International Journal of Innovative Computing, Information and Control
Volume **15**, Number **3**, June **2019**

ICIC International ©2019 ISSN 1349-4198
pp. **955–967**

CLASSIFICATION OF HIGH-DIMENSIONAL DATA: A RANDOM-MATRIX REGULARIZED DISCRIMINANT ANALYSIS APPROACH

Bin Ye and Peng Liu

School of Information and Control Engineering
China University of Mining and Technology
No. 1, Daxue Road, Quanshan District, Xuzhou 221116, P. R. China
yebin@cumt.edu.cn

Received September 2018; revised January 2019

Abstract. *Linear discriminant analysis (LDA) is one of the most popular parametric classification methods in machine learning and data mining tasks. Although it performs well in many applications, LDA is impractical for high-dimensional data sets which are now routinely generated everywhere in modern society. A primary reason for the inefficiency of LDA for high-dimensional data is that the sample covariance matrix is no longer a good estimator of the actual covariance matrix when the dimension of feature vector p is close to or even larger than the sample size n. Here, we propose to regularize LDA classifier by employing a truly consistent estimator of the high-dimensional covariance matrix. Using the theoretical tools from random matrix theory, the covariance matrices in high-dimensions are estimated in a linear or nonlinear shrinkage manner depending on the comparison between the dimension p and the sample size n. Besides improving the flexibility, numerical simulations demonstrate that the regularized discriminant analysis using random matrix theory yields higher accuracy than the existing competitors for a wide variety of synthetic and real data sets.*

Keywords: Linear discriminant analysis, High-dimensional data, Random matrix theory, Classification, Covariance matrix

1. **Introduction.** Linear discriminant analysis is a well-established supervised learning technique applicable in a variety of areas [1, 2, 3]. As a model-based classifier, it aims to allocate a data point into one of the predefined classes on the basis of a number of feature variables. Compared with other classification algorithms such as random forests or support vector classifier, the model constructed in LDA is more interpretable and easy to make predictions.

In the present era of "Big Data", with the rapid development of information technology, high-dimensional data sets are now generated and collected in almost all fields – engineering, physics, education, e-commerce, genomics [4, 5, 6, 7], etc. The most direct manifestation of high-dimensional data is that its dimension p is not fixed but becomes large together with the sample size n, which is called "large n, large p" asymptotics. Thus, high-dimensional data will transcend the boundary of classical multivariate statistics where we implicitly assume that the dimension of feature vector p is fixed while the sample size n tends to infinity. High-dimensional data bring great challenges to statistical learning techniques, including LDA. Linear discriminant classifier becomes inefficient in high-dimensional settings. One important reason is that the sample covariance matrix $\boldsymbol{S}$ in high dimensions is singular (noninvertible) or very close to being singular. It

DOI: 10.24507/ijicic.15.03.955

is no longer a good approximation to the population covariance matrix $\boldsymbol{\Sigma}$ in the high dimensional asymptotics and leads to high misclassification error rates.

To cope with the singularity of sample covariance matrices, the procedure of ridge regression or diagonal loading is proposed [8]. By artificially adding a positive diagonal matrix to the singular sample covariance matrix, it converts a singular sample covariance matrix into an invertible covariance. Similar modifications have been proposed by Friedman to regularize the covariance estimation in LDA, which bring forth the popular regularized discriminant analysis [9]. However, how to choose the optimal regularization parameter is a long-standing research problem. By ignoring the correlations among some features in small sample size, a diagonal linear discriminant analysis is proposed in [1] which performs better on gene expression data than the other classifiers such as nearest-neighbor classifier or decision trees. Ledoit and Wolf derived an asymptotic optimal formula to estimate the regularization parameter and proposed a consistent estimator for the precision matrix, i.e., the inverse of the covariance matrix [10]. However, the method applies only to the situation where the dimension p is less than the sample size n. A novel algorithm for RDA is presented for high-dimensional data in [11], which can estimate the optimal regularization parameters from a large set of candidates efficiently. A maximum uncertainty LDA-based method is proposed in [12]. It is based on a straightforward stabilization of the within-class scatter matrix and has been applied to face recognition. In [13], a new approach which uses a generalization of the Moore-Penrose pseudo inverse of the sample covariance matrix is proposed to remove the problem of singularity and to improve the quality of classification.

Random matrix theory as a powerful theoretical framework is believed to meet the challenges of high-dimensional data, since the "large p, large n" settings in high-dimensional data analysis fall exactly into the realm of random matrix theory. Bun et al. used tools from RMT to build consistent "rotationally invariant" estimators for large correlation matrices when there is no prior information on the structure of the underlying process [14]. RMT also provides a direct way to de-noise sample covariance matrix $\boldsymbol{K}$ by using Marchenko-Pastur law. And the de-noised sample matrix can be used as an estimator for the population covariance matrix $\boldsymbol{\Sigma}$ [15]. Motivated by these developments in random matrix theory, we propose to regularize the linear discriminant classifier by optimally shrinking the eigenvalues of the sample covariance matrix while keeping the eigenvectors unchanged. An extensive simulation analysis is conducted to test the performance of our algorithm in various high-dimensional settings. Experimental results show that our algorithm is more flexible and obtains lower misclassification rates for a variety of data sets, namely a handwritten digit data set and three microarray data sets.

The rest is organized as follows. Some preliminary results of LDA are introduced in Section 2. Our regularized discriminant analysis based on random matrix theory is presented in Section 3. In Section 4 and Section 5, our method is compared with other popular classifiers for the synthetic data sets and some real world data sets. Some concluding remarks are given in Section 6.

2. Preliminaries to Linear Discriminant Analysis.

For classification problems, linear discriminant analysis is a supervised learning method, where one or more new data points (observations) are classified into one of the predefined classes (groups) based on the observed features (variables). LDA is based on the assumption that every probability density within the k-th class is following a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, i.e., the p-dimensional joint probability density function for the k-th class can be modelled as:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)\,\boldsymbol{\Sigma}_k^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}}, \quad k = 1, 2, \ldots, K \tag{1}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and the covariance matrix for class k, respectively. It is assumed further in LDA that the variables for each class share the same covariance matrix, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, $k = 1, \ldots, K$. For a new observed data vector $\boldsymbol{x} \in \mathbb{R}^{1\times p}$, the posterior probability $P(G = k|\boldsymbol{x})$ that $\boldsymbol{x}$ belongs to class k can be obtained by using Bayes' rule

$$P(G = k|\boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\pi_l} \tag{2}$$

where π_k is the prior probability of class k. The optimal classification is obtained by selecting the class k which maximizes the class posteriors $P(G = k|\boldsymbol{x})$,

$$\hat{G}(\boldsymbol{x}) = \arg\max\nolimits_k P(G = k|\boldsymbol{x}) \tag{3}$$

Another equivalent, yet simple, description of the decision rule in Equation (3) is

$$\hat{G}(\boldsymbol{x}) = \arg\max\nolimits_k \delta_k(\boldsymbol{x}) = \arg\max\nolimits_k \left\{ \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k^{\mathrm{T}} - \frac{1}{2}\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k^{\mathrm{T}} + \log \pi_k \right\} \tag{4}$$

In practice, the mean vector $\boldsymbol{\mu}_k$, the covariance matrix $\boldsymbol{\Sigma}$ and the prior probability π_k in Equation (4) are estimated using the training data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ which consists of n labelled observations of the p-dimensional feature vectors. In particular, the covariance matrix $\boldsymbol{\Sigma}$ can be set equal to the overall sample covariance $\boldsymbol{K} = \frac{1}{n-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}(\boldsymbol{X} - \overline{\boldsymbol{X}})$ with $\overline{\boldsymbol{X}}$ denoting the sample mean. The sample covariance matrix $\boldsymbol{K}$ converges almost surely to $\boldsymbol{\Sigma}$ in the case where $p \ll n$. However, $\boldsymbol{K}$ is not a good estimator of $\boldsymbol{\Sigma}$ in the high-dimensional cases, which will result in poor classification performance of LDA. In the following section, two regularization methods will be proposed to improve the performance of LDA.

3. **Random Matrix Regularized Linear Discriminant Analysis.**

Technology advances nowadays have made data collection easier and faster, resulting in data sets with many observations and high dimensions. High-dimensional data sets are characterized by a large quantity of variables p relative to the sample size n. To deal with high-dimensional data sets in machine learning, the process of dimension reduction is usually introduced to reduce the number of variables. In many cases, however, the number of variables after dimension reduction is still quite large. Thus, regularization is needed to maintain their performance as usual.

It has been well known for a long time that, in high-dimensional situations, the sample covariance matrix $\boldsymbol{K}$ is not a good estimator of the covariance $\boldsymbol{\Sigma}$. In the cases where p is close to, or even larger than n, the sample covariance matrix $\boldsymbol{K}$ will become ill-conditioned or even singular. So the precision matrix $\boldsymbol{\Sigma}^{-1}$ in Equation (4) is badly estimated and results in inefficient classifications. In this section, we propose to regularize the linear discriminant analysis by using a consistent estimator of the covariance $\boldsymbol{\Sigma}$ from random matrix theory.

3.1. **Estimation of the covariance based on random matrix theory.**

Random matrix theory is concerned with the study of the diverse properties of matrices (most notably, statistics of matrix eigenvalues) with entries drawn randomly from various probability distributions traditionally referred to as the random matrix ensembles [16]. It has found an extraordinary variety of physical, mathematical and engineering applications, including quantum chaos, complex networks, number theory and wireless communications [17, 18]. Using the mathematical apparatus of random matrix theory, universal statistical properties of a variety of physical systems could be compared and classified [19].

More specifically, to estimate the covariance matrix in high dimensions, a ridge-type shrinkage estimation of a large dimensional precision matrix has been derived based on the asymptotic results from random matrix theory [20]. In [21], Bai et al. have developed a strongly consistent estimator based on the method of moments. Similarly, an approach based on the idea of dimensionality reduction through an ensemble of isotropically random unitary matrices is proposed in [22]. And closed form analytical expressions for the covariance matrix and its inverse in terms of the eigen-decomposition are derived. However, the way to find the optimal loading parameter is not explained in detail. In addition, the algorithms are often computationally intensive, especially in high dimensions.

By considering the number of variables p relative to the sample size n in the high-dimensional setting, two different methods from random matrix theory, namely, the rotational invariant estimation method and the eigenvalues clipping method, are employed to estimate the covariance.

3.1.1. *The case of $n \geq p$.* As above, we denote the $p \times p$ population covariance matrix by $\mathbf{\Sigma}$. And the sample covariance matrix which is obtained from the training data matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{K}$. In the case where the number of variables p is close to the sample size n, the sample covariance matrix $\boldsymbol{K}$ is ill-conditioned or near singular.

To overcome the near singularity of $\boldsymbol{K}$, the rotational invariant estimator is proposed, which can be seen as an optimal nonlinear shrinkage procedure. Before we go into the rotational invariant estimation procedure, we first shift the sample vectors in $\boldsymbol{X}$ to zero mean, to eliminate the effect of different scales. By doing so, we are actually handling the empirical correlation matrix $\boldsymbol{C}$. It has been demonstrated in [14] that $\boldsymbol{C}$ and $\boldsymbol{K}$ share identical statistically properties when $n \to \infty$, $p \to \infty$ up to a rank one perturbation. So we shall work with $\boldsymbol{K}$ henceforth.

In the rotational invariant estimator, the spectral decomposition of $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \sum_{i=1}^{p} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^{\dagger} \tag{5}$$

where μ_i, $i = 1, \ldots, p$, are the real eigenvalues of $\mathbf{\Sigma}$ and $\boldsymbol{v}_i$, $i = 1, \ldots, p$ are the corresponding eigenvectors. Similarly, the sample covariance matrix $\boldsymbol{K}$ can be decomposed as

$$\boldsymbol{K} = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger} \tag{6}$$

with the eigenvalues λ_i and the corresponding eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{K}$. The rotational invariant estimator is expected to find an estimator $\Xi(\boldsymbol{K})$ of the population covariance matrix $\mathbf{\Sigma}$ from $\boldsymbol{K}$ in a rotationally invariant way. More formally, the estimator $\Xi(\boldsymbol{K})$ satisfies:

$$\mathbf{\Omega}\Xi(\boldsymbol{K})\mathbf{\Omega}^{\dagger} = \Xi(\mathbf{\Omega}\boldsymbol{K}\mathbf{\Omega}^{\dagger}) \tag{7}$$

for any rotation matrix $\mathbf{\Omega}$. It has been shown that any rotational invariant estimator $\Xi(\boldsymbol{K})$ shares the same eigenbasis as $\boldsymbol{K}$ [23], that is,

$$\Xi(\boldsymbol{K}) = \sum_{i=1}^{p} \xi_i \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger} \tag{8}$$

where the eigenvalues $[\xi_i]_{i=1,\ldots,p}$ are the quantities that the rotational invariant estimator wishes to estimate.

Given the sample covariance $\boldsymbol{K}$ with the eigenvalues λ_i and the corresponding eigenvectors $\boldsymbol{u}_i$, $i \in \{1, \dots, p\}$, an optimal rotational invariant estimator is

$$\hat{\Xi}(\boldsymbol{K}) = \sum_{i=1}^{p} \hat{\xi}(\lambda_i) \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger} \tag{9}$$

Here $\hat{\xi}(\lambda_i)$ can be found using the Marchenko-Pastur law in random matrix theory and they are given by the following nonlinear mapping [24]

$$\hat{\xi}(\lambda_i) = \frac{\lambda_i}{\left|1 - q + q\lambda_i \lim\limits_{z \to \lambda_i - \mathrm{i}0^-} \mathfrak{g}_{\boldsymbol{K}}(z)\right|^2} \tag{10}$$

where $\mathfrak{g}_{\boldsymbol{K}}(z)$ is the Stieltjes transform in random matrix theory and $q = \frac{p}{n}$. The Stieltjes transform $\mathfrak{g}_{\boldsymbol{K}}(z)$ is useful to study the spectral properties of random matrices in high dimensions. The resolvent of a real symmetric matrix $\boldsymbol{K}_{p \times p}$ is defined as

$$G_{\boldsymbol{K}}(z) = (z\boldsymbol{I}_p - \boldsymbol{K})^{-1} \tag{11}$$

with $z = \lambda - \mathrm{i}\eta$ lying in the lower half of the complex plane. Since $\mathfrak{g}_{\boldsymbol{K}}(z)$ is usually not explicitly solvable, it will be computed numerically. For any large but finite p and n, the limiting Stieltjes transform $\mathfrak{g}_{\boldsymbol{K}}(z)$ is replaced by its discrete form $\mathfrak{g}_{\boldsymbol{K}}^{p}(z)$. The normalized trace of Equation (11) is defined as the Stieltjes transform

$$\mathfrak{g}_{\boldsymbol{K}}^{p}(z) = \frac{1}{p}\mathrm{Tr}[G_{\boldsymbol{K}}(z)] = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{z - \lambda_i} \tag{12}$$

here λ_i, $i = 1, 2, \dots, p$, are the eigenvalues of $\boldsymbol{K}$. This leads to

$$\hat{\xi}(\lambda_i) = \frac{\lambda_i}{|1 - q + qz_i \mathfrak{g}_{\boldsymbol{K}}^{p}(z_i)|^2} \tag{13}$$

with z_i set to be $z_i = \lambda_i - \mathrm{i}p^{-1/2}$.

In order to correct the systematic underestimation of the small eigenvalues of $\boldsymbol{K}$, we need to rescale them by multiplying a factor shown below:

$$\hat{\xi}(\lambda_i) = \hat{\xi}(\lambda_i) \times \max\left(1, \frac{|1 - q + qz_i \mathfrak{g}_{\boldsymbol{K}}^{\mathrm{iw}}(z_i)|^2}{\lambda_i/(1 + \alpha(\lambda_i - 1))}\right) \tag{14}$$

where $\alpha = 1/(1 + 2q\tau)$ with τ being a constant parameter which is assigned to 10 in the numerical implementations and $\mathfrak{g}_{\boldsymbol{K}}^{\mathrm{iw}}(z)$ is the Stieltjes transform of an inverse-Wishart matrix with some parameter τ, as follows:

$$\mathfrak{g}_{\boldsymbol{K}}^{\mathrm{iw}}(z) = \frac{z(1+\tau) - \tau(1-q) \pm \sqrt{(\tau(1-q) - z(1+\tau))^2 - z(z + 2q\tau)(2\tau + 1)}}{z(z + 2q\tau)} \tag{15}$$

Together with Equations (9), (13) and (14), one obtains a complete procedure for the optimal rotational invariant estimator of the covariance in high dimensions. It is rather simple and works perfectly when the sample size n is larger than the number of variables p.

3.1.2. *The case of $n < p$.* In the case of $n < p$, the method of eigenvalues clipping is used to correct the sample eigenvalues. This method is different from the rotational invariant estimator and is an intuitive application of the Marchenko-Pastur law in random matrix theory.

Consider an $n \times p$ random matrix $\boldsymbol{R}$ whose elements come from an independent standard Gaussian distribution. The Marchenko-Pastur law describes the asymptotic behavior of the eigenvalues of the $p \times p$ Wishart matrix $\boldsymbol{W} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$ when both n and p tend to infinity [24]. For $q = \frac{p}{n} \in (0, \infty)$, the largest eigenvalue λ_+ of $\boldsymbol{W}$ converges in probability to $(1 + \sqrt{q})^2$.

In the eigenvalues clipping method, all the eigenvalues beyond the largest expected eigenvalue λ_+ are interpreted as signal while the others are noise. To infer the covariance matrix $\boldsymbol{\Sigma}$ from the sample covariance matrix $\boldsymbol{K}$, we first decompose the matrix $\boldsymbol{K}$ and keep eigenvectors unchanged. Then apply the following scheme to correct the eigenvalues

$$\Xi^{\text{clip}} = \sum_{i=1}^{p} \xi_i \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger}, \qquad \xi_i = \begin{cases} \lambda_i, & \text{if } \lambda_i \geq (1 + \sqrt{q})^2 \\ \bar{\lambda} & \text{otherwise} \end{cases} \tag{16}$$

here $\bar{\lambda}$ is set to be a constant such that the trace of Ξ^{clip} is equal to that of $\boldsymbol{K}$ [25]. This eigenvalues clipping method for covariance estimation has also been found in a number of applications such as gas identification and immunogen design [26, 27].

3.2. Random matrix regularized discriminant analysis (RMRDA) algorithm.

Combining the linear discriminant analysis with the consistent covariance estimator given above, we have the regularized linear discriminant classifier based on random matrix theory.

Following the descriptions in Section 2, we begin with the estimation of $\boldsymbol{\mu}_k$, π_k and $\boldsymbol{\Sigma}$ for each class on the basis of the training data $\boldsymbol{X}_{train}$. However, $\boldsymbol{\Sigma}$ may be ill-conditioned or even singular in the high-dimensional cases. There are two ways to address this problem. The first one is to use rotationally invariant estimator to estimate $\boldsymbol{\Sigma}$ in the case where n is close to p. And the second is the eigenvalue clipping method in the case where p is larger than n. Then, for each test data in $\boldsymbol{X}_{test}$, we compute the decision function δ_k, $k = 1, 2, \ldots, K$ with respect to each class. Finally, we choose the k-th class as the right class with which the decision function in Equation (4) gets its maximum. Our algorithm can be applied effectively not only to the situation where the dimension p is close to the sample size n but also to the situation where p exceeds n.

The pseudocodes for our algorithm are shown in Algorithm 1.

4. Analysis of the Synthetic Data.

In this section, we use the synthetic data to compare the classification performance of our proposed method with other existing methods including DLDA [1], MDMP [28] and smDLDA [29]. We will consider the simulated data generated from three multivariate normal distributions: $N(\mu_1, \boldsymbol{\Sigma})$, $N(\mu_2, \boldsymbol{\Sigma})$ and $N(\mu_3, \boldsymbol{\Sigma})$. And the mean value of the 1st class is set to be $\mu_1 = 0$, while for μ_2 its first 100 values are set to 0.5 and the rest are 0. For the 3rd class, its mean value is $\mu_3 = -\mu_2$.

4.1. Synthetic data models.

A block diagonal covariance matrix is proposed in [30] to mimic the real world data sets and is popularly used in the discriminant analysis algorithm testing. The variables in this model are positively or negatively correlated and the correlations decay as a function of the distance between any pair of variables.

Algorithm 1 Random matrix regularized discriminant analysis (RMRDA)

```
Input: The training data $\boldsymbol{X}_{train}$ and the test data $\boldsymbol{X}_{test}$
Output: The average correct classification rate (ACCR)
Divide the labelled samples in $\boldsymbol{X}_{train}$ to $K$ groups
for $k = 1 : K$ do
    Compute $\boldsymbol{\mu}_k$ and $\pi_k$ in Equation (4)
end for
Compute the sample covariance matrix $\boldsymbol{\Sigma}$ in Equation (4)
if $n \geq p$ then
    Estimate $\boldsymbol{\Sigma}$ using rotational invariant estimator in Subsection 3.1.1
else
    Estimate $\boldsymbol{\Sigma}$ using eigenvalues clipping method in Subsection 3.1.2
end if
for each data vector $x$ in $\boldsymbol{X}_{test}$ do
    for $k = 1 : K$ do
        Compute $\delta_k(\boldsymbol{x})$ in Equation (4)
    end for
    Classify $\boldsymbol{x}$ into the $k$-th class satisfying $\arg\max_k \delta_k(\boldsymbol{x})$
end for
Compute the average correct classification rate
return
```

Similarly, we construct the covariance matrix as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_\rho & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{-\rho} & 0 & \cdots & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_\rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \boldsymbol{\Sigma}_{-\rho} \end{bmatrix}_{p \times p} \tag{17}$$

with

$$\boldsymbol{\Sigma}_\rho = \begin{bmatrix} 1 & \rho & \cdots & \rho^{24} \\ \rho & 1 & \cdots & \rho^{23} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{24} & \rho^{23} & \cdots & 1 \end{bmatrix}_{25 \times 25} \tag{18}$$

where the $(i, j)_{\text{th}}$ entry σ_{ij} in the block matrix $\boldsymbol{\Sigma}_\rho$ is:

$$\sigma_{ij} = \rho^{|i-j|}, \quad 1 \leq i, j \leq 25 \tag{19}$$

Obviously, the correlations between the variables become stronger gradually with the increasing of ρ. To simulate various correlations between the variables and without loss of generality, we will set $\rho = 0.1, 0.3, 0.6$ and 0.8.

4.2. **Simulation results.** In the numerical experiments, we use the covariance model in Equation (17) with different values of ρ to generate the training data. And the number of variables p is set to 1000. Each of the three classes contains the same number of training samples n_k. We also generate additional 1200 samples as test data set. The average correct classification rate (ACCR) for each algorithm is obtained by averaging over 100 runs and the standard deviation is also calculated.

We compare the performance of our algorithm with several other competitors. The average correct classification rates and their standard deviations for different settings are shown in Tables 1 and 2. It can be seen from Table 2 that the LDA classifier performs badly when n is less than p. It also can be seen from Tables 1 and 2 that our algorithm works better than most of the competitors and is only worse than smDLDA in few cases. When the correlations between the variables become stronger, our algorithm is superior to other classifiers and the standard deviation stays very low.

To further demonstrate the performance of our regularized classifier, we consider the case where ρ is fixed to 0.7 and the sample size n varies from 300 to 1,650 with step size 150. The results are shown in Figure 1. It shows that the average correct classification rate

TABLE 1. ACCR and standard deviation for different algorithms ($n = 1200$, $p = 1000$)

	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
LDA	0.776(0.017)	0.790(0.015)	0.887(0.009)	0.972(0.005)
DLDA	0.979(0.009)	0.972(0.008)	0.934(0.022)	0.824(0.045)
MDMP	0.922(0.016)	0.896(0.040)	0.825(0.041)	0.705(0.076)
smDLDA	0.987(0.005)	0.978(0.003)	0.941(0.018)	0.837(0.045)
RMRDA	0.984(0.003)	0.978(0.004)	0.993(0.002)	0.999(0.001)

TABLE 2. ACCR and standard deviation for different algorithms ($n = 900$, $p = 1000$)

	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
LDA	0.361(0.075)	0.293(0.052)	0.312(0.072)	0.328(0.145)
DLDA	0.980(0.008)	0.974(0.011)	0.917(0.022)	0.823(0.051)
MDMP	0.876(0.044)	0.881(0.032)	0.784(0.059)	0.653(0.089)
smDLDA	0.985(0.008)	0.981(0.012)	0.933(0.027)	0.849(0.044)
RMRDA	0.979(0.005)	0.977(0.004)	0.944(0.007)	0.966(0.008)

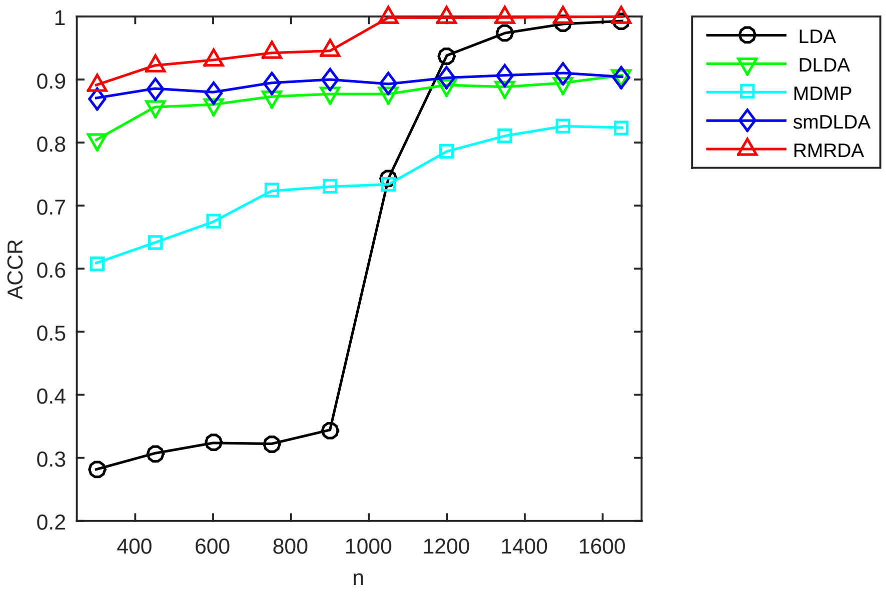

FIGURE 1. ACCR for the classifiers with different sample size n ($p = 1000$)

Table 3. ACCR and standard deviation for RMRDA with unbalanced data ($n = 900$)

Class size and dimension	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
$n_1 = 200,\ n_2 = 300$ $n_3 = 400,\ p = 1000$	0.969(0.007)	0.964(0.005)	0.933(0.007)	0.954(0.008)
$n_1 = 250,\ n_2 = 300$ $n_3 = 350,\ p = 1000$	0.982(0.003)	0.970(0.0039)	0.943(0.007)	0.946(0.006)

Table 4. ACCR and standard deviation for RMRDA with unbalanced data ($n = 1200$)

Class size and dimension	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
$n_1 = 300,\ n_2 = 400$ $n_3 = 500,\ p = 1000$	0.984(0.004)	0.977(0.004)	0.990(0.006)	0.999(0.001)
$n_1 = 350,\ n_2 = 400$ $n_3 = 450,\ p = 1000$	0.985(0.003)	0.980(0.004)	0.990(0.004)	0.999(0.001)

for every classifier rises with the increase of n. Moreover, our algorithm always performs better than other classifiers. In the case of $n < p$, the classification accuracy of LDA is far below expectation. While in the case of $n \geq p$, the classification accuracy of LDA increases sharply and even exceeds that of DLDA, DMDP or smDLDA. The behaviors of our algorithm on unbalanced data are also tested and given in Tables 3 and 4. Again, relatively high classification accuracies for different settings are obtained.

5. **Analysis of Real World Data Sets.**

To examine the performance of our methods in real world applications, we will consider four real world data sets, including a handwritten digit data set and three microarray data sets. A brief description of these real world data sets is provided below.

5.1. **Description of the data set.**

5.1.1. *Handwritten digit data set.* The Multiple Feature (Mfeat) data set is a multi-class data set described in [31], which consists of handwritten digits (from '0' to '9') obtained from Dutch utility maps. The data set includes six different feature sets of the same data, such as Fourier coefficients of the character shapes (fou), Karhunen-Love coefficients (kar), profile correlations (fac), pixel averages (pix), Zernike moments (zer) and morphological features (mor). We have chosen the fac and pix feature sets in our experiments. Each feature set is composed of altogether 2000 digitized images and is divided into ten classes. The fac feature set is described by 216 features while pix is described by 240 features.

5.1.2. *Microarray data set.* Three types of microarray data are used in the study, which are prostate, leukemia and SRBCT, respectively. Prostate data set is collected from patients undergoing radical prostatectomy, which has 10,510 genes and 102 samples. Among the 102 samples, there are 52 prostate tumor samples and 50 normal prostate samples [32].

Golub et al. have presented methods for classifying a leukemia data set consisting of acute myeloid leukemia (AML) samples and acute lymphoblastic leukemia (ALL) samples [33]. Recently, Armstrong et al. have reported that the difference in gene expression is robust enough to classify leukemias correctly as mixed-lineage leukemia (MLL), AML or

ALL [34]. The data set contains 11,225 genes for 28 AML samples, 24 ALL samples and 20 MLL samples.

Small round blue cell tumors (SRBCT) of childhood has been applied for classification in [35]. The data set is divided into four classes, including 29 samples of Ewing's sarcoma (EWS), 25 samples of rhabdomyo sarcoma (RMS), 11 samples of Burkitt's lymphoma (BL), and 18 samples of neuroblastoma (NB). Each sample is described by 2,308 genes.

In order to reduce the computational cost and reveal the practical performance for microarray data sets, a gene selection method is necessary. The BW method, which is a frequently used gene selection method proposed by Dudoit et al. [1], has been applied for discriminant analysis of microarray data. Using this metric, we are able to choose the top p features with the largest BW ratios in the experiments. To choose an appropriate value of p, we have compared the classification accuracy of our algorithm for various p and fixed n. The results are shown in Figure 2. We can see that our algorithm achieves the best classification accuracy when p is approximately equal to 240. So we will choose 240 genes to test the performance of our RMRDA algorithm.

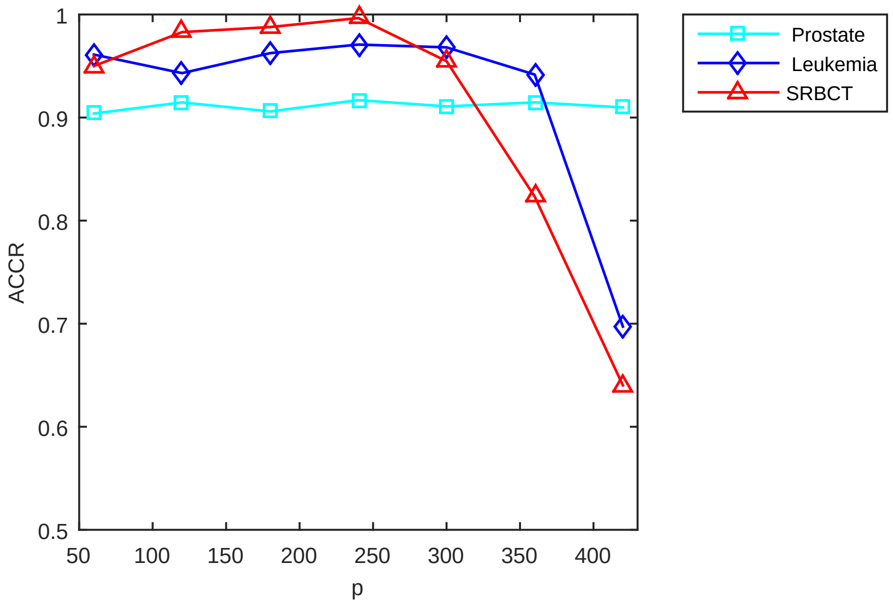

FIGURE 2. ACCR for RMRDA with different p on the microarray data

5.2. **Experimental results.** Each of the real world data set above is split into a training data set and a testing data set for cross validation purpose. The sample size n and the dimension p of the training and testing data sets are summarized in Table 5. As above, each procedure runs 100 times and the average correct classification rates are obtained together with the standard deviations.

TABLE 5. Summary of training and testing data sets

Dataset	Class	Dimension	Training set	Testing set
Mfeat-pix	10	240	260	600
Mfeat-fac	10	216	260	600
Prostate	2	240	51	51
Leukemia	3	240	36	36
SRBCT	4	240	41	42

The classification results for the handwritten digit data set are presented in Table 6. We can see that LDA has a poor performance when the sample size n is close to p. The classification accuracy of MDMP slightly exceeds that of RMRDA on pix data set, but RMRDA still outperform other competitors. For the fac data set, RMRDA shows the best classification performance and maintains high classification accuracy.

Table 6. ACCR and standard deviation for the Mfeat data set

	LDA	DLDA	MDMP	smDLDA	RMRDA
Mfeat-pix	0.402(0.033)	0.918(0.012)	0.940(0.011)	0.918(0.012)	0.932(0.011)
Mfeat-fac	0.071(0.036)	0.887(0.010)	0.884(0.013)	0.887(0.010)	0.951(0.009)

The classification results for the microarray data sets are shown in Table 7. It can be seen that the classification performance of LDA classifier is extremely poor compared with other classifiers while RMRDA always keeps higher classification accuracy and its standard deviation is lower.

Table 7. ACCR and standard deviation for the microarray data set

	LDA	DLDA	MDMP	smDLDA	RMRDA
Prostate	0.517(0.190)	0.834(0.030)	0.922(0.048)	0.830(0.031)	0.911(0.039)
Leukemia	0.319(0.195)	0.904(0.047)	0.901(0.039)	0.906(0.046)	0.965(0.027)
SRBCT	0.201(0.136)	0.978(0.036)	0.835(0.061)	0.977(0.038)	0.994(0.018)

The experimental results demonstrate that, for the high-dimensional data classification problems, our proposed algorithm performs better than the other popular discriminant analysis algorithms. We will attribute the good performance of our algorithm to the flexible estimation of the population covariance matrix in high dimensions.

6. Conclusions.

Linear discriminant analysis is a widely used method for classification. However, it may fail when the number of the features is close to or larger than the sample size. We propose a regularized discriminant analysis method based on random matrix theory. It can handle the high-dimensional data sets. Compared with other popular classifiers, it shows competitive and satisfying performance when evaluated on both the synthetic data sets and the real world data sets.

It has been pointed out recently that the Marchenko-Pastur law is more suitable for the data set in which the ratio of its dimension p to the sample size n is between 0.1 and 10 [36]. This explains, at least in part, why our proposed algorithm does not work well for the ultrahigh-dimensional data sets ($p/n \gg 10$). So how to classify the ultrahigh-dimensional data sets merits further research. Moreover, the proposed regularization method may be further extended to quadratic discriminant analysis with minimal effort.

Acknowledgment. The authors wish to express their gratitude to the Advanced Analysis and Computation Center of CUMT for the award of CPU hours to accomplish this work. This work is supported by the Science and Technology Program of Xuzhou (Grant No. KC18069) and the Postgraduate Research & Practice Program of Education & Teaching Reform of CUMT.

REFERENCES

[1] S. Dudoit, J. Fridlyand and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, vol.97, no.457, pp.77-87, 2002.

[2] O. C. Hamsici and A. M. Martinez, Bayes optimality in linear discriminant analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.30, no.4, pp.647-657, 2008.

[3] M. K. Ng, L. Z. Liao and L. Zhang, On sparse linear discriminant analysis algorithm for high-dimensional data classification, *Numerical Linear Algebra with Applications*, vol.18, no.2, pp.223-235, 2011.

[4] F. Li and N. R. Zhang, Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics, *Journal of the American Statistical Association*, vol.105, no.491, pp.1202-1214, 2010.

[5] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega and P. Vandergheynst, The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains, *IEEE Signal Processing Magazine*, vol.30, no.3, pp.83-98, 2013.

[6] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski and M. E. Martone, Big data from small data: Data-sharing in the 'long tail' of neuroscience, *Nature Neuroscience*, vol.17, no.11, pp.1442-1447, 2014.

[7] M. A. Ahmed, Y. F. Hassan and A. Elsayed, Transfer learning using rough sets for medical data classification, *ICIC Express Letters*, vol.12, no.7, pp.645-654, 2018.

[8] B. D. Carlson, Covariance matrix estimation errors and diagonal loading in adaptive arrays, *IEEE Trans. Aerospace and Electronic Systems*, vol.24, no.4, pp.397-401, 1988.

[9] J. H. Friedman, Regularized discriminant analysis, *Journal of the American Statistical Association*, vol.84, no.405, pp.165-175, 1989.

[10] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, vol.88, no.2, pp.365-441, 2004.

[11] J. Ye and T. Wang, Regularized discriminant analysis for high dimensional, low sample size data, *Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, pp.454-463, 2006.

[12] C. E. Thomaz, E. C. Kitani and D. F. Gillies, A maximum uncertainty LDA-based approach for limited sample size problems-with application to face recognition, *Journal of the Brazilian Computer Society*, vol.12, no.2, pp.7-18, 2006.

[13] T. Gorecki and M. Luczak, Linear discriminant analysis with a generalization of the Moore-Penrose pseudoinverse, *International Journal of Applied Mathematics and Computer Science*, vol.23, no.2, pp.463-471, 2013.

[14] J. Bun, J. P. Bouchaud and M. Potters, Cleaning large correlation matrices: Tools from random matrix theory, *Physics Reports*, vol.666, pp.1-109, 2017.

[15] J. Bai and S. Shi, Estimating high dimensional covariance matrices and its applications, *Annals of Economics and Finance*, vol.12, no.2, pp.199-215, 2011.

[16] T. Gühr, A. Muller-Groeling and H. A. Weidenmüller, Random-matrix theories in quantum physics: Common concepts, *Physics Reports*, vol.299, nos.4-6, pp.189-425, 1998.

[17] P. J. Forrester, N. C. Snaith and J. J. M. Verbaarschot, Developments in random matrix theory, *Journal of Physics A: Mathematical and General*, vol.36, no.12, 2003.

[18] B. Ye, L. Qiu, X. S. Wang and T. Guhr, Spectral statistics in directed complex networks and universality of the Ginibre ensemble, *Communications in Nonlinear Science and Numerical Simulation*, vol.20, no.3, pp.1026-1032, 2015.

[19] M. Müller, G. Baier, A. Galka, U. Stephani and H. Muhle, Detection and characterization of changes of the correlation structure in multivariate time series, *Physical Review E*, vol.71, no.4, 2005.

[20] C. Wang, G. Pan, T. Tong and L. Zhu, Shrinkage estimation of large dimensional precision matrix using random matrix theory, *Statistica Sinica*, vol.25, no.3, pp.993-1008, 2015.

[21] Z. Bai, J. Chen and J. Yao, On estimation of the population spectral distribution from a high-dimensional sample covariance matrix, *Australian and New Zealand Journal of Statistics*, vol.52, no.4, pp.423-437, 2010.

[22] T. L. Marzetta, G. H. Tucci and S. H. Simon, A random matrix-theoretic approach to handling singular covariance estimates, *IEEE Trans. Information Theory*, vol.57, no.9, pp.6256-6271, 2011.

[23] J. Bun, R. Allez and J. P. Bouchaud, Rotational invariant estimator for general noisy matrices, *IEEE Trans. Information Theory*, vol.62, no.12, pp.7475-7490, 2016.

[24] A. Edelman and N. R. Rao, Random matrix theory, *Acta Numerica*, vol.14, pp.233-297, 2005.

[25] L. Laloux, P. Cizeau and M. Potters, Random matrix theory and financial correlations, *International Journal of Theoretical and Applied Finance*, vol.3, no.3, pp.391-397, 2000.

[26] M. Hassan and A. Bermak, Robust Bayesian inference for gas identification in electronic nose applications by using random matrix theory, *IEEE Sensors Journal*, vol.16, no.7, pp.2036-2045, 2016.

[27] A. A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I. Hsing and M. R. McKay, Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design, *Journal of Virology*, vol.88, no.13, pp.7628-7644, 2014.

[28] M. S. Srivastava and T. Kubokawa, Comparison of discrimination methods for high dimensional data, *Journal of the Japan Statistical Society*, vol.37, no.1, pp.123-134, 2007.

[29] T. Tong, L. Chen and H. Zhao, Improved mean estimation and its application to diagonal discriminant analysis, *Bioinformatics*, vol.28, no.4, pp.531-537, 2012.

[30] Y. Guo, T. Hastie and R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, vol.8, no.1, pp.86-100, 2007.

[31] R. P. W. Duin, *UCI Machine Learning Repository*, Department of Applied Physics, Delft University of Technology, Netherlands, 2013.

[32] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, vol.1, no.2, pp.203-209, 2002.

[33] T. R. Golub, D. K. Slonim, P. Tamayo et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, vol.286, no.5439, pp.531-537, 1999.

[34] S. A. Armstrong, J. E. Staunton, L. B. Silverman et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, *Nature Genetics*, vol.30, pp.41-47, 2002.

[35] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, vol.7, pp.673-679, 2001.

[36] O. Ledoit and M. Wolf, Numerical implementation of the QuEST function, *Computational Statistics and Data Analysis*, vol.115, pp.199-223, 2017.