Pattern Recognition Letters 138 (2020) 651–656

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Robust discriminant analysis using multi-directional projection pursuit

Hsin-Hsiung Huang[a,*], Teng Zhang[b]

[a] Department of Statistics and Data Science, University of Central Florida, United States
[b] Department of Mathematics, University of Central Florida, United States

ARTICLE INFO

Article history:
Received 30 July 2019
Revised 20 July 2020
Accepted 8 September 2020
Available online 18 September 2020

Keywords:
Classification
Dimension reduction
Optimal scores
Projection pursuit
Robustness

ABSTRACT

While linear discriminant analysis (LDA) is a widely used classification method, it is highly affected by outliers which commonly occur in various real datasets. Therefore, several robust LDA methods have been proposed. However, they either rely on robust estimation of the sample means and covariance matrix which may have noninvertible Hessians or can only handle binary classes or low dimensional cases. The proposed robust discriminant analysis is a multi-directional projection-pursuit approach which can classify multiple classes without estimating the covariance or Hessian matrix and work for high dimensional cases. The weight function effectively gives smaller weights to the points more deviant from the class center. The discriminant vectors and scoring vectors are solved by the proposed iterative algorithm. It inherits good properties of the weight function and multi-directional projection pursuit for reducing the influence of outliers on estimating the discriminant directions and producing robust classification which is less sensitive to outliers. We show that when a weight function is appropriately chosen, then the influence function is bounded and discriminant vectors and scoring vectors are both consistent as the percentage of outliers goes to zero. The experimental results show that the robust optimal scoring discriminant analysis is effective and efficient.

1. Introduction

Linear discriminant analysis (LDA) [10] is a widely used classification method, which searches a linear boundary with the optimal discrimination between two classes [13]. However, the classical LDA could be improved by solving the following issues. First, LDA is sensitive to the outlying observations. There are two approaches of robust discriminant analysis [26]. Approach I: To replace the unknown population parameters, group means (μ_1, μ_2) and covariance matrix (Σ) by estimators of multivariate location and scatter [6,11,23,25,29–32,35,36] such as Robust Regularized LDA (RRLDA) [5,11,12], Robust Mixture Discriminant Analysis (RMDA) [2], and Robust Linear Discriminant Analysis (Linda) [31]. Croux and Dehon [6] proposed robust linear discriminant analysis using s-estimators of the means and covariance. Kim et al. [22] proposed a robust LDA method which searches the worst-case performance of a discriminant over all possible means and covariances. Guo et al. [12] developed shrunken centroids regularized discriminant analysis. However, it does not work well for high dimensional data, since there may exist ill-conditioned covariance matrices in this setting [26]. Approach II: To replace the unknown univariate population parameters along the projections, $a^T\mu_1$, $a^T\mu_2$, and $a^T\Sigma a$ by univariate estimators of location and scatter. Approach II is a projection-pursuit approach. Pires and Branco [26] proposed the robust LDA with Projection Pursuit (LDAPP), which is a one-directional projection pursuit of location and scatter, and hence only can handle binary classification. In this study, we proposed a novel robust discriminant analysis by multi-directional projection-pursuit of optimal scores, and named it as the *robust optimal scoring discriminant analysis* (ROSDA), which can classify multiple classes. It provides a supervised projection of the predictors by using the Ψ loss function which is less sensitive to outliers in the fashion of the previous work of robust principal component analysis [19,20]. The Ψ loss function results in a weight function which effectively gives smaller weights to the points more deviant from the class center. ROSDA inherits good properties of the Ψ loss function and multi-directional projection pursuit for reducing the influence of outliers on estimating the discriminant directions and producing robust classification which is less sensitive to outliers. We propose an iteration algorithm to solve the ROSDA problem effectively and efficiently, since ROSDA is a multi-directional projection-pursuit approach which can classify multiple classes without estimating the covariance or Hessian matrix and work for high dimensional data. We show that the proposed algorithm has the same computational complexity as LDA and RRLDA. Moreover, we derived the associ-

* Corresponding author.
E-mail addresses: hsin.huang@ucf.edu (H.-H. Huang), teng.zhang@ucf.edu (T. Zhang).

https://doi.org/10.1016/j.patrec.2020.09.013
0167-8655/

ated influence function and its properties. All the technical proofs are in Appendix.

2. Methodology

The linear discriminant analysis can be reformulated as an optimal scoring least square problem [14]. Let Y denote a $n \times K$ matrix of dummy variables for the K classes. Y_{ik} an indicator variable defined by $y_{ik} = 1$, if the ith observation belongs to the kth class; $y_{ik} = 0$ otherwise. Assuming that data matrix X is an $n \times p$ with each row X_i corresponding to ith observation in $\mathbb{R}^p$ sampled from K classes, and the data has been centered. The optimal scoring discriminant analysis (OSDA) [14] finds the discriminant scores by finding at most $K-1$ discriminant directions, where K is the number of classes. The OSDA problem is expressed as

$$\min_{\Theta,B} \frac{1}{n}\|Y\Theta - XB\|_F^2, \ \text{s.t.}\ \Theta^\top Y^\top Y\Theta = nI_{Q\times Q}, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\Theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^{K\times Q}$, $B = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^{p\times Q}$, $k = 1, \ldots, Q \le K-1$, θ_k are the scoring vectors, β_k are the discriminant vectors. B is a matrix mapping X_i, $i = 1, \ldots, n$ from $\mathbb{R}^p$ to $\mathbb{R}^Q$, and the rows of Θ represent the K classes in $\mathbb{R}^Q$. Though Problem (1) is not convex, the OSDA problem could be solved iteratively by a decomposition of the least squares problem in θ_k and β_k, since the estimates of β_k and θ_k have closed forms individually [14,15].

2.1. The asymptotic expectation of OSDA

In this section, we show the asymptotic expectation of Problem (1) when Ψ is the identity function. Following Hastie et al. [14], Θ is the eigenvector matrix of $Y^\top P_X Y$. We assume that for all $1 \le k \le K$, each sample is from the kth class with probability $1/K$. In addition, the samples in the kth class follow from a distribution with mean μ_k and covariance I. Then $Y^\top P_X Y/n \xrightarrow{prob.} [\mu_1, \ldots, \mu_K]^\top [\sum_{k=1}^K (I + \mu_k\mu_k^\top)]^{-1}[\mu_1, \ldots, \mu_K]$ ($\xrightarrow{prob.}$: convergence in probability) with $P_X = X(X^\top X)^{-1}X^\top$ [28]. As a result, $\Theta \in \mathbb{R}^{K\times Q}$ is the matrix given by the Q top right eigenvectors of the matrix $[\mu_1, \ldots, \mu_K] \in \mathbb{R}^{p\times K}$, and B is given by $(X^\top X)^{-1}X^\top Y \xrightarrow{prob.} [\sum_{k=1}^K (I + \mu_k\mu_k^\top)]^{-1}[\mu_1, \ldots, \mu_K]\Theta$. Assuming that the singular values of the matrix $[\mu_1, \ldots, \mu_K]$ is $\sigma_1 \ge \sigma_2 \ge \ldots$, then B is the matrix given by $\text{diag}(\frac{\sigma_1}{\sigma_1^2+K}, \frac{\sigma_2}{\sigma_2^2+K}, \ldots, \frac{\sigma_Q}{\sigma_Q^2+K})V^\top$, where $V \in \mathbb{R}^{p\times Q}$ is the top Q left eigenvectors of the matrix $[\mu_1, \ldots, \mu_K] \in \mathbb{R}^{p\times K}$.

3. Robust optimal scoring discriminant analysis

3.1. ROSDA with Ψ function

Like other least squares problems, the OSDA is easily influenced by outliers. Inspired by Huang et al. [19] and Huang and Yeh [20], we propose a supervised dimension reduction method that can be considered as a robust version of OSDA using Ψ function. Higuchi and Eguchi [18] discussed the Ψ loss function and their parameter tuning procedures. Here are a few choices of $\Psi(z)$ and the corresponding weight function $\Psi'(z)$ (the first derivative with respect to z): 1) $\Psi_0(z) = z$ with $\Psi_0'(z) = 1$, 2) $\Psi_1(z) = (1 - e^{-\zeta z})/\zeta$ with $\Psi_1'(z) = e^{-\zeta z}$, 3) $\Psi_2(z) = -\frac{1}{\zeta}\log\{\frac{1+e^{-\zeta(z-\xi)}}{2}\}$ with $\Psi_2'(z) = 1 - \frac{1}{1+e^{-\zeta(z-\xi)}}$. We used Ψ_1 in our data analysis since it requires only one tuning parameter and provides weights which reduce the influences of outliers effectively. Note that $\lim_{\zeta\to 0}\Psi_1 = \Psi_0$, the identify function, is the loss function of the OSDA problem. The Ψ function is applied to the loss function in Problem (1) in order to reduce the influence of outliers. The proposed ROSDA solves the following problem

$$\min_{\Theta,B} \frac{1}{n}\sum_{i=1}^n \Psi(z_i), \ \text{s.t.}\ \Theta^\top Y^\top Y\Theta = nI_{Q\times Q}, \tag{2}$$

where $\Psi(z_i)$ is a monotonic increasing concave, and differentiable function of $z_i = \|Y_i\Theta - X_iB\|_2^2$ with the Euclidean norm $\|\cdot\|_2$, B is the $p \times Q$ matrix with columns consisting of the Q coefficient vectors $\beta_1, \ldots, \beta_Q$, and Θ is the $K \times Q$ matrix with columns consisting of the Q scoring vectors $\theta_1, \ldots, \theta_Q$ with $1 \le Q \le K-1$. We aim to use Ψ functions for reducing influence of outliers on the residual sum of squares (RSS) [18–20]. The Ψ function assigns higher weights to the points with small RSS and gives smaller weights to the points with large RSS.

An iterative algorithm for solving Θ and B in the minimization problem (2) is introduced as follows. First, calculate the weight matrix $W^{(j-1)} = \text{diag}(W_1^{(j-1)}, \ldots, W_n^{(j-1)}) \in \mathbb{R}^{n\times n}$ with the element $W_i^{(j-1)} = \Psi'(\|Y_i\Theta^{(j-1)} - X_iB^{(j-1)}\|_2^2)$, and $\Psi'(z) = \frac{d\Psi(z)}{dz}$. In the jth iteration, the weights $W_1, W_2, \ldots, W_n$ are given by the estimates of $\Theta^{(j-1)}$ and $B^{(j-1)}$ from the $(j-1)$th iteration. Second, in the jth iterative update, the target problem is

$$\min_{\Theta,B} \frac{1}{n}\sum_{i=1}^n \|W_i^{(j-1)}(Y_i\Theta - X_iB)\|_2^2, \ \text{s.t.}\Theta^\top D\Theta = I_{Q\times Q},$$

where $D = \frac{1}{n}Y^\top Y$. As a result, Problem (2) is reduced to

$$\min_{\Theta,B} \frac{1}{n}\|W^{(j-1)}Y\Theta - W^{(j-1)}XB\|_F^2, \ \text{s.t.}\ \Theta^\top D\Theta = I_{Q\times Q}. \tag{3}$$

Here $n \times n$ matrix $P_{W^{(j-1)}X}$ denotes a projection matrix of the subspace spanned by the p columns of $W^{(j-1)}X$. Problem (3) is equivalent to

$$\min_{\Theta} \frac{1}{n}\|(I_{n\times n} - P_{W^{(j-1)}X})W^{(j-1)}Y\Theta\|_F^2, \ \text{s.t.}\ \Theta^\top D\Theta = I_{Q\times Q}. \tag{4}$$

Let P_Y be a matrix of size $\mathbb{R}^{n\times K}$ such that $P_Y^\top P_Y = I_{K\times K}$, and P_Y has the same column space as Y. Here we choose P_Y as the matrix of the left singular vectors of Y. Let $\Theta_0 \in \mathbb{R}^{K\times Q}$ be defined such that $\frac{1}{\sqrt{n}}Y\Theta = P_Y\Theta_0$, then $\Theta_0^\top\Theta_0 = I_{Q\times Q}$ and the problem can be written as

$$\min_{\Theta_0} \|(I_{n\times n} - P_{W^{(j-1)}X})W^{(j-1)}P_Y\Theta_0\|_F^2, \ \text{s.t.}\ \Theta_0^\top\Theta_0 = I_{Q\times Q}. \tag{5}$$

Therefore, Θ_0 can be obtained by the Q smallest right singular vectors of $(I_{n\times n} - P_{W^{(j-1)}X})W^{(j-1)}P_Y$, and then $\Theta = \frac{1}{\sqrt{n}}D^{-1}Y^\top P_Y\Theta_0$. Consequently, we have Algorithm 1 and the following properties.

Algorithm 1 Robust OSDA (ROSDA).

1: Input: Y is a $n \times K$ matrix, X is a $n \times p$ matrix, and $Q = K-1$.
2: Output: $(B_{p\times Q}^{(j)}, \Theta_{K\times Q}^{(j)})$.
3: Initialize $B^{(0)}$ and $\Theta^{(0)}$ as matrices of 1's.
4: Standardize X.
5: Compute $D = \frac{1}{n}Y^\top Y$.
6: Run a loop until the criterion, |old RSS − RSS|/RSS ≤ Tolerance and the number of iterations reach a maximum number, is achieved.
7: Compute $z_i^{(j-1)} = \|Y_i\Theta^{(j-1)} - X_iB^{(j-1)}\|_2^2$.
8: In jth iteration, compute the weight matrix $W = \text{diag}(W_1, \ldots, W_n) \in \mathbb{R}^{n\times n}$ with $W_i = \Psi'(z_i^{(j-1)})$.
9: Assign RSS as the old RSS
10: Compute $B^{(j)} = (X^\top WX)^{-1}X^\top WY\Theta^{(j-1)}$ by backward substitution.
11: Compute P_Y as the matrix of the left singular vectors of Y.
12: Compute Θ_0 as the Q smallest right singular vectors of $(I_{n\times n} - P_{WX})WP_Y$.
13: Compute $\Theta = \frac{1}{\sqrt{n}}D^{-1}Y^\top P_Y\Theta_0$.
14: Update RSS as $\frac{1}{n}\sum_{i=1}^n \Psi(z_i)$, where $z_i = \|Y_i\Theta^{(j)} - X_iB^{(j)}\|_2^2$.
15: Repeat the above procedure 5–13 until the stopping criterion is satisfied.
16: For each data X_i, predict classes by finding which Θ_k is closest to X_iB for $1 \le k \le K$.

Proposition 1. *Assume that the $\Psi(z)$ function is differentiable with $0 < \Psi'(0) < \infty$, and $\Psi(z)$ is strictly increasing and strictly concave in $z \in \mathbb{R}^+$. Then the empirical loss functions generated by Algorithm 1 are strictly decreasing:*

$$L(B^{(1)}, \Theta^{(1)}; \Psi) > L(B^{(2)}, \Theta^{(2)}; \Psi) > \dots,$$

where $L(B^{(j)}, \Theta^{(j)}; \Psi) = \frac{1}{n}\sum_{i=1}^{n} \Psi(z_i)$, *and* $z_i = \|Y_i\Theta^{(j)} - X_iB^{(j)}\|_2^2$.

It is necessary to have strict concavity to get strict inequality. The strict decrease of the loss function guarantees the convergence of Algorithm 1 regardless of where the initial point starts.

Theorem 2. *Given a dataset $\{(X_i, Y_i)\}_{i=1}^n$ or (X, Y) in a matrix form, the optimal solution (B, Θ) satisfies the following stationary equation:*

$$B = (X^\top WX)^{-1}X^\top WY\Theta, \ \Theta = \frac{1}{\sqrt{n}}D^{-1}Y^\top P_Y\Theta_0, \tag{6}$$

where Θ_0 can be obtained by the Q smallest right singular vectors of $(I - P_{WX})WP_Y$, P_Y is the matrix of the left singular vectors of Y, $W_i = \Psi'(z_i)$, $z_i = \|Y_i\Theta - X_iB\|_F^2$, and $D = \frac{1}{n}Y^\top Y$, B is the $K \times Q$ coefficient matrix with columns $\beta_1, \dots, \beta_Q$ and Θ is the $K \times Q$ scoring matrix with columns $\theta_1, \dots, \theta_Q$, and $\Psi'(z) = \frac{d\Psi(z)}{dz}$.

Notice that if (B, Θ) is a solution for Eq. (2), then $(BU, \Theta U)$ is also a solution for all $Q \times Q$ orthogonal matrix U. Therefore, in this study, the solution (B, Θ) expresses the collection of all the $(BU, \Theta U)$ in which the classification is invariant.

3.2. Computational cost

In this section, we describe the computation cost per iteration in Algorithm 1. Step 7 requires a computational cost of $O(nKQ + npQ)$; step 10 requires $O(p^2n + p^3 + pnK + pKQ)$; step 11 requires $O(nK\min(n, K))$; the calculation of $P_Y = Y(Y^TY)^{-0.5}$ requires $O(nK^2 + K^3)$; the calculation in step 12 with given P_Y is

$$WP_Y - (WX)(X^TW^TWX)^{-0.5}(X^TW^T)WP_Y,$$

which has a complexity of $O(nK + np^2 + p^3 + npK)$ (note that W is a diagonal matrix); the calculation of step 13 with given P_Y is $O(K^3 + K^2n)$. Therefore, the computational cost per iteration is in the order of $O(\max(n, p)(K^2 + p^2))$ by adding them together with the implicit assumption that $n > K$. In comparison, the computational cost of LDA or RRLDA per iteration is about the calculation of sample covariance matrix or robust covariance matrix and its matrix inversion, which is in the order of $O(\max(n, p)p^2)$. As a result, ROSDA, LDA, and RRLDA have the same computational complexity.

3.3. Influence function

The influence function (IF) of an estimator is an asymptotic version of its sensitivity curve. It is an approximation to the behavior of the estimator when the sample contains a small fraction ϵ of identical outliers. Assuming that for any distribution H_0 in $\mathbb{R}^p$, the statistical functional T outputs an estimator in $\mathbb{R}^q$, then the influence function is defined by $\mathrm{IF}(z_0, T, H_0) = \lim_{\epsilon\to 0}\frac{T(H_1) - T(H_0)}{\epsilon}$, where $H_1 = (1-\epsilon)H_0 + \epsilon\Delta_{z_0}$ and Δ_{z_0} is a Dirac measure putting all its weights on $z_0 \in \mathbb{R}^p$. We first consider the generic estimation $T(H) = \mathrm{argmin}_{w\in L}\mathbb{E}_{z\in H_0} f(z, w)$, where L is a subspace in $\mathbb{R}^q$ and f is a function from $\mathbb{R}^p \times \mathbb{R}^q$ to $\mathbb{R}$. Then we have the following generic result.

Theorem 3. *Let $F(w) = \mathbb{E}_{z\in H_0} f(z, w)$, P_L be the projector to the subspace L, and $^+$ represents the Moore Penrose inverse (pseudo inverse), then* $\mathrm{IF}(z_0, T, H_0) = [P_L\mathrm{Hessian}_w F(w)|_{w=T(H_0)}P_L]^+[P_L\nabla_w f(z_0, w)|_{w=T(H_0)}]$.

It can be applied to the following setting of our estimator: z corresponds to the samples (X_i, Y_i) in our notation; w corresponds to the estimation (Θ, B); $f(z, w)$ is $\Psi(\|Y\Theta - XB\|_2^2)$ as in the objective function; the constraint $w \in L$ corresponds to the assumption $\Theta^TY^TY\Theta = I$ (when this constraint in not linear, for perturbation result we may assume that L is the tangent plane of the set $\Theta^TY^TY\Theta = I$). Theorem 3 shows that the influence function strongly depends on $f_w(z_0, w)|_{w=T(H_0)}$. Assuming that $(\hat\Theta, \hat B)$ is estimated under distribution H_0, $f_w(z_0, w)|_{w=T(H_0)}$ is

$$\nabla_{\Theta,B}\Psi(\|Y_0\hat\Theta - X_0\hat B\|_2^2) = \Psi'(\|Y_0\hat\Theta - X_0\hat B\|_2^2)\nabla_{\Theta,B}\|Y_0\hat\Theta - X_0\hat B\|_2^2,$$

and only the part $\Psi'(\|Y_0\hat\Theta - X_0\hat B\|_2^2)$ depends on the outlier (X_0, Y_0). The outlier (X_0, Y_0) have a large value of $\|Y_0\Theta - X_0B\|_2^2$. If we choose Ψ such that $\Psi'(z)$ is small and bounded (and decreases to zero) as $z \to \infty$ (which holds for the choices Ψ_1 and Ψ_2), $f_w(z_0, w)|_{w=T(H_0)}$ is small. As a result, the influence defined by the influence function would be small and bounded. The above analysis gives useful insights about the robustness of our estimator. While it is possible to write down the Hessian explicitly (and the influence function) for our estimator, the calculation is rather complicated and does not give too much additional information or intuition about the robustness, so we skip the calculation here.

As a special case, we may assume that ith class contains an equivalent number of samples from $N(\mu_i^*, I)$. When the classes are well-separated, i.e., $\|\mu_i - \mu_j\|$ is much larger than 1. We reformulate the problem as follows. Let $y_i \in [1, \dots, K]$ be the index of the class for the ith sample. With the goal is to find the K centers $\mu_1, \dots, \mu_K$ for all K classes and a matrix $B \in \mathbb{R}^{p\times K-1}$, we solve

$$\underset{\mu_1,\dots,\mu_K,\Theta}{\mathrm{argmin}} \sum_{i=1}^{n}\Psi(\|X_iB - \theta_{y_i}\|^2), \ \text{such that}\ \Theta^\top D\Theta = I_{Q\times Q}.$$

The estimation satisfies $\hat\theta_i \approx B\mu_i^*$ (as shown in Section 2.1 with $\sigma_i > \ > 1$ for all $1 \le i \le Q$), and the solution B is a matrix with image in the span of $\{\mu_i^*\}_{i=1}^K$. If we introduce an outlier (X_0, Y_0), then the component $\nabla_w f(z_0, w)|_{w=T(H_0)}$ in Theorem 3 would be $\Psi'(\|X_0\hat B - \hat\theta_{k_0}\|^2) \approx \Psi'(\|(X_0 - \mu_{k_0}^*)\hat B\|^2)$, where $k_0 \in [1, \dots, K]$ is the index of class Y_0. If (X_0, Y_0) is an outlier, X_0 is not close to the center of the class $\mu_{y_0}^*$, and $\Psi'(\|(X_0 - \mu_{k_0}^*)\hat B\|^2)$ would be small since $X_0 - \mu_{k_0}^*$ is large.

Theorem 3 is proved using Lemma 1, with $F_1 = \mathbb{E}_{z\in H_0} f(z, w)$ and $F_2 = \mathbb{E}_{z\in H_1} f(z, w)$, with L being the tangent space of the constraint set. Here, Lemma 1 is a perturbation result stating that for two comparable functions, their roots are close to each other as well.

Lemma 1. *Assuming that F_1: $\mathbb{R}^n \to \mathbb{R}^n$ is a twice differentiable functions, x_1 is a root of $F_1(x)$, and $\nabla F_1(X_1) \neq 0$. Let $F_2 - F_1$ be another twice differentiable function such that its absolute value and its Lipschitz constant $L_{F_2-F_1}$ are bounded above by $C\epsilon$ and $C'\epsilon$ respectively in a neighborhood of x_1, and $z \in \mathbb{R}^n$ be a value defined by $z = [\nabla F_1(x)]^{-1}((F_2 - F_1)(x_1))$. Then there exists a root of F_2, denoted by x_2, such that $\|x_2 - (x_1 + z)\| \le C''\epsilon$, where C'' is a constant depending on F_1 and x_1.*

The proofs of Lemma 1 is rather technical and deferred to the Appendix.

3.4. Illustrative examples

Ionosphere data: The Johns Hopkins University Ionosphere dataset [27] contains 351 observations on 32 quantitative electromagnetic signal predictors and binary classes: good and bad radars. The error rates of the whole dataset for LDA and the proposed ROSDA with $\zeta = 0.6$ are 0.1054 and 0.0969, respectively. The proposed ROSDA improves classification while it projects the

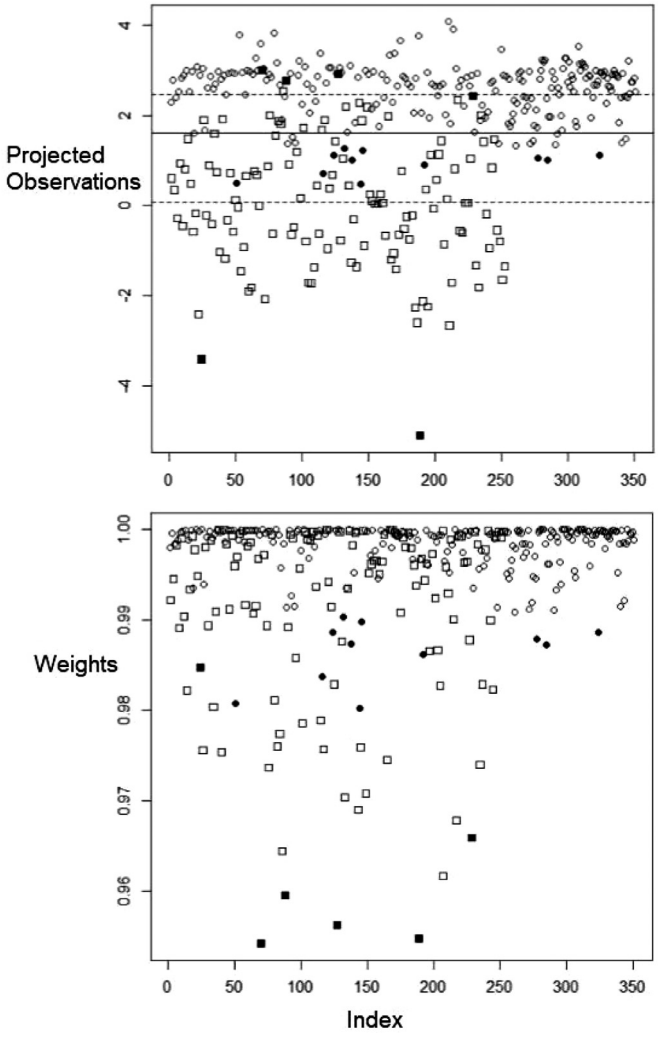

Fig. 1. The top panel is the scatter plot of the projected observations in which circles and squares stand for groups Good and Bad, respectively, the solid circles and squares are the observations with the 5% smallest weights of groups Good and Bad individually, the dashed lines are the group means (0.0690, 2.4708) (groups: Bad, Good), and the solid line is the classification boundary. The bottom panel is the plot of weights corresponding to the scatter plot. The weights of the points on the wrong side of categories are apparently smaller. □: the points of Group Bad; ○: the points of Group Good; ■: the points of Group Bad with the bottom 5% weights; •: the points of Group Bad with the bottom 5% weights.

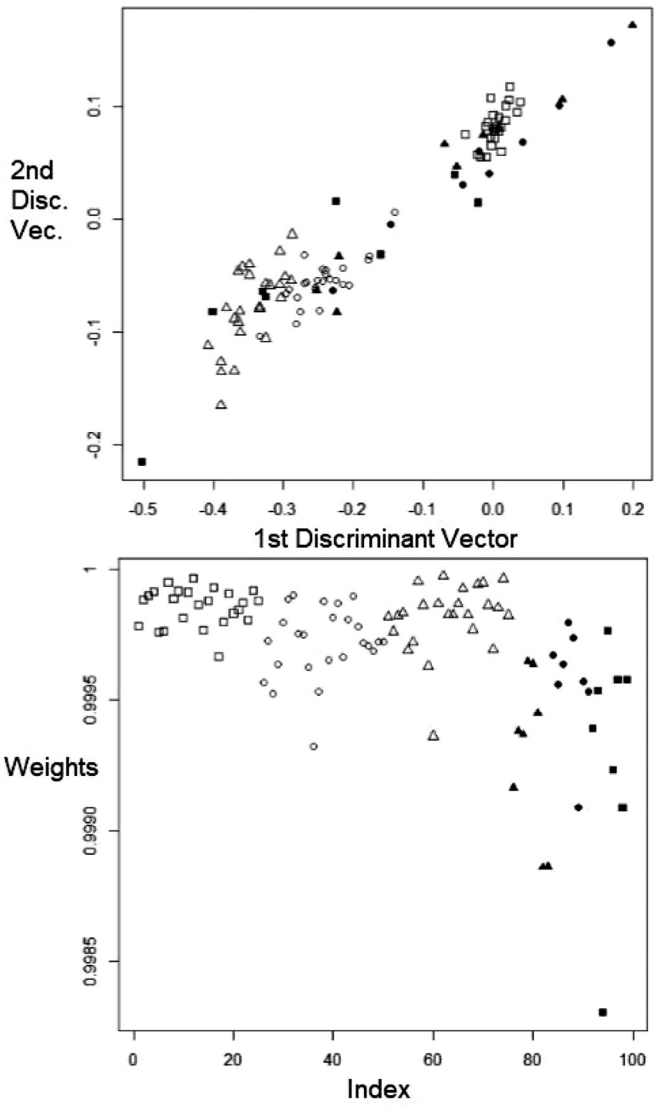

Fig. 2. The top panel is the projected training X on the first two discriminant vectors, and the down panel is the weights of the training observations. Square: Setoda, circle: Versicolor, triangle: Virginica, and the solid points are simulated 24 outliers corresponding to each class. The 24 simulated outliers have smaller weights than the original observations.

data on the discriminant vector. The bottom 5% smallest weights correspond to the observations far from their group centroids or closer to the opposite class, and the observations of Group Good with weights smaller than the 5% quantile of the weights in each group are located in the opposite side of the classification boundary (Fig. 1).

Iris data with simulated outliers The iris data consists of 150 observations, four predictors and three classes: Setosa, Versicolor, and Virginica [10]. We simulated 16 points from a multivariate normal distribution with the mean and covariance as the sample mean and sample covariance of the observations of Setoda, and labeled 8 of them as Versicolor and the rest 8 points as Virginica. Moreover, we simulated 4 points from a multivariate normal distribution with the mean and covariance as the sample mean and sample covariance of the observations of Versicolor, and simulated 4 points from a multivariate normal distribution with the mean and covariance as the sample mean and sample covariance of the observations of Virginica, and then labeled these 8 points as Setosa. Therefore, total 24 simulated points were plugged into the original Iris dataset as outliers.

We first divided the data into two equal portions (75,75) as the training and test sets, and then fitted the models of LDA and ROSDA on the training set, and predict the classes on the test set. The ROSDA with tuning parameter $\zeta = 0.06$. The test error of LDA is 24 misclassified observations, while the ROSDA has only 12 misclassified points by projecting data into the first two discriminant vectors. Notice that the test set does not contain the 24 simulated outliers. The weights of the simulated outliers are generally smaller than the original points (Fig. 2).

4. Experiments for model comparisons

We evaluated the proposed ROSDA with LDA [33], RRLDA, LDAPP, RMDA [2], and Linda [31] with their best tuning parameters using four simulated datasets and seven real-world data applications.

Table 1
The test accuracy rates of each data set using ROSDA, LDA, RRLDA, LDAPP, Linda, and RMDA. Notice that LDAPP can only classify binary classes ($K = 2$), where K is the number of classes. Simulation 1 is an example of Gaussian noises and Simulation 2 is an example of heavy tailed distribution ($t_{d.f=1}$) in low dimensional spaces. Simulation 3 is an example of Gaussian noises in a high dimensional space.

Data	ROSDA	LDA	RRLDA	LDAPP	Linda	RMDA
Sim 1	1.0000	0.8750	1.0000	n.a.	1.0000	0.9875
Sim 2	0.8500	0.7000	0.8625	n.a.	0.8625	0.5750
Sim 3	0.5875	0.5500	0.9500	n.a.	n.a.	n.a.
Iris	0.9714	0.9286	0.9286	n.a.	0.9571	0.7000
Hemo.	0.8000	0.3500	0.7000	0.4000	0.5000	0.9500
Biting	0.8571	0.7857	0.8571	0.8571	0.9286	0.5000
DMD	0.8718	0.8461	0.8461	0.8461	n.a.	0.6154
Anor.	0.5714	0.5000	0.6429	n.a.	0.5714	0.3571
Fish	0.9063	0.9063	0.7185	n.a.	n.a.	n.a.
Heart	0.8333	0.8000	0.8667	0.8333	0.8333	0.7000
Faces	0.9286	0.8214	0.7500	n.a.	n.a.	0.2857

4.1. Simulated data

Simulation 1 consists of classes 1, 2 and 3 in $\mathbb{R}^2$. Each class i has 80 points sampled from identical and independent normal distributions with mean μ_i and variance 1, where $\mu_1 = (0, 2\sqrt{3})$, $\mu_2 = (3, -\sqrt{3})$, and $\mu_3 = (-3, -\sqrt{3})$. We add 20 outlying points into class 1. Theses outliers are sampled from identical and independent normal distributions with mean $(-\sqrt{3}, -3)$ and variance 1. In Simulation 2, we simulate 80 points from the t distribution with degree of freedom 1, $t_{d.f.=1}$ for each of classes 1, 2, and 3 with centers $\mu_1 = (0, 3\sqrt{3})$, $\mu_2 = (6, -2\sqrt{3})$, and $\mu_3 = (-6, -2\sqrt{3})$, and then insert 20 $t_{d.f.=1}$ outliers labeled class 1 of center $(-9, -5\sqrt{3})$. We sampled 80 out of the original 240 data points as the test set, and trained the classification models on the rest data points including the outliers. Simulation 3 is multivariate Gaussian in $\mathbb{R}^{200}$ with the group means as in Simulation 1 in the first two dimensions and zeros in the rest dimensions.

The fourth simulation example is the Iris dataset with simulated outliers. We generated 6 outliers by replacing the 132th data point in variable 1, the 16th measure in variable 2, the 119th measure in variable 3, and the 101th, 110th, and 145th measures in variables 4 with -10. All of these measures were identified as outliers by the χ^2 outliers test [9], and the substitute values are more further from the average of each variable in order to enhance the effect of outliers. The accuracy rates of the test sets are reported in Table 1. The results shows that the outliers strongly influences the classification of LDA and the proposed ROSDA reduced the effect of the outliers.

4.2. Real data applications

The first example is the Hemophilia data [16], which consists of $n_1 = 30$ observations of normal women and $n_2 = 45$ of hemophilia A carriers, with two variables (AHF activity and AHF antigen). We sample 20 out of the original 75 data points as the test set, and train the classification models on the rest data points combined with the generated outliers. The second application is the Biting Flies data [21]. The data set consists of two groups (Leptoconops torrens and Leptoconops carteri) of 70 flies with seven variables. There are five outliers (the 15th, 36th, 51st, 59th, 60th) of variable wing width in second group found by Van Aelst and Willems [32]. We sample 30 out of the 65 data points excluding the five outliers as the test set, and train the classification models on the rest data points. The third is the Duchenne Muscular Dystrophy (DMD) data set [1]. This data set contains measurements of $n_1 = 127$ DMD carriers and $n_2 = 67$ noncarriers. There are ten outliers identified by the χ^2 outliers test [9]. We sample 60 out of the 184 data points excluding the 10 outliers as the test set, and train the classification models on the rest data points. The fourth is the anorexia dataset having 72 observations and two variables and the response of three levels. The fifth is the Fish Catch dataset containing 159 observations with six variables and the response-Species of seven levels. The anorexia and fish datasets were analyzed and used as an example for Linda (Package rrcov in R) [29–31]. The sixth is the heart failure clinical records data [3] with 299 observations and seven numerical covariates and a binary response–death status. We sampled 30 observations as the test set, and used the rest as the training set. The seventh is the Yale Face Database B [24]. We used the first four people (60 images with size 192 by 168 from each person) as four categories and the fifth person as the outliers. We inserted 10 images of the fifth person to each category, then transformed the images with the Daubechies d4 wavelet [8,34], and used the vectorized HL4 coefficients with length $12 \times 10 = 120$ as the input data for each image. We sampled 28 images as the test set and used the rest as the training set.

5. Discussion

The proposed method efficiently reduces the influence of outliers by classifying observation projected in the $Q = K - 1$ discriminant directions with K categories. As the tuning parameter ζ in the Ψ function is zero, ROSDA is equivalent to LDA. Therefore, the classification performance of LDA is the lower bound of ROSDA. Our experiments' results (Table 1) show that only the proposed ROSDA works and provides classification accuracy better than the classical LDA for all the applications including high dimensional ($n \leq p$) data, the number of categories greater than 2, and correlated features. Unlike LDAPP, which applies one-directional projection pursuit so that it can only deal with binary classification. The proposed ROSDA is an extension of LDAPP using multi-directional projection pursuit method, and it is applicable to multiple classes. In the low dimensional data, our method is comparable or even better than RRLDA. In the high dimensional data with sparse signals, our method performs worse than RRLDA, which uses minimum covariance determinant estimator [5] and projection pursuit [7] with the L_1 penalty regularization as the case in Simulation 3 with 2 dimensions of signals and 198 dimensions of noises. However, RRLDA suffers from multiple overlapping classes with low dimensions as the case in the Fish data with seven classes. Linda and RMDA both fail in high dimensional data due to the ill-conditioned inverting Hessian matrix. Nonetheless, we used the regularized target function [4] as a ridge version for B and P_{WX} [17] as $B = (X^\top WX + \lambda I_p)^{-1} X^\top WY\Theta$ and $P_{WX} = WX((WX)^\top WX + \lambda I_p)^{-1}(WX)^\top$, where $\lambda \in [0, \infty)$ can be used. That may be the reason that ROSDA outperforms other methods in the Yale Face example. Additionally, it is the solution of the weighted penalized optimal scoring problem [15] with the regularized loss function $L(\beta_k, \theta_k; \Psi) + \lambda \beta_k^\top \beta_k$ for high dimensional cases. Additionally, the properties in Section 3.3 do not rely on any specific distributions, so that they can be applied to general applications.

Declaration of Competing Interest

The authors have confirmed that there is no conflict of interests.

Acknowledgments

This research was supported by the National Science Foundation (grant number 1924792).

Appendix A

In this section, we will prove Proposition 1, Theorem 2, Theorem 3, and Lemma 1.

A1. Proof of Proposition 1

Without loss of generality, we consider the difference between $L(B^{(2)}, \Theta^{(2)}; \Psi)$ and $L(B^{(1)}, \Theta^{(1)}; \Psi)$.

$$L(B^{(2)}, \Theta^{(2)}; \Psi) - L(B^{(1)}, \Theta^{(1)}; \Psi) = \sum_{i=1}^{n} \Psi(z_{i2}) - \sum_{i=1}^{n} \Psi(z_{i1}) \leq \sum_{i=1}^{n} \Psi'(z_{i1})[z_{i2} - z_{i1}], \quad (7)$$

where $z_{i1} = z(X_i, Y_i, B^{(1)}, \Theta^{(1)})$, $z_{i2} = z(X_i, Y_i, B^{(2)}, \Theta^{(2)})$, and the last inequality holds due to the concavity of Ψ, $\Psi(z)$ is strictly increasing and differentiable with $\Psi'(z) > 0$ for all $z \geq 0$, and the IRLS algorithm solve $(B^{(2)}, \Theta^{(2)})$ by minimizing $\sum_{i=1}^{n} W_i^{(1)} z_{i2}$, where $W_i^{(1)} = \Psi'(z_{i1})$. Therefore, $\sum_{i=1}^{n} W_i^{(1)} z_{i2} \leq \sum_{i=1}^{n} W_i^{(1)} z_{i1}$. Since Ψ is strictly concave, the inequality in Eq. (7) is strict if $z(X_i, Y_i, B^{(2)}, \Theta^{(2)}) \neq z(X_i, Y_i, B^{(1)}, \Theta^{(1)})$ for at least one $i = 1, \ldots, n$. If $z(X_i, Y_i, B^{(2)}, \Theta^{(2)}) \neq z(X_i, Y_i, B^{(1)}, \Theta^{(1)})$ for all i, then the stationary point of (B, Θ) is achieved.

A2. Proof of Theorem 2

Taking the first derivative of $\sum_{i=1}^{n} \Psi(z(X_i, Y_i, B, \Theta))$ with respect to B and let it to zero, and we obtain $\sum_{i=1}^{n} \left[X_i^\top \Psi'(z(X_i, Y_i, B, \Theta))(Y_i\Theta - X_iB)\right] = 0$. In a matrix form, the above equation is $X^\top W(Y\Theta - XB) = 0$, and then $B = (X^\top WX)^{-1}X^\top WY\Theta$. For Θ, please see the argument after Eq. (5).

A3. Proof of Lemma 1

Applying the second-order Taylor series of F_1 at x_1 (assuming that its second directional derivative is bounded above by C_{F_1}) and the Lipschitz assumption of $F_2 - F_1$, we have

$$\begin{aligned} &F_2(x_1 + z) = F_1(x_1 + z) + (F_2(x_1 + z) - F_1(x_1 + z)) \\ \leq &F_1(x_1 + z) + (F_2(x_1) - F_1(x_1)) + L_{F_2-F_1}\|z\| \\ = &F_1(x_1) + z^\top \nabla F_1(x_1) + \frac{1}{2}C_{F_1}\|z\|^2 + (F_2(x_1) - F_1(x_1)) + L_{F_2-F_1}\|z\| \\ = &L_{F_2-F_1}\|z\| + \frac{1}{2}C_{F_1}\|z\|^2. \end{aligned}$$

When $(F_2 - F_1)(x_1)$ and $L_{F_2-F_1}$ are in the order of $O(\epsilon)$, then it implies that $F_2(x_1 + z)$ is in the order of $O(\epsilon^2)$. Since $\nabla F_2(x_1 + z) = \nabla F_2(x_1) + O(\epsilon) = \nabla F_1(x_1) + O(\epsilon)$, it implies that $\nabla F_2(x_1 + z)$ is nonzero and invertible when ϵ is small. The lemma is then proved by applying the Brouwer fixed-point theorem to $g(x) = -[\nabla F_2(x_1 + z)]^{-1}F_2(x) + x$ to the region $B(x_1 + z, C''\epsilon^2)$, a ball centered at $x_1 + z$ with radius $C''\epsilon^2$.

References

[1] D.F. Andrews, A.M. Herzberg, Data, Springer-Verlag, New York, 1985.
[2] C. Bouveyron, S. Girard, Robust supervised classification with mixture models: learning from data with uncertain labels, Pattern Recognit. 42 (11) (2009) 2649–2658.
[3] D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 16.
[4] L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (4) (2011) 406–413.
[5] C. Croux, G. Haesbroeck, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, J. Multivariate Anal. 71 (1999) 161–190.
[6] C. Croux, C. Dehon, Robust linear discriminant analysis using s-estimators, Can. J. Stat. 29 (3) (2001) 473–493.
[7] C. Croux, P. Filzmoser, M. Oliveira, Algorithms for projection pursuit robust principal component analysis, Chemom. Intell. Lab. Syst. 87 (2) (2007) 218–225.
[8] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.
[9] W.J. Dixon, Analysis of extreme values, Ann. Math. Stat. 21 (4) (1950) 488–506.
[10] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[11] M. Gschwandtner, P. Filzmoser, C. Croux, G. Haesbroeck, rrlda: Robust regularized linear discriminant analysis, R package version 1.1 (2012).
[12] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, Biostatistics 8 (1) (2007) 86–100.
[13] D.J. Hand, Classifier technology and the illusion of progress, Stat. Sci. 21 (1) (2006) 1–15.
[14] T. Hastie, R. Tibshirani, A. Buja, Flexible discriminant analysis by optimal scoring, J. Amer. Statist. Assoc. 89 (428) (1994) 1255–1270.
[15] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1) (1995) 73–102.
[16] J.D.F. Habbema, J. Hermans, K. Van den Broeck, A stepwise discriminant analysis program using density estimation, in: Proceedings in computational statistics, Physica-Verlag, Vienna, 1974, pp. 101–110.
[17] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[18] I. Higuchi, S. Eguchi, Robust principal component analysis with adaptive selection for tuning parameters, J. Mach. Learn. Res. 5 (2004) 453–471.
[19] S.Y. Huang, Y.R. Yeh, S. Eguchi, Robust kernel principal component analysis, J. Neural Comput. 21 (11) (2009) 3179–3213.
[20] H.H. Huang, Y.R. Yeh, Iterative algorithm for robust kernel principal component analysis, Neurocomputing 74 (18) (2011) 3921–3930.
[21] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Upper Saddle River, 2002.
[22] S.J. Kim, A. Magnani, S. Boyd, Robust fisher discriminant analysis, Adv. Neural Inf. Process. Syst 18 (2005) 659–666.
[23] C.-N. Li, Y.-H. Shao, N.Y. Deng, Robust l1-norm two-dimensional linear discriminant analysis, Neural Netw. 65 (2015) 92–104.
[24] K.-C. Lee, J. Ho, D.J. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[25] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, Y. Li, A non-greedy algorithm for L1-norm LDA, IEEE Trans. Image Process. 26 (2) (2016) 684–695.
[26] A.M. Pires, J.A. Branco, Projection-pursuit approach to robust linear discriminant analysis, J. Multivar. Anal. 101 (2010) 2464–2485.
[27] V.G. Sigillito, S.P. Wing, L.V. Hutton, K.B. Baker, Classification of radar returns from the ionosphere using neural networks, Johns Hopkins APL Tech. Dig. 10 (1989) 262–266.
[28] E. Slutsky, Über stochastische asymptoten und grenzwerte, Metron (in German) 5 (3) (1925) 3–89.
[29] V. Todorov, Robust selection of variables in linear discriminant analysis, Stat. Methods Appl. 15 (2007) 395–407.
[30] V. Todorov, A.M. Pires, Comparative performance of several robust linear discriminant analysis methods, Revstat Stat. J. 5 (2007) 63–83.
[31] v. Todorov, P. Filzmoser, An object oriented framework for robust multivariate analysis, J. Stat. Softw. 32 (3) (2009) 1–47.
[32] S. Van Aelst, G. Willems, Inference for robust canonical variate analysis, Adv. Data Anal. Classif. 4 (2010) 181–197.
[33] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, New York, 2002.
[34] B. Whitcher, waveslim: Basic wavelet routines for one-, two-, and three-dimensional signal processing, R package version 1.8.2 (2020).
[35] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, T. Yin, L1-norm distance linear discriminant analysis based on an effective iterative algorithm, IEEE Trans. Circuits Syst. Video Technol. 28 (1) (2016) 114–129.
[36] F. Zhong, J. Zhang, Linear discriminant analysis based on l1-norm maximization, IEEE Trans. Image Process. 22 (8) (2013) 3018–3027.