ROBUST LINEAR DISCRIMINANT ANALYSIS WITH A LAPLACIAN ASSUMPTION ON PROJECTION DISTRIBUTION

Shujian Yu[1], *Zheng Cao*[1], *Xiubao Jiang*[2]

[1]Department of Electrical and Computer Engineering, University of Florida
[2]2012 Lab, Huawei Technologies Co., LTD

ABSTRACT

Linear discriminant analysis (LDA) is typically carried out using Fisher's method, which relies heavily on the estimation of sample mean vectors and covariance matrices. However, Fisher LDA is vulnerable to outliers as it happens to other multivariate statistical methods. In this paper, we analyzed the optimal discriminant design based on the criterion of minimizing total misclassification rate, assuming that the projected samples follow Laplacian distribution. The corresponding optimization objective can be approximated as a linear programming problem. We illustrated the relations of our proposed discriminant to Fisher LDA and minimax probability machine (MPM) from the perspective of projection-pursuit. Experiments on 6 real world benchmark dataset from UCI repository validate the effectiveness of our method.

Keywords—Linear discriminant analysis (LDA), Laplacian distribution, Minimax probability machine (MPM).

I. INTRODUCTION

Linear discriminant analysis (LDA) methods seek to find a projection vector (or discriminant) that yields optimal discrimination between distinct groups (or classes) of observations [1]. Among them, the most popular one is Fisher LDA [2]. The basic idea of Fisher LDA is to project all the samples into a lower dimensional space that maximizes the between-class separability while minimizing their within-class variability. However, Fisher's method is vulnerable to outliers as it happens to other multivariate statistical methods [3].

To handle this, vast of efforts have been made on robust LDA in the last decades, mostly for binary classification. Early works focused on replacing sample mean vectors and pooled sample covariance matrix with their robust counterparts (normally called "plug-in method") [4], [5]. Recent works paid more attentions on projection-pursuit (PP) approach, which is initiated in [6] and further developed in [7], [8]. In general, PP techniques search for low-dimensional projections of higher-dimensional data where an objective function called *projection index* (PI) is maximized [7]. Other relevant works for robust LDA attempt to optimize misclassification rate under worst-case scenario. [9], [10] proposed minimax probability machine (MPM) to find the discriminant that can maximize the probability of correct classification in the worst-case setting. On the other hand, by explicitly incorporating a model of data uncertainty in a classification problem, [11] also developed a robust Fisher LDA model which can be carried out using convex optimization.

In this paper, we firstly analyzed the optimal discriminant design based the criterion of minimizing total misclassification rate, assuming that the projected samples follow Laplacian distribution[1]. After that, we presented a novel robust LDA method and also illustrated its connections to Fisher LDA and MPM from the PP perspective.

The rest of this paper is organized as follows. In section II, we briefly reviewed related works on LDA and its robustification. In section III, we proposed a novel robust LDA method and discussed the corresponding optimization problem. Following that, we illustrated the connections between our method with Fisher LDA and MPM. Section IV presented experimental comparison between our method, standard LDA and MPM, using 6 real world benchmark dataset from UCI repository. This paper is concluded in section V.

II. BACKGROUNDS

Before discussing our method, we gave a short review on some related backgrounds. Throughout this paper, we only consider samples coming from two pre-specified classes (or populations).

II-A. Fisher LDA for binary classification

Fisher LDA finds a linear discriminant that yields optimal discrimination between two classes [1]. Suppose a dataset $\mathcal{X}$ is given for which each sample $\mathbf{x} \in \mathbb{R}^D$ either $\mathbf{x} \in \mathcal{C}_1$ or $\mathbf{x} \in \mathcal{C}_2$, where $\mathcal{C}_i (i = 1, 2)$ denotes a class. Then for a linear discriminant characterized by $\boldsymbol{w} \in \mathbb{R}^n$, the degree of discrimination is measured by the Fisher discriminant ratio:

$$\begin{aligned} \mathcal{J}(\boldsymbol{w}) &= \frac{\boldsymbol{w}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{w}} \\ &= \frac{(\boldsymbol{w}^T\boldsymbol{\mu}_2 - \boldsymbol{w}^T\boldsymbol{\mu}_1)^2}{\boldsymbol{w}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{w}} \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ ($\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$) denote the mean and covariance matrix of classes $\mathcal{C}_1$ and $\mathcal{C}_2$.

A discriminant that maximizes (1) is thus given by:

$$\boldsymbol{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \tag{2}$$

Having found a discriminant $\boldsymbol{w}$, the linear discriminant procedure for any new $\mathbf{x}$ is as follows:
a) classify $\mathbf{x}$ into $\mathcal{C}_1$ if $\boldsymbol{w}^T\mathbf{x} \geq \gamma$

[1]The reasons behind the choice of Laplacian is elaborated in III-A

b) classify $\mathbf{x}$ into $\mathcal{C}_2$ if $\boldsymbol{w}^T\mathbf{x} < \gamma$
where γ is an estimated cutoff point (for Fisher LDA, it is the midpoint of two means). In practice, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unknown and estimated from given samples. Thus Fisher LDA is highly vulnerable to problem data or small data size.

II-B. Robust LDA with Projection-pursuit (PP) approaches

The pioneering works on multivariate statistics using PP approaches was initiated in [7], [8]. The most promising advantage of PP approaches lies in their capability to overcome the so called "curse of dimensionality" [12]. In general, PP techniques search for low-dimensional projections of higher-dimensional data where the *projection index* (PI) is maximized [3], [7].

Recalling the problem of robust LDA, one representative PI was proposed in [13]:

$$\mathcal{I}_1(\boldsymbol{w}) = \frac{|L(\boldsymbol{w}^T\mathbf{X}_1) - L(\boldsymbol{w}^T\mathbf{X}_2)|}{S(\boldsymbol{w}^T\mathbf{X}_1, \boldsymbol{w}^T\mathbf{X}_2)} \tag{3}$$

where $L(\cdot)$ denotes location estimator, $S(\cdot)$ is a dispersion (or scale) estimator, and $\mathbf{X}_1 = \{\mathbf{x}_n, n \in \mathcal{C}_1\}$ ($\mathbf{X}_2 = \{\mathbf{x}_n, n \in \mathcal{C}_2\}$) contains all the training samples in class $\mathcal{C}_1$ ($\mathcal{C}_2$).

Another well-known PI is the squared standardized distance between the projected observations of the two classes ($i.e.$, in the Fisher sense):

$$\mathcal{I}_2(\boldsymbol{w}) = \frac{(L(\boldsymbol{w}^T\mathbf{X}_1) - L(\boldsymbol{w}^T\mathbf{X}_2))^2}{\theta S^2(\boldsymbol{w}^T\mathbf{X}_1) + (1-\theta)S^2(\boldsymbol{w}^T\mathbf{X}_2)} \tag{4}$$

where $L(\cdot)$ and $S(\cdot)$ have the same meaning as in (3), θ is the prior probability of one class.

In fact, $L(\cdot)$ and $S(\cdot)$ can have different pairs of choices. It is not difficult to find that when $L(\cdot)$ and $S(\cdot)$ are sample mean and sample standard deviation, Fisher's solution is obtained in (4). Review works on performance comparison of different pairs of $L(\cdot)$ and $S(\cdot)$ are available in [14], [13].

II-C. Minimax probability machine (MPM)

MPM [9], [10] is the workhorse of robust LDA considering worst-case scenarios. Under MPM framework, for all possible choices of class-conditional densities with given mean and covariance matrix, the worst-case (maximum) misclassification rate of future data is minimized. This can be formulated as:

$$\begin{aligned} \max_{\alpha, \boldsymbol{w}\neq 0, \gamma} \quad & \alpha \\ s.t. \quad & \inf_{\mathbf{x}\sim(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)} P(\boldsymbol{w}^T\mathbf{x} \geq \gamma) \geq \alpha \\ & \inf_{\mathbf{x}\sim(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} P(\boldsymbol{w}^T\mathbf{x} \leq \gamma) \geq \alpha \end{aligned} \tag{5}$$

Following the results in [15], it can be proved that the optimization problem (5) is equivalent to:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}} + \sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_2\boldsymbol{w}} \\ s.t. \quad & \boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 1 \end{aligned} \tag{6}$$

The optimization problem (6) is a convex optimization problem, more precisely a second order cone programm [16].

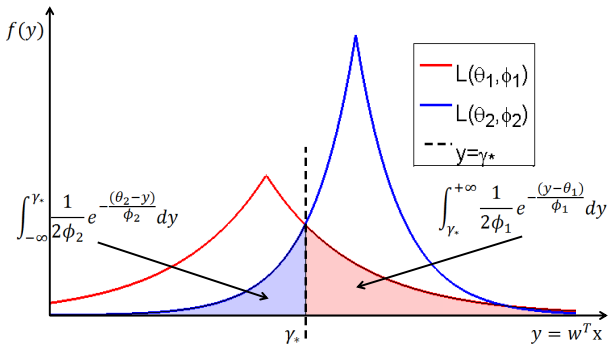


Fig. 1: An analogy for projected distribution of two classes.

III. ROBUST LINEAR DISCRIMINANT ANALYSIS

In this section, we elaborated the reasons behind the choice of Laplacian distribution. After that, a novel robust LDA using Laplacian assumption is presented.

III-A. Why Laplacian

The main reasons for the selection of Laplacian distribution are threefold: 1) Regarding distribution of linear projection $\mathbb{R}^D \mapsto \mathbb{R}^d$ ($d < D$), [17] pointed out the projected distribution is close to a single Gaussian under suitable conditions when $d = 1$. This conclusion was later questioned by [18], which mathematically validated the fact that almost all linear projections look like a scale-mixture of spherical Gaussians (this reduces to scale-mixture of Gaussians for $d = 1$), if the coefficient of eccentricity is small; 2) It is not difficult to prove that the standard exponential power of family is a subset of the classes of scale-mixture of Gaussians (see [19], [20]); and 3) Within the exponential power of family, Laplacian distribution has the simplest formulation and the definite integral of Laplacian distribution has explicit expression on its parameters [21], which makes it computationally tractable.

III-B. Optimal LDA for Laplacian projected data

Suppose projected samples $y = \boldsymbol{w}^T\mathbf{x}$ that come from two Laplacian distributions are characterized by parameters (θ_1, ϕ_1) and (θ_2, ϕ_2) (see Fig.1):

$$f_1(y|\theta_1, \phi_1) = \frac{1}{2\phi_1}\exp\left(-\frac{|y-\theta_1|}{\phi_1}\right) \tag{7}$$

$$f_2(y|\theta_2, \phi_2) = \frac{1}{2\phi_2}\exp\left(-\frac{|y-\theta_2|}{\phi_2}\right) \tag{8}$$

where $\theta \in (-\infty, +\infty)$ and $\phi > 0$ are location and scale parameters, respectively. Then the optimal cutoff point γ_* satisfies:

$$f_1(\gamma_*|\theta_1, \phi_1) = f_2(\gamma_*|\theta_2, \phi_2) \tag{9}$$

Assuming $\theta_2 > \theta_1$, with straightforward simplification, γ_* is given by:

$$\gamma_* = \frac{\phi_1\phi_2\ln(\phi_2/\phi_1) + \phi_2\theta_1 + \phi_1\theta_2}{\phi_1 + \phi_2}. \tag{10}$$

The misclassification probability $\mathbf{P}$ can be represented as:

$$\begin{aligned}\mathbf{P} &= \frac{1}{2}\left(\int_{\gamma_*}^{+\infty} f_1(y)dy + \int_{-\infty}^{\gamma_*} f_2(y)dy\right) \\ &= \frac{1}{2}\left[\frac{1}{2}\exp\left(\frac{\theta_1-\gamma_*}{\phi_1}\right) + \frac{1}{2}\exp\left(\frac{\gamma_*-\theta_2}{\phi_2}\right)\right]. \quad (11)\end{aligned}$$

Combining (10) and (11), we have:

$$\mathbf{P} = \frac{1}{4}\left[\left((\phi_1/\phi_2)^{\frac{\phi_2}{\phi_1+\phi_2}} + (\phi_2/\phi_1)^{\frac{\phi_1}{\phi_1+\phi_2}}\right)\exp\left(\frac{\theta_1-\theta_2}{\phi_1+\phi_2}\right)\right]. \quad (12)$$

Since the maximum likelihood (ML) estimator $\hat{\theta}_k(k=1,2)$ to θ_k is sample median (Med) and the ML estimator $\hat{\phi}_k(k=1,2)$ to ϕ_k is mean absolute deviation (MAD) [22]:

$$\hat{\phi}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}^{N} |\boldsymbol{w}^T\mathbf{x}_n - \hat{\theta}_k|, k=1,2. \quad (13)$$

The optimal discriminant criterion can be formulated directly:

$$\begin{aligned}\min_{\boldsymbol{w}} \quad & \left((\hat{\phi}_1/\hat{\phi}_2)^{\frac{\hat{\phi}_2}{\hat{\phi}_1+\hat{\phi}_2}} + (\hat{\phi}_2/\hat{\phi}_1)^{\frac{\hat{\phi}_1}{\hat{\phi}_1+\hat{\phi}_2}}\right)\exp\left(\frac{\hat{\theta}_1-\hat{\theta}_2}{\hat{\phi}_1+\hat{\phi}_2}\right) \quad (14)\\ s.t. \quad & \hat{\theta}_k = Med(\boldsymbol{w}^T\mathbf{x}_n, n\in\mathcal{C}_k), k=1,2. \\ & \hat{\phi}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}^{N} |\boldsymbol{w}^T\mathbf{x}_n - \hat{\theta}_k|, k=1,2.\end{aligned}$$

As can be seen, the estimation of sample mean and sample standard deviation in Fisher LDA is substituted with sample Med and MAD in (14). Compared with sample mean and standard deviation, the sample Med and MAD are relatively less sensitive to tail behavior of error distributions or outliers [23], [24], which immediately brings the benefits of robustness. Unfortunately, the objective (14) is difficult to solve. In the next section, we show that (14) can be approximated with a linear programming problem.

III-C. Approximation and solution to robust LDA

Regarding the term $(\phi_1/\phi_2)^{\frac{\phi_2}{\phi_1+\phi_2}} + (\phi_2/\phi_1)^{\frac{\phi_1}{\phi_1+\phi_2}}$ in (12), it is easy to prove and visually demonstrate that

$$1 < (\phi_1/\phi_2)^{\frac{\phi_2}{\phi_1+\phi_2}} + (\phi_2/\phi_1)^{\frac{\phi_1}{\phi_1+\phi_2}} \leq 2 \quad (15)$$

As an approximation, it is reasonable to assume that the term $\exp\left(\frac{\theta_1-\theta_2}{\phi_1+\phi_2}\right)$ dominates misclassification probability $\mathbf{P}$. (14) is thus reduced to:

$$\begin{aligned}\max_{\boldsymbol{w}} \quad & \frac{\left|\hat{\theta}_1 - \hat{\theta}_2\right|}{\hat{\phi}_1+\hat{\phi}_2} \quad (16)\\ s.t. \quad & \hat{\theta}_k = Med(\boldsymbol{w}^T\mathbf{x}_n, n\in\mathcal{C}_k), k=1,2. \\ & \hat{\phi}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}^{N_k} |\boldsymbol{w}^T\mathbf{x}_n - \hat{\theta}_k|, k=1,2.\end{aligned}$$

However, the objective (16) with respect to $\boldsymbol{w}$ is not differentiable, which makes it still difficult to be solved. To tackle this, we further approximate $\hat{\theta}_k$ with $\boldsymbol{w}^T Med(\mathbf{x}_n, n\in\mathcal{C}_k) = \boldsymbol{w}^T\mathbf{m}_k(k=1,2)$, where the i-th element of $\mathbf{m}_k$ is the median of $\{\mathbf{x}_{ni}, n\in\mathcal{C}_k\}$, and $\mathbf{x}_{ni}$ is the i-th element of $\mathbf{x}_n$. Finally the optimization problem (16) goes down to:

$$\max_{\boldsymbol{w}} \frac{\left|\boldsymbol{w}^T\mathbf{m}_1 - \boldsymbol{w}^T\mathbf{m}_2\right|}{\frac{1}{N_1}\sum_{n\in\mathcal{C}_1}^{N_1}|\boldsymbol{w}^T\mathbf{x}_n - \boldsymbol{w}^T\mathbf{m}_1| + \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}^{N_2}|\boldsymbol{w}^T\mathbf{x}_n - \boldsymbol{w}^T\mathbf{m}_2|}$$

which is equivalent to:

$$\begin{aligned}\min_{\boldsymbol{w}} \quad & \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}^{N_1}|\boldsymbol{w}^T\mathbf{x}_n - \boldsymbol{w}^T\mathbf{m}_1| + \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}^{N_2}|\boldsymbol{w}^T\mathbf{x}_n - \boldsymbol{w}^T\mathbf{m}_2| \quad (17)\\ s.t. \quad & \boldsymbol{w}^T(\mathbf{m}_1 - \mathbf{m}_2) = 1\end{aligned}$$

Obviously, (17) can be formulated as a linear programming problem [16]:

$$\begin{aligned}\min_{\boldsymbol{w}} \quad & \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}^{N_1} s_n + \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}^{N_1} s_n \quad (18)\\ s.t. \quad & \boldsymbol{w}^T(\mathbf{m}_1 - \mathbf{m}_2) = 1 \\ & \boldsymbol{w}^T(\mathbf{x}_n - \mathbf{m}_1) \leq s_n, n\in\mathcal{C}_1 \\ & -\boldsymbol{w}^T(\mathbf{x}_n - \mathbf{m}_1) \leq s_n, n\in\mathcal{C}_1 \\ & \boldsymbol{w}^T(\mathbf{x}_n - \mathbf{m}_2) \leq s_n, n\in\mathcal{C}_2 \\ & -\boldsymbol{w}^T(\mathbf{x}_n - \mathbf{m}_2) \leq s_n, n\in\mathcal{C}_2\end{aligned}$$

III-D. Relations to Fisher LDA and MPM

Fisher LDA involves solving the following problem:

$$\max_{\boldsymbol{w}} \mathcal{J}_{Fisher}(\boldsymbol{w}) = \frac{|\boldsymbol{w}^T\boldsymbol{\mu}_2 - \boldsymbol{w}^T\boldsymbol{\mu}_1|}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{\Sigma}_2\boldsymbol{w}}} \quad (19)$$

On the other hand, the discriminants of MPM and our proposed robust LDA are obtained by solving (20) and (21):

$$\max_{\boldsymbol{w}} \mathcal{J}_{MPM}(\boldsymbol{w}) = \frac{|\boldsymbol{w}^T\boldsymbol{\mu}_2 - \boldsymbol{w}^T\boldsymbol{\mu}_1|}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}} + \sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_2\boldsymbol{w}}} \quad (20)$$

$$\max_{\boldsymbol{w}} \mathcal{J}_{ours}(\boldsymbol{w}) = \frac{|\boldsymbol{w}^T Med(\mathbf{X}_2) - \boldsymbol{w}^T Med(\mathbf{X}_1)|}{MAD(\boldsymbol{w}^T\mathbf{X}_1) + MAD(\boldsymbol{w}^T\mathbf{X}_2)} \quad (21)$$

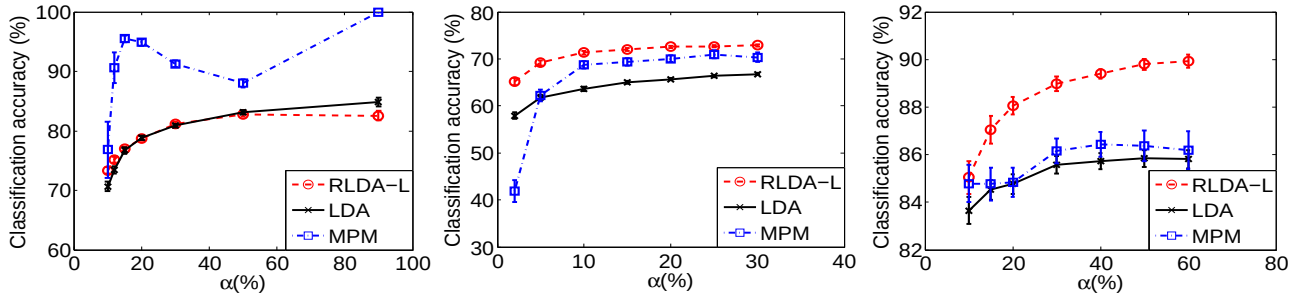
As can be seen, both MPM and our method share a common PI of the following form:

$$\mathcal{I}_3(\boldsymbol{w}) = \frac{|L(\boldsymbol{w}^T\mathbf{X}_1) - L(\boldsymbol{w}^T\mathbf{X}_2)|}{S(\boldsymbol{w}^T\mathbf{X}_1) + S(\boldsymbol{w}^T\mathbf{X}_2)} \quad (22)$$

where $L(\cdot)$ and $S(\cdot)$ denote location estimator and dispersion estimator respectively, as in (3) and (4). On the other hand, the PI of Fisher LDA can be represented as:

$$\mathcal{I}_4(\boldsymbol{w}) = \frac{|L(\boldsymbol{w}^T\mathbf{X}_1) - L(\boldsymbol{w}^T\mathbf{X}_2)|}{\sqrt{S^2(\boldsymbol{w}^T\mathbf{X}_1) + S^2(\boldsymbol{w}^T\mathbf{X}_2)}} \quad (23)$$

Generally, Fisher LDA, MPM and our discriminant all involve seeking an optimal $\boldsymbol{w}_*$ with which we can have a good separation between the two projected sets $\boldsymbol{w}^T\mathbf{X}_1$ and $\boldsymbol{w}^T\mathbf{X}_2$



(a) *Ionosphere* (b) *Pima* (c) *Wholesale*

Fig. 2: Classification accuracy for *Ionosphere*, *Pima* and *Wholesale* benchmarks versus size of the training set. The dashed line represents our RLDA-L results, the solid line is the LDA results and the dash-dot line denotes MPM results. The vertical bars represent the standard errors.

with small dispersions, yet the selected local and dispersion estimators are different for our method and MPM, and the way in which the sample mean and sample variance are combined is also different for MPM and Fisher LDA.

IV. EXPERIMENTS

In this section, we evaluate and compare the performance of our proposed robust LDA with a Laplacian assumption (RLDA-L, for short) with two benchmark methods, *i.e.*, LDA and MPM. To this end, 6 benchmark binary classification dataset from UCI repository, i.e., *Ionosphere*, *Magic*, *Pima*, *Skin*, *Vertebral* and *Wholesale*, are selected. Each set is randomly partitioned into 90% training and 10% testing as conducted in [9], [10], [25] (except for *Magic* and *Skin*, where only 100 samples from each class are used for training). Ensemble average results over 50 random partitions are reported in Table I. It is well known that a larger training set typically provides better testing performance. Denote α the size of training set, as a fraction of the total number of samples (for example, $\alpha = 0.3$ means that each dataset is randomly partitioned into 30% training and 70% testing). To test the robustness (or coherence), we repeat the same experimental procedure for each set, except that only 10% data is used for training. The corresponding results are listed in Table II. As can be seen, the performance of RLDA-L and MPM is consistently better than LDA. This phenomena confirmed the widely existence of non-Gaussian projections in real world as well as the feasibility of our projection assumption. Furthermore, it is interesting to find that the MPM can achieve extremely good results (even 100%) given enough training samples. However, The advantages of MPM does not exist as the training set become smaller.

To further verify this, we change the value of α within a reasonable range as conducted in [11], and test the performance of all the methods on *Ionosphere*, *Pima* and *Wholesale*. Fig.2 summarizes the classification results. For each of the tested methods, and for each value of α, we plot the average classification accuracy as well as the standard errors of mean. As can be expected, apart from sensitivity to small training data size, another drawback for MPM lies in its large variance. Compared with MPM, the performance of RLDA-L is much more stable.

Table I: Classification accuracy on UCI real data ($\alpha = 90\%$).

	RLDA-L	LDA	MPM
Ionosphere (%)	82.17 (5.90)	85.77 (4.91)	**100** (0)
Magic (%)	75.06 (2.10)	77.20 (0.96)	**78.76** (2.93)
Pima (%)	72.08 (4.27)	68.70 (5.18)	**100** (0)
Skin (%)	**94.19** (0.26)	93.60 (0.55)	91.99 (4.09)
Vertebral (%)	73.48 (8.48)	72.13 (8.39)	**100** (0)
Wholesale (%)	89.64 (3.71)	85.23 (5.22)	**100** (0)

Table II: Classification accuracy on UCI real data ($\alpha = 10\%$).

	RLDA-L	LDA	MPM
Ionosphere (%)	73.10 (4.39)	69.11 (5.38)	**83.33** (25.99)
Magic (%)	75.06 (2.10)	77.20 (0.96)	**78.76** (2.93)
Pima (%)	**70.61** (3.28)	63.68 (2.97)	69.10 (5.57)
Skin (%)	**94.19** (0.26)	93.60 (0.55)	91.99 (4.09)
Vertebral (%)	68.22 (5.41)	67.14 (4.07)	**70.10** (7.70)
Wholesale (%)	**86.12** (5.38)	84.45 (4.08)	85.20 (4.82)

V. CONCLUSIONS AND FUTURE WORKS

We present a novel robust LDA method assuming Laplacian distributions for projected samples. Under this assumption, the optimal discriminant is investigated. In addition, we show that the proposed method can be carried out using linear programming, making implementation very convenient. Experiments validate the effectiveness of our method compared with other benchmarks.

Besides, the proposed method has several interesting implications and extensions: a) unlike majority of previous works using prior knowledge on class-conditional distribution (e.g. [26]), we demonstrate the feasibility of making assumptions on projection distribution; b) it is particularly interesting to find that our method share the same projection index with MPM, given that the two methods are derived from different theoretical perspectives; c) the proposed method can be easily extended for feature selection with ℓ_1 penalty[2].

[2]This can be done by adding ℓ_1 penalty on $\boldsymbol{w}$ in (18). Again, the new objective can be formulated as a linear programming problem. According to our experimental results (not shown in the paper), the performance of feature selection is satisfactory.

VI. REFERENCES

[1] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 4, no. 4.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[3] A. M. Pires and J. A. Branco, "Projection-pursuit approach to robust linear discriminant analysis," *Journal of Multivariate Analysis*, vol. 101, no. 10, pp. 2464–2485, 2010.

[4] R. H. Randles, J. D. Broffitt, J. S. Ramberg, and R. V. Hogg, "Generalized linear and quadratic discriminant functions using robust estimates," *Journal of the American Statistical Association*, vol. 73, no. 363, pp. 564–568, 1978.

[5] D. M. Hawkins and G. J. McLachlan, "High-breakdown linear discriminant analysis," *Journal of the American Statistical Association*, vol. 92, no. 437, pp. 136–143, 1997.

[6] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," 1974.

[7] P. J. Huber, "Projection pursuit," *The Annals of Statistics*, pp. 435–475, 1985.

[8] G. Li and Z. Chen, "Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 759–766, 1985.

[9] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "Minimax probability machine," in *Advances in Neural Information Processing Systems*, 2001, pp. 801–807.

[10] ——, "A robust minimax approach to classification," *The Journal of Machine Learning Research*, vol. 3, pp. 555–582, 2003.

[11] S.-J. Kim, A. Magnani, and S. Boyd, "Robust fisher discriminant analysis," in *Advances in Neural Information Processing Systems*, 2005, pp. 659–666.

[12] A. Pires, "Robust linear discriminant analysis and the projection pursuit approach," in *Developments in Robust Statistics*. Springer, 2003, pp. 317–329.

[13] Z.-Y. Chen and R. J. Muirhead, "A comparison of robust linear discriminant procedures using projection pursuit methods," *Lecture Notes-Monograph Series*, pp. 163–176, 1994.

[14] K. Joossens and C. Croux, "Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis," in *Theory and applications of recent robust methods*. Springer, 2004, pp. 131–140.

[15] A. W. Marshall and I. Olkin, "Multivariate chebyshev inequalities," *The Annals of Mathematical Statistics*, pp. 1001–1014, 1960.

[16] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[17] P. Diaconis and D. Freedman, "Asymptotics of graphical projection pursuit," *The Annals of Statistics*, pp. 793–815, 1984.

[18] S. Dasgupta, D. Hsu, and N. Verma, "A concentration theorem for projections," in *Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[19] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate laplace distribution," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 300–303, 2006.

[20] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 3, pp. 646–648, 1987.

[21] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.

[22] A. Childs and N. Balakrishnan, "Maximum likelihood estimation of laplace parameters based on general type-ii censored examples," *Statistical Papers*, vol. 38, no. 3, pp. 343–349, 1997.

[23] M. Svensén and C. M. Bishop, "Robust bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[24] P. J. Huber, *Robust statistics*. Springer, 2011.

[25] K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan, "The minimum error minimax probability machine," *The Journal of Machine Learning Research*, vol. 5, pp. 1253–1286, 2004.

[26] T. W. Anderson and R. Bahadur, "Classification into two multivariate normal distributions with different covariance matrices," *Annals of Mathematical Statistics*, vol. 33, no. 2, pp. 420–431, 1962.