. Specifically, if we denote the first discriminant direction obtained by Algorithm 1 as $\mathbf{w}_1$, then we need to compute subsequent discriminant directions $\mathbf{w}_r$ for $r > 1$. We adopt the deflation technique: the rth $(1 < r \le n)$ discriminant direction $\mathbf{w}_r$ is computed by using the deflated samples

$$\mathbf{x}_i^{\text{new}} = \mathbf{x}_i - \sum_{l=1}^{r-1}\mathbf{w}_l\left(\mathbf{w}_l^T\mathbf{x}_i\right) \tag{14}$$

where $\mathbf{w}_l$ is the lth discriminant direction and $\mathbf{x}_i$ is the ith data sample. Thus, (14) ensures that the new data samples are computed such that the information contained in the previously obtained discriminant directions is deducted. This gives the recursive RLDA Algorithm 2.

IV. Robust and Sparse Linear Discriminant Analysis

A. Problem Formulation

Although problem (4) is an upper bound of the Bhattacharyya error and can also achieve robustness, it lacks sparseness. A sparse discriminant direction ensures easier model interpretation and may reduce overfitting of the training data [59], and this can be ensured by adding an L_1-norm constraint on the discriminant direction. Therefore, in order

to obtain a sparse robust discriminant direction, we consider the following problem:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} -\sum_{i=1}^{c} m_i|\mathbf{w}^T\overline{\mathbf{x}_i} - \mathbf{w}^T\overline{\mathbf{x}}| + \lambda\mathbf{w}^T\mathbf{S}_w\mathbf{w} \quad \text{s.t. } \|\mathbf{w}\|_1 < s_0 \tag{15}$$

where $\lambda > 0$ is a tuning parameter, the L_1-norm inequality constraint controls the sparsity, and s_0 is the sparsity upper bound. We describe the above-mentioned problem (15) as robust and sparse LDA or RSLDA for brevity.

Since problem (15) contains the L_1-norm term in both its objective and constraint, we consider the following minimization form:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} -\sum_{i=1}^{c} m_i|\mathbf{w}^T\overline{\mathbf{x}_i} - \mathbf{w}^T\overline{\mathbf{x}}| + \delta\|\mathbf{w}\|_1 + \lambda\mathbf{w}^T\mathbf{S}_w\mathbf{w} \tag{16}$$

where $\delta, \lambda > 0$ are the tuning parameters. The geometric meaning of problem (16) is clear. Compared to RLDA, (16) has an extra L_1-norm term $\delta\|\mathbf{w}\|_1$, which controls the sparsity while preventing the model from overfitting at the same time. Therefore, compared to RLDA, the formulation of (16) will provide a sparse linear discriminant direction.

B. ADMM Algorithm of RSLDA for One Discriminant Direction

First, we reformulate problem (16) as its ADMM form

$$\min_{\mathbf{w},\mathbf{y},\mathbf{z}} -\|\mathbf{y}\|_1 + \delta\|\mathbf{w}\|_1 + \lambda\mathbf{z}^T\mathbf{S}_w\mathbf{z} \quad \text{s.t. } \mathbf{X_0}^T\mathbf{z} - \mathbf{y} = 0,\ \mathbf{w} - \mathbf{z} = 0 \tag{17}$$

where $\mathbf{X_0} = (m_1(\overline{\mathbf{x}_1} - \overline{\mathbf{x}}), \ldots, m_c(\overline{\mathbf{x}_c} - \overline{\mathbf{x}})) \in \mathbb{R}^{n\times c}$. Similar to RLDA, we consider solving the following smoothing problem:

$$\min_{\mathbf{w},\mathbf{y},\mathbf{z}} -\sum_{i=1}^{c}\left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \delta\|\mathbf{w}\|_1 + \lambda\mathbf{z}^T\mathbf{S}_w\mathbf{z} \quad \text{s.t. } \mathbf{X_0}^T\mathbf{z} - \mathbf{y} = 0,\ \mathbf{w} - \mathbf{z} = 0 \tag{18}$$

where $\varepsilon > 0$ is a small enough real number.

The Lagrangian of problem (18) is then written as

$$L_{0RS}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{v}_1, \mathbf{v}_2) = \lambda\mathbf{z}^T\mathbf{S}_w\mathbf{z} - \sum_{i=1}^{c}\left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \delta\|\mathbf{w}\|_1 + \mathbf{v}_1^T\left(\mathbf{X_0}^T\mathbf{z} - \mathbf{y}\right) + \mathbf{v}_2^T(\mathbf{w} - \mathbf{z}) \tag{19}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the dual variables.

As in the case of RLDA, the augmented Lagrangian is given by

$$L_{\rho RS}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{v}_1, \mathbf{v}_2) = \lambda\mathbf{z}^T\mathbf{S}_w\mathbf{z} - \sum_{i=1}^{c}\left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \delta\|\mathbf{w}\|_1 + \mathbf{v}_1^T\left(\mathbf{X_0}^T\mathbf{z} - \mathbf{y}\right) + \mathbf{v}_2^T(\mathbf{w} - \mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{X_0}^T\mathbf{z} - \mathbf{y}\right\|_2^2 + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 \tag{20}$$

Algorithm 3 ADMM Algorithm for RSLDA

Input: The training input data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, parameters λ, $\rho > 0$; small positive tolerances ϵ^{pri1}, ϵ^{pri2}, $\epsilon^{\text{dual}} > 0$; small parameter $\varepsilon > 0$.

Process:

I. Initialization. Set the iteration number $k = 0$ and randomly initialize $\mathbf{w}^0$, $\mathbf{u}_2^0 \in \mathbb{R}^n$, $\mathbf{y}^0$, $\mathbf{u}_1^0 \in \mathbb{R}^c$.

II. **Repeat**

(a) $\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \lambda(\mathbf{z}^T\mathbf{S}_w\mathbf{z} - 1) + \frac{\rho}{2}\|\mathbf{X_0}^T\mathbf{z} - \mathbf{y}^k + \mathbf{u}_1^k\|_2^2 + \frac{\rho}{2}\|\mathbf{w}^k - \mathbf{z} + \mathbf{u}_2^k\|_2^2$;

(b) $\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} -\sum_{i=1}^{c}(y_i^2 + \varepsilon)^{\frac{1}{2}} + \frac{\rho}{2}\|\mathbf{X_0}^T\mathbf{z}^{k+1} - \mathbf{y} + \mathbf{u}_1^k\|_2^2$;

(c) $\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} \delta\|\mathbf{w}\|_1 + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}^{k+1} + \mathbf{u}_2^k\|_2^2$;

(d) $\mathbf{u}_1^{k+1} = \mathbf{u}_1^k + (\mathbf{X_0}^T\mathbf{z}^{k+1} - \mathbf{y}^{k+1})$;

(e) $\mathbf{u}_2^{k+1} = \mathbf{u}_2^k + (\mathbf{w}^{k+1} - \mathbf{z}^{k+1})$.

Until the stopping criterion is satisfied. Specifically, the iteration stops if the following is satisfied:

$\|\mathbf{X_0}^T\mathbf{z}^{k+1} - \mathbf{y}^{k+1}\|_2 \le \epsilon^{\text{pri1}}$;

$\|\mathbf{w}^{k+1} - \mathbf{z}^{k+1}\|_2 \le \epsilon^{\text{pri2}}$;

$\|\mathbf{X_0}(\mathbf{y}^{k+1} - \mathbf{y}^k) + (\mathbf{w}^{k+1} - \mathbf{w}^k)\| \le \epsilon^{\text{dual}}$.

Output: $\mathbf{w}^* = \mathbf{w}^{k+1}$.

or

$$L_{\rho RS}(\mathbf{z}, \mathbf{y}, \mathbf{w}; \mathbf{u}_1, \mathbf{u}_2) = \lambda\mathbf{z}^T\mathbf{S}_w\mathbf{z} - \sum_{i=1}^{c}\left(y_i^2 + \varepsilon\right)^{\frac{1}{2}} + \delta\|\mathbf{w}\|_1 + \frac{\rho}{2}\left\|\mathbf{X_0}^T\mathbf{z} - \mathbf{y} + \mathbf{u}_1\right\|_2^2 + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z} + \mathbf{u}_2\|_2^2 \tag{21}$$

where ρ is the proper positive penalty parameter and $\mathbf{u}_1$ and $\mathbf{u}_2$ are the scaled dual variables satisfying $\mathbf{u}_1 = (1/\rho)\mathbf{v}_1$ and $\mathbf{u}_2 = (1/\rho)\mathbf{v}_2$.

We present the ADMM algorithm that solves RSLDA in Algorithm 3, and the rationality of the stopping criteria of Algorithm 3 is given in Section II of the Supplementary Material. It should be noted that the only difference between Algorithms 1 and 3 is Step II(c), that is, the solution of $\mathbf{w}$. We now give the solutions of iterative steps (a)–(c) in Algorithm 3.

For solving (a) in Algorithm 3, the solution of $\mathbf{z}$ is exactly the same as in RLDA, that is, $\mathbf{z}^{k+1} = \mathbf{G}^{-1}\mathbf{g}^k$.

For solving (b) in Algorithm 3, the solution of $\mathbf{y}$ is given by (13).

For solving (c) in Algorithm 3, it can be seen that the involved problem is separable with respect to its own variable components, and hence, its solution can be obtained elementwisely by direct computation. Specifically

$$\mathbf{w}_l^{k+1} = S_{\delta/\rho}\left[\left(\mathbf{z}^{k+1} - \mathbf{u}_2^k\right)_l\right] \tag{22}$$

for $l = 1, 2, \ldots, n$, where

$$S_\kappa(a) = \begin{cases} a - \kappa, & a > \kappa \\ 0, & |a| \le \kappa \\ a + \kappa, & a < -\kappa. \end{cases}$$

By observing the solution (22) of $\mathbf{w}$ in each iteration, we can obtain a sparse solution for RSLDA, which is also the main difference between RLDA and RSLDA. To obtain multiple discriminant directions, similar to RLDA, we employ the recursive procedure as in Algorithm 2 but with Algorithm 1 in Step II(b) replaced by Algorithm 3.

V. Discussion

To further clarify the significance of our method, in this section, we discuss the difference between our RLDA and RSLDA and the closely related supervised LDA methods in detail.

A. Relationship

1) Difference From LDA-L1: LDA-L1 was proposed in [30]–[32]. It considers maximizing the L_1-norm between-class scatter and minimizing the L_1-norm within-class scatter, which results in more robust performance. Since the problem involves the division operation of two L_1-norm formulations and the objective is neither convex nor differential, a simple iterative algorithm regarding its rationality was given. However, the algorithm is based on the gradient ascending (GA) technique of nonconvex surrogate functions, and hence, the optimal solution cannot be guaranteed, and the proper step size was hard to choose in practice. Different from LDA-L1, our RLDA and RSLDA consider the minus formulations rather than the ratio one, which makes our optimization problems capable of choosing an appropriate regularization term and avoids solving the division form of the optimization problem with an L_1-norm. Moreover, our RSLDA not only possesses robustness but also has sparseness, which is not the case in LDA-L1.

2) Difference From L1-LDA: Similar to LDA-L1, L1-LDA [34] also considers the ratio form of LDA based on the L_1-norm but uses the whole data scatter $\mathbf{S}_t$ instead of the within-class scatter. L1-LDA is derived via the Bayes error bound optimization. However, the Bayes error bound theoretical derivation of L1-LDA is established under the rigorous condition that each class has an equal number of samples; this is, however, not always suitable. In contrast, our RLDA is an improvement of L1-LDA since it is realized through the Bhattacharyya error bound optimization without the requirement that each class has an equal number of samples. Furthermore, L1-LDA is solved through an iterative technique by converting it to a series of quadratic problems while our RLDA and RSLDA are solved through ADMM.

3) Difference From SULDA-ADMM and SULDA_ℓ_1: Sparse uncorrelated LDA (SULDA)-ADMM and SULDA_ℓ_1 are two algorithms that solve the SULDA presented in [39]. SULDA incorporates the sparseness into an equivalent characterization of the ULDA [60] transformation by seeking the sparse solution with the minimum L_1-norm. In contrast to SULDA-ADMM, SULDA_ℓ_1 is the accelerated linearized Bregman algorithm. Therefore, SULDA-ADMM usually takes more time than SULDA_ℓ_1 to obtain a comparable result but achieves a more accurate solution. However, although SULDA has sparseness, it does not consider robustness. Compared to SULDA, rather than using the L_2-norm, our RLDA and RSLDA use the L_1-norm to characterize the between-class scatter. Therefore, our RSLDA possesses robustness and sparseness simultaneously.

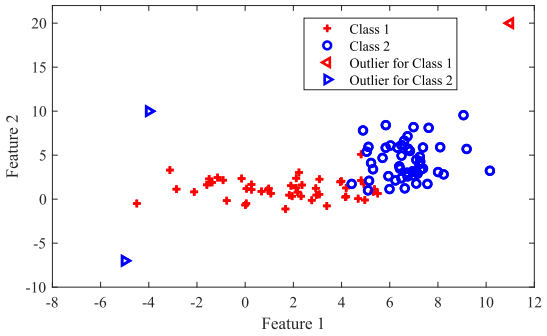


Fig. 1. Artificial data set 1. The data set contains two classes, with each class having 50 samples, and one outlier for Class 1 and two outliers for Class 2.

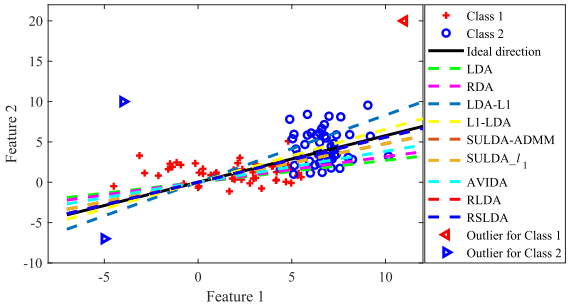


Fig. 2. First discriminant direction obtained by LDA, RDA, LDA-L1, L1-LDA, SULDA-ADMM, SULDA_ℓ_1, AVIDA, RLDA, and RSLDA on artificial data set 1.

4) Difference From AVIDA: Absolute value inequalities discriminant analysis (AVIDA) proposed in [61] is also constructed based on the L_1-norm but uses the difference of the L_1-norm between-class scatter and the L_1-norm within-class scatter. It also considers an extra L_1-norm regularization term and, therefore, makes AVIDA sparse. By using the successive linear algorithm (SLA), AVIDA is transferred into a series of support vector machine (SVM)-type linear programming problems and hence is very time consuming. However, our RLDA and RSLDA are solved through ADMM, which is much more efficient than SLA. In addition, the derivation of our RLDA is theoretically guaranteed.

B. Analysis of Computational Cost

We now present the time complexity analysis of RLDA and RSLDA when one discriminant direction is learned. For RLDA, the time complexity of Step II(a) is $O(n^3)$ since its solution is given by $G^{-1}g$. The computation cost of Step II(b) for one variable is $O(nc)$; therefore, computing Step II(b) takes $O(T_0nc^2)$, where T_0 is the iteration number of Step II(b). The time complexity of Step II(c) is $O(n)$. Therefore, the total time complexity of RLDA is $O(n^3 + T_0nc^2)$. For RSLDA, the only difference between it and RLDA is Step II(c), where its time complexity is also $O(n)$. Therefore, the computation cost of RSLDA is also $O(n^3 + T_0nc^2)$.

The time complexity of LDA-L1 is $O(T((m + c)n))$ [31], [32], where T is the iteration number. For L1-LDA, its time complexity is $O(T(n^3 + n^2(m + c) + nc))$ [34]. For LDA, the time complexity is $O((m + c + 1)n)$.

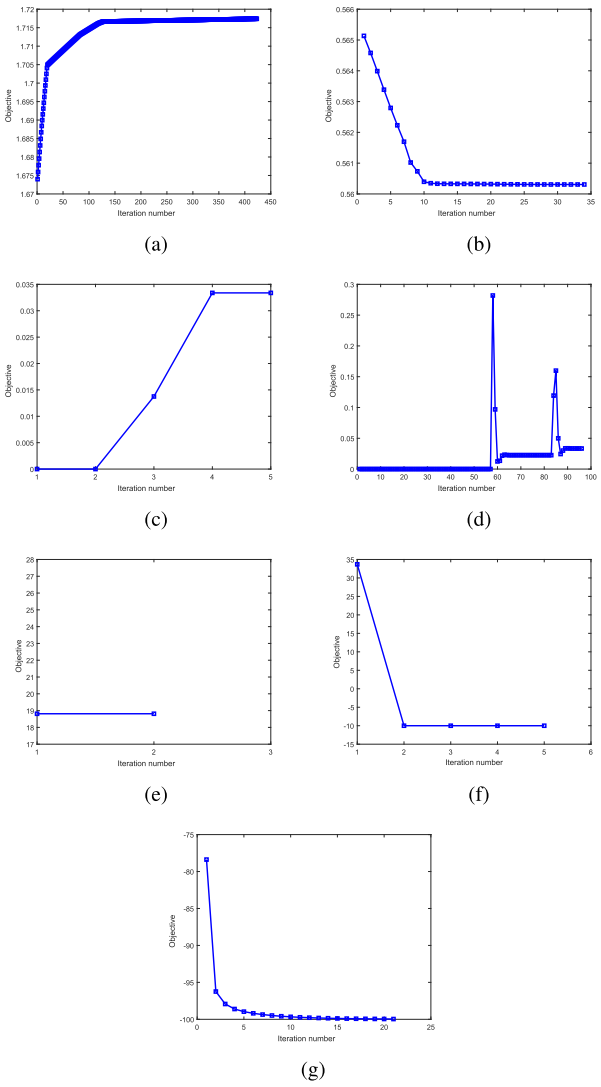


Fig. 3. Variation of objective values along the iteration numbers for LDA-L1, L1-LDA, SULDA-ADMM, SULDA_ℓ_1, AVIDA, RLDA, and RSLDA on artificial data 1. (a) LDA-L1. (b) L1-LDA. (c) SULDA-ADMM. (d) SULDA_ℓ_1. (e) AVIDA. (f) RLDA. (g) RSLDA.

For SULDA_ℓ_1, its time complexity is $O(n^3/q + q^2T)$, where q is the rank of $\mathbf{S}_b$ and T is the iteration number. For SULDA-ADMM, the main computation cost is approximately $O(4n\gamma + 2n + \gamma)$, where γ is the rank of $\mathbf{S}_t$.

For AVIDA, the main computational cost involves solving the linear programming problems, which depends on the number of unknown variables and linear constraints. With a fast interior method that can be used for solving the linear programming, the time complexity to compute one discriminant vector is approximately $O(T((2n+m)^{3.5})(n+2m)^2)$, where T is the iteration number. Therefore, AVIDA has the highest computation cost.

VI. Experimental Results

In this section, experimental results are illustrated to evaluate the performance of the proposed methods. Several related DR methods, including PCA [1], PCA-L1 [29], LDA [8], [9], RDA (regularized LDA) [17], LDA-L1 [30]–[32], L1-LDA [34], SULDA-ADMM [39], SULDA_ℓ_1 [39], and AVIDA [61] are used for comparison. The regularization parameter for RDA and two parameters ν and μ for AVIDA are all selected from the set $\{10^{-5}, 10^{-4}, \ldots, 10^{-1}, 1, 10\}$, and the learning rate parameter for LDA-L1 is chosen from the set $\{10^{-8}, 10^{-7}, \ldots, 10^{2}\}$. For SULDA-ADMM, we set the parameters $\delta = 1$ and $\rho = 2$, and for SULDA_ℓ_1, we set $\delta = 0.9$ and $\tau = 1$, as chosen in [39]. To avoid 0 being a denominator in L1-LDA, a small enough positive number is added in the algorithm of L1-LDA. For our RLDA and RSLDA, parameters ρ and λ are selected from the sets $\{0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000\}$ and $\{0.1, 0.5, 1, 5, 10, 50, 100\}$, respectively, and we choose δ for RSLDA from the set $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. In all the experiments, the parameters for all the methods are optimally selected from their own sets.

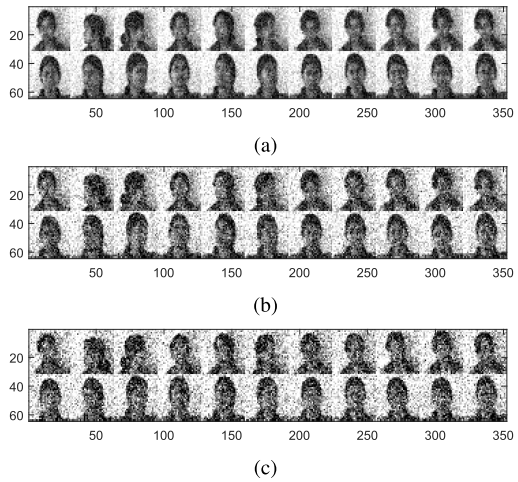


Fig. 4. Sample faces from the Indian females database with noise added to the whole face. Here, the noise is zero-mean Gaussian distributed with variances 0.01, 0.03, and 0.05, respectively. (a) Samples with noise being zero-mean Gaussian distributed with variance 0.01. (b) Samples with noise being zero-mean Gaussian distributed with variance 0.03. (c) Samples with noise being zero-mean Gaussian distributed with variance 0.05.

All the methods are carried out on a PC with a P4 2-GHz CPU and 2-GB RAM memory using MATLAB 2016a. The behaviors of various methods are investigated on two artificial data sets as well as on three contaminated human face databases, including the Indian females database, the ORL database, and the IMM database. For the face databases, original face images are upsampled to 32×32 pixels with 256 gray levels per pixel. Each image is represented by a 1024-D row vector and further scaled to $[0, 1]$ by dividing by 255. To test the performance of various methods on the face recognition, we first project test images to a new space obtained by DR methods; then, the nearest neighbor classifier under the Euclidean metric is applied to identify the newly created image.

A. Artificial Data Sets

1) Artificial Data 1: First, we investigate the robustness of the proposed RLDA and RSLDA on 2-D data. The data contain two classes, with 50 samples in each class. Class 1 is generated from a Gaussian distribution of mean $(\sqrt{3}, 1)$ and covariance $\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$, while Class 2 is generated from a Gaussian distribution

TABLE I

COMPARISON OF DIFFERENT METHODS IN TERMS OF AVERAGE ACCURACIES (%) ALONG DIFFERENT NUMBERS OF TRAINING SAMPLES PER CLASS ON THE CONTAMINATED INDIAN FEMALES DATABASE WITH EACH FACE IMAGE COVERED WITH ZERO-MEAN GAUSSIAN NOISE DISTRIBUTED WITH VARIANCE 0.01

Method	Acc& Time	Number of training samples per class									Mean
		2	3	4	5	6	7	8	9	10	
WithoutDR	Acc	54.04	55.68	58.44	59.09	60.00	60.23	60.61	61.36	63.64	59.23
PCA	Acc	46.27	46.02	53.90	50.60	55.09	58.41	53.64	53.64	57.27	52.76
	Time	0.0001	0.0002	0.0002	0.0003	0.0005	0.0006	0.0007	0.009	0.0011	
PCA-L1	Acc	29.90	49.32	49.87	49.09	49.09	50.23	52.12	52.27	49.09	47.89
	Time	0.0010	0.0012	0.0024	0.0025	0.0028	0.0038	0.0048	0.0067	0.0104	
LDA	Acc	44.44	28.98	48.70	52.23	57.27	56.82	53.03	59.09	40.91	49.06
	Time	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0011	
RDA	Acc	43.94	55.68	62.34	54.54	50.91	52.27	56.06	61.36	54.54	54.63
	Time	0.0001	0.0001	0.0001	0.0002	0.0003	0.0003	0.0004	0.0005	0.0013	
LDA-L1	Acc	46.46	46.02	51.95	53.03	52.73	54.54	54.55	59.09	68.18	54.06
	Time	0.0013	0.0014	0.0022	0.0029	0.0029	0.0030	0.0035	0.0035	0.0037	
L1-LDA	Acc	50.00	53.41	55.19	56.82	54.54	59.09	61.61	61.36	63.63	57.29
	Time	1.4495	1.4827	1.5044	1.5070	1.5218	1.5492	1.6468	1.6475	1.6851	
SULDA-ADMM	Acc	63.13	69.88	67.53	66.67	72.73	64.77	74.24	75.00	77.27	70.14
	Time	0.0625	0.9292	0.9751	0.1196	0.1263	0.1555	0.1838	0.1966	0.2152	
SULDA_ℓ_1	Acc	55.05	60.80	61.69	60.61	69.09	61.36	69.70	72.73	63.64	63.85
	Time	0.0264	0.0370	0.0393	0.0405	0.0667	0.0726	0.0940	0.0944	0.1017	
AVIDA	Acc	53.03	59.09	68.83	69.70	63.64	72.73	69.15	68.18	75.18	67.17
	Time	5.0011	8.0735	9.5099	18.5889	26.5226	27.2290	47.7882	57.3626	61.8797	
RLDA	Acc	73.74	**77.84**	79.87	**86.36**	**90.00**	**88.64**	**90.91**	**95.45**	**95.45**	**86.47**
	Time	0.2574	0.2841	0.3142	0.3223	0.3838	0.3904	0.3987	0.4049	0.4717	
RSLDA	Acc	**76.77**	77.27	**87.66**	83.33	84.55	86.36	**90.91**	93.18	**95.45**	86.16
	Time	0.2347	0.2440	0.2409	0.2529	0.2578	0.2602	0.3544	0.3663	0.3895	

of mean $(4\sqrt{3}, 4)$ and covariance $\begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$, as shown in Fig. 1. One extra outlier of Class 1 and two extra outliers of Class 2 are also added for the purpose of evaluating the robustness of various methods. Here, we only consider supervised DR methods.

Obviously, when no outliers are added, the included angle between the ideal projection direction of the above-mentioned data and the x-axis is 30°. We now apply LDA, RDA, LDA-L1, L1-LDA, SULDA-ADMM, SULDA_ℓ_1, AVIDA, and our RLDA and RSLDA to the above-mentioned artificial data and plot the directions of these projection discriminant vectors in Fig. 2. Furthermore, we obtain their included angles θ between their first projection discriminant vectors and the x-axis and obtain $\theta_{\text{LDA}} = 15.16°$, $\theta_{\text{RDA}} = 17.55°$, $\theta_{\text{LDA-L1}} = 39.76°$, $\theta_{\text{L1-LDA}} = 33.71°$, $\theta_{\text{SULDA\_}\ell_1} = 26.44°$, $\theta_{\text{SULDA-ADMM}} = 25.60°$, $\theta_{\text{AVIDA}} = 21.00°$, $\theta_{\text{RDA}} = 29.14°$, and $\theta_{\text{RSLDA}} = 29.14°$. Clearly, the results suggest the robustness of our proposed RLDA and RSLDA. To evaluate the convergence of the proposed methods, for the methods that need iteration, we also record the variation of the objective values along the iteration numbers, as shown in Fig. 3. The results demonstrate the convergence of RLDA and RSLDA. To further demonstrate the sparseness of RSLDA, we also apply it to a 6-D artificial data set with sparseness and intrinsic outliers. The results show the sparseness of RSLDA and the convergence of RLDA and RSLDA on this data set (see Section III-A of the Supplementary Material). We finally compute the Bayes error, the Bhattacharyya error, and our RLDA error on this data set, which are 0.260867, 0.404764, and 0.481119, respectively. This result shows that our estimation is slightly larger than that of the Bhattacharyya estimation.

B. Human Face Databases

To further show the effectiveness of the proposed methods, we apply various methods to three different types of contaminated human face databases. The classification accuracy on the projected face image space is used as an indicator to test the performance of various methods, and the nearest neighbor classifier with the Euclidean metric is used.

1) Indian Females: The Indian females database contains 22 distinct subjects and each subject has 11 different images. The original size of each image is 640 × 480 pixels, with 256 gray levels per pixel. Here, we upsample all the images into 32 × 32 pixels.[1] A random subset with p $(2, 3, \ldots, 10)$ images per subject is taken to form the training set, while the rest of the data comprise the test set. For each given p, the average result over ten random splits is considered. To test the robustness of various methods, we add three different levels of random noise to each whole face: the noise is zero-mean Gaussian distributed with variances 0.01, 0.03, and 0.05, and Fig. 4 shows some training sample faces with different noise.

In this experiment, we fix the reduced dimension to 50 for each method, except for LDA, RDA, SULDA_ℓ_1, and SULDA-ADMM for which the number of dimensions is reduced to at most $c - 1$, where c is the number of subjects. In addition, we also show the classification results when no DR methods are applied and refer to them as "WithoutDR." For each training number, Table I lists the average results over ten random experiments for the contaminated Indian females database when the noise level is 0.01, and the mean

[1] http://www.optimal-group.org/Resources/Code/RSLDA.html

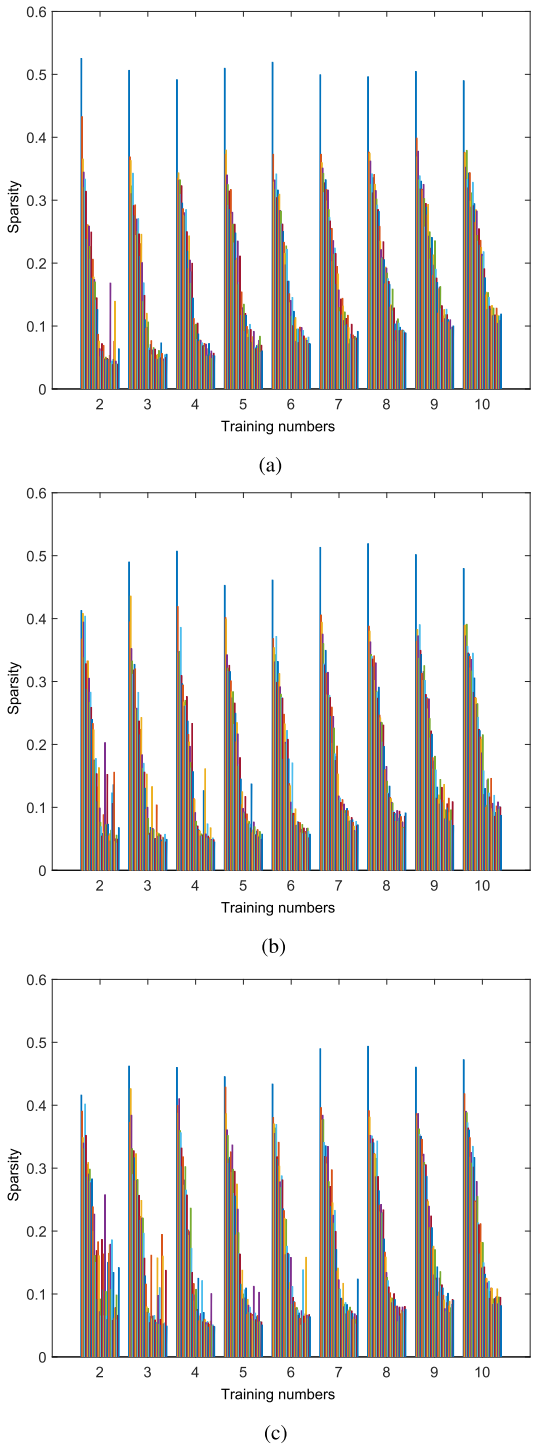


Fig. 5. Average sparsity over 10 random experiments of the Indian females database when noise is added on each whole training face image, for training number 2, 3, . . . , 10 on 50 discriminant directions. Here, the noise is zero-mean Gaussian distributed with variances 0.01, 0.03, and 0.05, respectively. (a) Sparsity results with noise being zero-mean Gaussian distributed with variance 0.01. (b) Sparsity results with noise being zero-mean Gaussian distributed with variance 0.03. (c) Sparsity results with noise being zero-mean Gaussian distributed with variance 0.05.

accuracy is over all training numbers. The best mean accuracy is shown in bold. By observing the results, we can see that as the training number per class increases, the accuracies are basically ascending for all of the methods. However, our RLDA and RSLDA perform much better than the other methods. For example, even when the training number is 3 per class, the accuracies of RLDA and RSLDA are higher than those of the other methods when their training numbers are 10 per class. The mean accuracies of RLDA and RSLDA over all training numbers are almost 20% higher than the rest of the methods. Furthermore, our RLDA and RSLDA both behave better than WithoutDR, which demonstrates the effectiveness of the proposed methods on resisting outliers even when the dimension is reduced. In addition, we see that RSLDA can achieve comparable behavior to RLDA, and thus, we conclude that the sparseness imposed on RLDA does not cause much discriminant information loss. We also study the corresponding results for the cases with noise variances of 0.03 and 0.05, which are given in Tables S2 and S3 (in the Supplementary Material). The results support a similar conclusion.

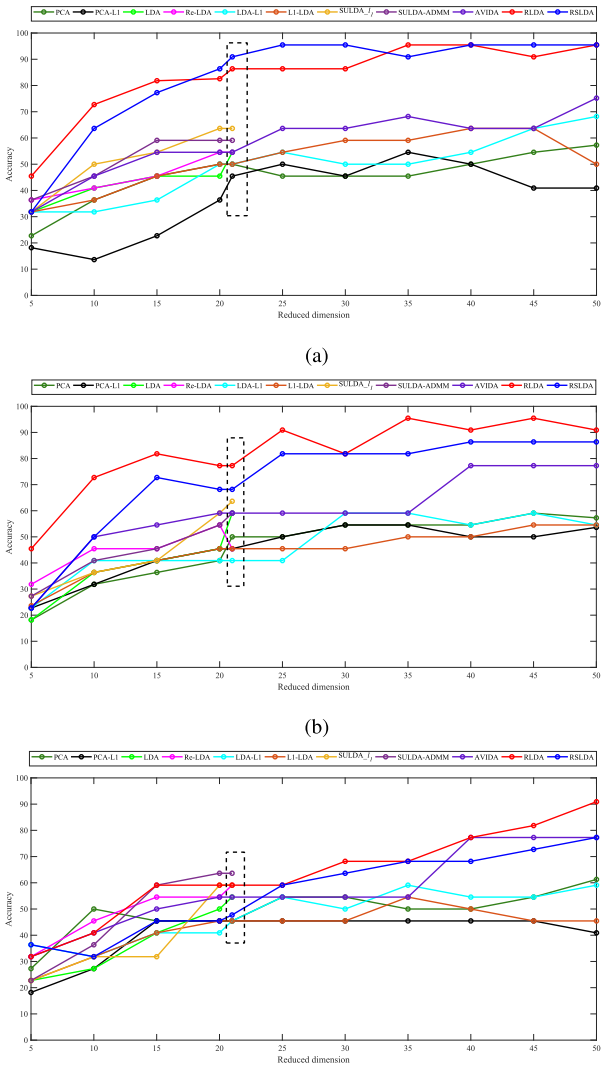


Fig. 6. Accuracies of the Indian females database for different reduced dimensions when noise is added on each training face image. Here, the noise is zero-mean Gaussian distributed with variances 0.01, 0.03, and 0.05, respectively. (a) Accuracies versus dimensions results with noise being zero-mean Gaussian distributed with variance 0.01. (b) Accuracies versus dimensions results with noise being zero-mean Gaussian distributed with variance 0.03. (c) Accuracies versus dimensions results with noise being zero-mean Gaussian distributed with variance 0.05.

We also present the computational time for each method in these three tables. For each method, it is clear that as the number of training samples increases, the computational time also increases. For methods without iteration, such as PCA, LDA, and RDA, they run the fastest. Among the iteration methods, PCA-L1 and LDA-L1 run faster than the others since their update rules are achieved by adding a vector. SULDA-ADMM and SULDA_ℓ_1 have less computational time than our RLDA and RSLDA since they obtain all the discriminant directions once for all. L1-LDA and AVIDA have the lowest computational speed due to the solving of quadratic problems and linear programming problems, respectively.

To measure the sparsity of the proposed RSLDA, we give the ratio of nonzero elements in each obtained discriminant direction. For each training number, there are 50 such ratios that are obtained since 50 discriminant directions are set. We depict these ratios by computing average values over ten random experiments for the contaminated Indian females database, as shown in Fig. 5. By observing the figure, we can see that in each discriminant direction, at most 50% of its elements are nonzero. Therefore, RSLDA can achieve a fair amount of sparseness. As the number of reduced dimensions increases, nonzero elements decrease. This reflects that RSLDA can extract the most useful information effectively.

For the purpose of investigating how the reduced dimension affects classification accuracies, we compare the recognition abilities when applying the proposed methods and the other methods for DR with the feature dimension varying in the set $\{5, 10, \ldots, 50\}$. Since LDA and RLDA can reduce the dimension to at most 21, we also give the corresponding accuracies of 21 dimensions. The results are shown in Fig. 6. From the figure, we can see that for all of the methods, the accuracies all have ascending trends in terms of the increase in reduced dimensions in general. However, regardless of the noise level changes, our RLDA and RSLDA always perform the best. For example, when the noise level is 0.01, it is can be seen that both RLDA and RSLDA are much better than the other methods for almost all of the dimensions. We also note that when the reduced dimension is 21, the accuracies of RLDA and RSLDA are at least 20% higher than those of LDA and RDA. The above-mentioned results demonstrate the superiority of our RLDA and RSLDA. Similar results are achieved for the cases when noise variances are 0.03 and 0.05. We also conduct an objective convergence experiment on the database in Section III-D of the Supplementary Material. The results show that the proposed methods are convergent by choosing appropriate parameters.

To further verify the robust performance of our RLDA and RSLDA, we add two other different types of noise to the ORL database and the IMM database[2] and test their discriminative ability. The results support the above-mentioned observation that both RLDA and RSLDA are effective and possess robustness, while RSLDA has sparseness (Section III-C of the Supplementary Material).

[2]http://www.optimal-group.org/Resources/Code/RSLDA.html

VII. Conclusion

This paper proposed a robust discriminant analysis criterion (RLDA) based on the L_1-norm. Its criterion is an upper bound of the theoretical framework of the Bhattacharyya optimality. RLDA can be easily extended to its sparse version RSLDA by considering an extra sparse regularization term. This makes our methods more robust to outliers and noise than LDA and other related methods and capable of extracting sparse features. The experimental results show the superiority of RLDA and RSLDA. Our MATLAB code and slides can be downloaded from http://www.optimal-group.org/Resources/Code/RSLDA.html. In the future, we will study more efficient algorithms for solving RLDA and RSLDA.

Acknowledgment

The authors would like to thank the referees for their valuable comments that have largely improved the presentation of this paper and Prof. X. Zhang for the generous help.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.
[2] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[3] Z. Zhang, M. Zhao, and T. W. S. Chow, "Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2192–2205, Oct. 2013.
[4] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.
[5] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2005, pp. 1208–1213.
[6] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, Jan. 2010.
[7] Z. Zhang, S. Yan, and M. Zhao, "Pairwise sparsity preserving embedding for unsupervised subspace learning and classification," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4640–4651, Dec. 2013.
[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic, 1991.
[10] Z. Zhang, M. Zhao, and T. W. S. Chow, "Constrained large margin local projection algorithms and extensions for multimodal dimensionality reduction," *Pattern Recognit.*, vol. 45, no. 12, pp. 4466–4493, Dec. 2012.
[11] S. Zeiler, R. Nicheli, N. Ma, G. J. Brown, and D. Kolossa, "Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 2797–2801.
[12] Z. Zhang and W. S. Chow, "Tensor locally linear discriminative analysis," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 643–646, Nov. 2011.
[13] R. Huang, C. Liu, and J. Zhou, "Discriminant analysis via jointly $L_{2,1}$-norm sparse tensor preserving embedding for image classification," *J. Vis. Commun. Image Represent.*, vol. 47, pp. 10–22, Aug. 2017.
[14] J. Zhao, L. Shi, and J. Zhu, "Two-stage regularized linear discriminant analysis for 2-D data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1669–1681, Aug. 2015.
[15] R. H. Randles, J. D. Broffitt, J. S. Ramberg, and R. V. Hogg, "Generalized linear and quadratic discriminant functions using robust estimates," *J. Amer. Stat. Assoc.*, vol. 73, no. 363, pp. 564–568, 1978.
[16] S. Yu, Z. Cao, and X. Jiang, "Robust linear discriminant analysis with a Laplacian assumption on projection distribution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2567–2571.
[17] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.
[18] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, Dec. 2002.

[19] S.-J. Kim, A. Magnani, and S. Boyd, "Robust Fisher discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2005, pp. 659–666.

[20] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[21] Z. Wang, Q. Ruan, and G. An, "Projection-optimal local Fisher discriminant analysis for feature extraction," *Neural Comput. Appl.*, vol. 26, no. 3, pp. 589–601, Apr. 2015.

[22] Z. Zhang and T. W. S. Chow, "Robust linearly optimized discriminant analysis," *Neurocomputing*, vol. 79, no. 3, pp. 140–157, Mar. 2012.

[23] F. Z. Okwonu and A. R. Othman, "Comparative performance of classical Fisher linear discriminant analysis and robust Fisher linear discriminant analysis," *Matematika*, vol. 29, pp. 213–220, Jun. 2013.

[24] G.-F. Lu, J. Zou, and Y. Wang, "L1-norm and maximum margin criterion based discriminant locality preserving projections via trace Lasso," *Pattern Recognit.*, vol. 55, pp. 207–214, Jul. 2016.

[25] G.-F. Lu, G. Tang, and J. Zou, "Spare L1-norm-based maximum margin criterion," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 11–17, Jul. 2016.

[26] L. Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," *IEEE Trans. Neural Netw.*, vol. 6, no. 1, pp. 131–143, Jan. 1995.

[27] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant L_1 norm," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2571–2579, Aug. 2010.

[28] X. Li, S. Fei, and T. Zhang, "Weighted maximum scatter difference based feature extraction and its application to face recognition," *Mach. Vis. Appl.*, vol. 22, no. 3, pp. 591–595, 2011.

[29] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[30] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.

[31] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.

[32] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[33] C.-N. Li, Y.-H. Shao, and N.-Y. Deng, "Robust L1-norm two-dimensional linear discriminant analysis," *Neural Netw.*, vol. 65, pp. 92–104, May 2015.

[34] W. Zheng, Z. Lin, and H. Wang, "L1-norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 793–805, Apr. 2014.

[35] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[36] R. Glowinski and A. Marroco, "On the approximation by finite elements of order one, and resolution, penalisation-duality for a class of nonlinear Dirichlet problems," *ESAIM, Math. Model. Numer. Anal.-Math. Model. Numer. Anal.*, vol. 9, no. R2, pp. 41–76, 1975.

[37] B. He and X. Yuan, "On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method," *SIAM J. Numer. Anal.*, vol. 50, no. 2, pp. 700–709, Apr. 2012.

[38] D. Davis, "Convergence rate analysis of primal-dual splitting schemes," *SIAM J. Optim.*, vol. 25, no. 3, pp. 1912–1943, Sep. 2015.

[39] X. Zhang, S. Chu, and R. C. E. Tan, "Sparse uncorrelated linear discriminant analysis for undersampled problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1469–1485, Jul. 2016.

[40] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, Jan. 2016.

[41] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J. Sci. Comput.*, vol. 78, no. 1, pp. 29–63, Jan. 2019. doi: 10.1007/s10915-018-0757-z.

[42] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers Math. China*, vol. 7, no. 2, pp. 365–384, Apr. 2012.

[43] A. P. Liavas and N. D. Sidiropoulos, "Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5450–5463, Oct. 2015.

[44] R. Chartrand and B. Wohlberg, "A nonconvex ADMM algorithm for group sparsity with sparse groups," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6009–6013.

[45] B. Moghaddam, Y. Weiss, and S. Avidan, "Generalized spectral bounds for sparse LDA," in *Proc. 23rd Int. Conf. Mach. Learn.*, Jun. 2006, pp. 641–648.

[46] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, Jan. 2011.

[47] L. Tian and Q. Gu, "Communication-efficient distributed sparse linear discriminant analysis," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, pp. 1178–1187, Apr. 2017.

[48] J. Cai and X. Huang, "Modified sparse linear-discriminant analysis via nonconvex penalties," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4957–4966, Oct. 2018.

[49] H. Xiong, W. Cheng, J. Bian, W. Hu, Z. Sun, and Z. Guo, "DBSDA: Lowering the bound of misclassification rate for sparse linear discriminant analysis via model debiasing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 707–717, Mar. 2019. doi: 10.1109/TNNLS.2018.2846783.

[50] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. C-24, no. 3, pp. 281–289, Mar. 1975.

[51] J. Ye, "Least squares linear discriminant analysis," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 1087–1093.

[52] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 1982.

[53] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distribution," *Bull. Calcutta Math. Soc.*, vol. 35, no. 15, pp. 99–109, 1943.

[54] M.-J. Lai and J. Wang, "An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems," *SIAM J. Optim.*, vol. 21, no. 1, pp. 82–101, 2011.

[55] M.-J. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, Mar. 2013.

[56] X. Chen, J. Yang, Q. Ye, and J. Liang, "Recursive projection twin support vector machine via within-class variance minimization," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2643–2655, 2011.

[57] Y.-H. Shao, N.-Y. Deng, and Z.-M. Yang, "Least squares recursive projection twin support vector machine for classification," *Pattern Recognit.*, vol. 45, no. 6, pp. 2299–2307, Jun. 2012.

[58] Q. Tao, D. Chu, and J. Wang, "Recursive support vector machines for dimensionality reduction," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 189–193, Jan. 2008.

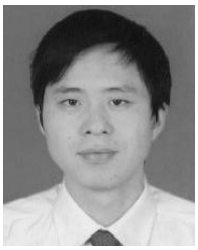
[59] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data." *Int. J. Appl. Math.*, vol. 39, no. 1, pp. 48–60, Jan. 2009.

[60] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.

[61] C.-N. Li, Z. R. Zheng, M.-Z. Liu, Y.-H. Shao, and W.-J. Chen, "Robust recursive absolute value inequalities discriminant analysis with sparseness," *Neural Netw.*, vol. 93, pp. 205–218, Sep. 2017.

Chun-Na Li received the master's and Ph.D. degrees from the Department of Mathematics, Harbin Institute of Technology, Harbin, China, in 2009 and 2012, respectively.

She is currently an Associate Professor with the School of Management, Hainan University, Haikou, China, and the Zhijiang College, Zhejiang University of Technology, Hangzhou, China. Her current research interests include data mining, machine learning, and optimization methods.



Yuan-Hai Shao received the B.S. degree in information and computing science from the College of Mathematics, Jilin University, Changchun, China, in 2006, and the master's degree in applied mathematics and the Ph.D. degree in operations research and management from the College of Science, China Agricultural University, Beijing, China, in 2008 and 2011, respectively.

He is currently a Professor with the School of Management, Hainan University, Haikou, China. His current research interests include support vector machines, nonparallel support vector machines, data mining, machine learning, and optimization methods. He has authored or coauthored over 100 refereed papers on these areas.

Wotao Yin received the Ph.D. degree in operations research from Columbia University, New York, NY, USA, in 2006.

From 2006 to 2013, he was with Rice University, Houston, TX, USA. He is currently a Professor with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA, USA. His current research interests include computational optimization and its applications in signal processing, machine learning, and other data science problems. He invented fast algorithms for sparse optimization and large-scale distributed optimization problems.

Dr. Yin was a recipient of the NSF CAREER Award in 2008, the Alfred P. Sloan Research Fellowship in 2009, and the Morningside Gold Medal in 2016, and has coauthored five papers receiving best paper-type awards. He is among the top 1

Ming-Zeng Liu received the master's and Ph.D. degrees from the School of Mathematical Sciences and the School of Automotive Engineering, Dalian University of Technology, Dalian, China, in 2008 and 2013, respectively.

He is currently an Associate Professor with the School of Mathematics and Physics Sciences, Dalian University of Technology. His current research interests include machine learning and computational geometry.