Sequential Linear Discriminant Analysis in High Dimensions Using Individual Discriminant Functions

Seungchul Baek
Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA
baek@umbc.edu

January 29, 2022

Abstract

High dimensional classification has been highlighted for last two decades and much research has been conducted in order to circumvent challenges encountered in high dimensions. While existing methods have focused mainly on developing classification rules assuming independence of covariates or using regularization on the sample covariance matrix or the sample mean vector or among others, we propose a novel approach that employs the “discriminatory power” of each covariate, selects a set of important variables yielding the lowest misclassification rate empirically, and constructs the optimal linear classifier with selected variables. We carry out simulation studies and analyze real data sets to illustrate the performance of our proposed classifier by comparing it with existing classifiers.

Keywords: Discriminatory power; High dimensional classification; Individual discriminant function; Linear discriminant analysis; Variable selection

1 Introduction

The goal of classification is to classify an observation into one of prespecified classes that are categorical response based on some covariates corresponding to the subject. The classification has been used in a variety of disciplines such as image/pattern recognition, bioinformatics, machine learning, statistics, data mining, among others. There have been a lot of
developments in classifiers, which include Fisher discriminant analysis, quadratic discriminant analysis, support vector machines, logistic regression, neural networks, nearest neighbor methods, and many others.

In high dimensions, i.e., the number of covariates \( p \) is greater than the number of samples \( n \) \((p > n)\), many classifiers yield poor performance. For example, under a typical binary classification assuming normality with a common covariance matrix for each group, Bickel and Levina (2004) proved that the misclassification error rate of Fisher discriminant analysis converges to \(1/2\) as \(p/n \to \infty\), which implies that it is no more than random guessing. This is mainly because the sample covariance matrix is ill-conditioned, which accordingly leads to an ill-conditioned precision matrix estimate that is a component of the decision rule. As a simple remedy, one may ignore correlations among covariates, which is called the naive Bayes (NB) rule or the independent rule. The singularity of the sample covariance matrix issue can be addressed by using NB rule in high dimensions, and the performance of NB rule is proven to be superior to one of Fisher discriminant analysis (Bickel and Levina 2004). Some variants of NB have been proposed, e.g., Fan and Fan (2008) proposed the features annealed independence rule (FAIR) which puts some selected variables to use, Greenshtein and Park (2009) presented a naive Bayes type classification rule equipped with nonparametric empirical Bayes, and the nearest shrunken centroid (NSC) method excluding some insignificant variables was proposed in Tibshirani et al. (2003), etc.

Another way to tackle singularity issue is to regularize the sample covariance matrix. This idea goes back to a seminal paper, Hoerl and Kennard (1970), in which they fixed the unstability of the precision matrix by adding positive quantities to diagonal terms of an ill-conditioned matrix in the context of linear regression. We denote \( \hat{\Sigma} \) to be a \(p \times p\) sample covariance matrix, then we may have a form of regularization \( \tilde{\Sigma} \) as follows:

\[
\tilde{\Sigma} = \hat{\Sigma} + \lambda I_p,
\]

where \( I_p \) is a \(p \times p\) identity matrix and \( \lambda \in [0, \infty) \). This idea and its extensions give rise to many high dimensional classifiers, e.g., shrunken centroids regularized linear discriminant
analysis (SCRDA) proposed in Guo et al. (2007), Hastie & Tibshirani (2004), and Friedman (1989).

In addition, a wide range of classifiers in high dimensions have been proposed: Kubokawa & Srivastava (2008) focused on estimating the precision matrix straightly instead of taking an inverse of the sample covariance matrix, Witten & Tibshirani (2011) proposed a classifier penalizing the direction vector of Fisher discriminant analysis, and Fan et al. (2012) presented a method to estimate the direction vector in a different way. For a broad overview of high dimensional classification, see Aoshima et al. (2018), Fan et al. (2011), Fan & Li (2006).

Despite numerous developments in high dimensional classifiers, there seems not exist the optimal classifier that outperforms any other classifiers. Each of them has some downsides, for example, some impose too stringent assumptions such as independent rule ignoring the whole heterogeneous structure, which may be against real data structure in practice, most of regularization processes need to adjust some tuning parameters, and some classification rules are overly involved compared to gains from those models, which may make hard to implement the method and interpret the results. It motivated us to come up with a simple approach leading to comparable performances with high dimensional classifiers in existing literature. Nevertheless, our proposed method, named as sequential linear discriminant analysis (SLDA), may be regarded as a special type of NB or NSC rule since we make use of variance estimate of each covariate separately and select some covariates. However, our proposed classifier is distinct in sense that it collects some significant covariates in a different approach compared to NSC and then establishes a classifier with selected features based on linear discriminant analysis, and in doing so, all dependencies among selected variables are reflected in our proposed classifier, which is also a difference with NB rule.

The rest of this paper is organized as follows. In Section 2, we formulate a binary classification problem in high dimensions and propose a new classifier (SLDA), while defining an individual discriminatory function and discriminatory power of each random variable in order to construct SLDA. We carry out numerical studies to show good performance of our proposed classifier compared to some existing methods using simulations and real data.
examples in Sections 3 and 4. We conclude with a brief discussion in Section 5.

2 Methodology

We begin this section with an introduction of the optimal classification in low dimensions under normality and common covariance assumption. In its extension to high dimensions, one need to face up to some challenges, which are briefly described. We propose a new classifier in high dimensions via a different approach in Section 2.3.

2.1 The Optimal Classifier in Low Dimensions

For simplicity, in this work we consider a binary classification problem, but our proposed method in what follows can be easily extended to multi-class classification problems. We denote \( Y \in \{0, 1\} \) to be the response and the covariates \( X \) is a \( p \)-dimensional random vector, i.e., \( X \in \mathbb{R}^p \). We assume that

\[
X|Y = 0 \sim N(\mu_0, \Sigma), \\
X|Y = 1 \sim N(\mu_1, \Sigma),
\]

where \( \mu_k = E(X|Y = k) \) is the \( p \)-dimensional mean vector for each group, \( k = 0, 1 \), and \( \Sigma \) is a \( p \times p \) common covariance matrix. In addition, we let \( \pi_1 = pr(Y = 1) \) and \( \pi_0 = pr(Y = 0) = 1 - \pi_1 \). It is easy to show that the optimal linear classifier, Bayes rule, is given by

\[
\delta^{Bayes}(x) = \arg\max_k \delta_k(x),
\]

where \( \delta_k(x) \equiv x^T \Sigma^{-1} \mu_k - (1/2) \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \), which is called the linear discriminant function, and \( x \) is an observation to be classified. For the binary classification, (1) can be
written as

\[ \delta^{\text{Bayes}}(x) = \begin{cases} 0, & \text{if } \delta_0(x) > \delta_1(x), \\ 1, & \text{otherwise}, \end{cases} \] (2)

which means that we classify the observation \( x \) into \( Y = 0 \) group if the log-ratio of two conditional densities evaluated at \( x \) is greater than 0, i.e.,

\[
\log \frac{f_0(x)}{f_1(x)} = \delta_0(x) - \delta_1(x) = \left( x - \frac{\mu_0 + \mu_1}{2} \right)^T \Sigma^{-1}(\mu_0 - \mu_1) - \log \frac{\pi_0}{\pi_1} > 0, \] (3)

where \( f_0(x) \) and \( f_1(x) \) is the conditional probability densities of \( x \) given that \( Y = 0 \) and \( Y = 1 \), respectively. Otherwise, if \( \delta_0(x) - \delta_1(x) < 0 \), \( x \) is classified into \( Y = 1 \) group.

In (\ref{eq:fisher}), \( \delta^F(x) \equiv \{x-(\mu_0+\mu_1)/2\}^T \Sigma^{-1}(\mu_0-\mu_1) \) is called the *Fisher discriminant function*. With the language of Fisher discriminant analysis, if \( \delta^F(x) > \log(\pi_0/\pi_1) \), \( x \) is labeled with \( Y = 0 \) group, otherwise, it is classified into \( Y = 1 \) group. Following Fisher’s original idea, a binary classification problem is to find out a direction vector \( w \) separating data best by maximizing

\[
\max_{w \neq 0} \frac{w^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)w^T}{w^T\Sigma w}, \] (4)

where this ratio is called the Rayleigh quotient \citep{Kent1979}. In other words, to maximize the Rayleigh quotient is to seek for the vector \( w \) which maximizes the ratio of the between class variability to within class variability. It can be shown that \( w = \Sigma^{-1}(\mu_0 - \mu_1) \), hence \( \delta^F(x) = w^T \{x - (\mu_0 + \mu_1)/2\} \). In a binary classification problem, Fisher linear discriminant analysis is equivalent to \cite{Duda2012} based on likelihood approach except that it does not assume normality.

Even though \cite{Duda2012} stems from somewhat strict assumption–normality with the same covariance matrix for each group, it shows robust performance on many real data sets in practice.
It may not be a surprise because the equivalence of (1) and (4) implies that they might be robust against deviation from normality assumption. Unfortunately, this property is not guaranteed in high dimensional setting, \( p > n \) \cite{BickelLevina2004,FanFan2008}. Two main obstacles resulting in this phenomenon in high dimensions come from unreliable estimates of the covariance matrix and the mean vector, which deteriorates the optimal classifier. We briefly go over some challenges in establishing the optimal decision rule in high dimensions, and then propose a new linear classifier that avoids those obstacles in the following subsections.

### 2.2 Challenges in High Dimensions

Suppose that we have \( n_0 \) independent and identically distributed (iid) samples from \( Y = 0 \) and \( n_1 \) iid observations from \( Y = 1 \) group, hence we have \( n = n_0 + n_1 \) observations, \( \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) where \( X_i = (X_{i1}, \ldots, X_{ip})^T \). Define some estimators as follows:

\[
\begin{align*}
\bar{X}_0 &= \frac{1}{n_0} \sum_{i=1}^{n} X_i (1 - Y_i), \\
\bar{X}_1 &= \frac{1}{n_1} \sum_{i=1}^{n} X_i Y_i, \\
\hat{\Sigma}_0 &= \frac{1}{n_0 - 1} \sum_{i=1}^{n} (1 - Y_i)(X_i - \bar{X}_0)(X_i - \bar{X}_0)^T, \\
\hat{\Sigma}_1 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n} Y_i(X_i - \bar{X}_1)(X_i - \bar{X}_1)^T, \\
\hat{\Sigma} &= \{(n_0 - 1)\hat{\Sigma}_0 + (n_1 - 1)\hat{\Sigma}_1\}/(n - 2), \\
\hat{\pi}_0 &= n_0/n, \quad \hat{\pi}_1 = n_1/n,
\end{align*}
\]

where \( \hat{\Sigma} \) is a pooled covariance estimator. Therefore, it is natural to plug those estimates in the linear discriminant function

\[
\hat{\delta}_k(x) = x^T\hat{\Sigma}^{-1}x_k - \frac{1}{2}x_k^T\hat{\Sigma}^{-1}x_k + \log(\hat{\pi}_k), \quad (5)
\]

where \( \text{diag}(\hat{\Sigma}) \) is a diagonal matrix whose components are diagonal elements of the pooled covariance matrix estimate and \( k = 0, 1 \). As noted, in high dimensions, \( \hat{\delta}_k \) is along with two
main problems: the one is that $\Sigma$ is singular, hence it can’t be inverted, and the other is that the estimates of mean vectors are not reliable because of insufficient sample size compared to the number of covariates. For singularity issue of the sample covariance matrix, we succinctly introduced some existing methods in Section 1. As a remedy for mean vector estimation, one may implement a James-Stein [James & Stein 1961] type shrinkage estimator, which has in general a form of $X^\text{JS}_k = (1 - \eta)X_k$ for some constants $\eta$, where $X^\text{JS}_k$ is a shrinkage estimator for mean vector of $Y = k$ group. For example, an $\eta$ in [Tong et al. 2012] is given by

$$\eta = \frac{(p - 2)(n_k - 1)}{n_k(n_k - 3)X_k^T \text{diag}(\hat{\Sigma})X_k},$$

where $\text{diag}(\hat{\Sigma})$ is a diagonal matrix whose components are diagonal elements of the pooled covariance matrix estimate and $k = 0, 1$. There are much literature featuring this approach, e.g., [Lin & Tsai 1973], [Efron & Morris 1973], [Gleser 1986].

### 2.3 Sequential Linear Discriminant Analysis (SLDA)

As noted in previous section, our interest hinges on construction of more accurate linear discriminant function (5) in high dimensions. Instead of using some regularization techniques on singular sample covariance matrix or sample mean vector or both of them, we propose a new classifier that is free of regularization and employs only reliable information from the data as they are. In high dimensions, it cannot be achievable, but with a subset of covariates whose size should be less than $n - 2$, one is able to construct the optimal linear discriminant rule. Hence, it can be viewed as a high dimensional classification rule with variable selection, but unlike existing literature, our proposed classifier is established by sequentially assessing a discriminatory power for each covariate.

From the property of multivariate normality, a single random variable $X_j$ for $j = 1, \ldots, p$ is distributed as $X_j|Y = 0 \sim N(\mu_0^T, \sigma_{jj})$ and $X_j|Y = 1 \sim N(\mu_1^T, \sigma_{jj})$, where $\sigma_{jj}$ is the $j$-th diagonal element of $\Sigma$, and $e_j$ is a $p$-dimensional vector with 1 for $j$-th component.
and zero for others. In light of the linear discriminant function, we define an individual discriminant function for a single random variable $X_j$.

**Definition 1.** An individual discriminant function for a random variable $X_j$ is defined as

$$
\delta_k(x_j) \equiv \frac{1}{\sigma_{jj}} x_j \mu_{jk} - \frac{1}{2\sigma_{jj}} \mu_{jk}^2 + \log(\pi_k),
$$

where $\mu_{jk} = e_j^T \mu_k$ for $k = 0, 1$, $j = 1, \ldots, p$.

As the estimation of $\sigma_{jj}$ and $\mu_{jk}$ in (6) does not suffer from high dimensionality, the individual discriminant function can be properly adopted to measure a discriminatory power, which is defined in Definition 2 below. For a binary classification, $\delta_0(x_j) - \delta_1(x_j)$ is the logarithm of the two marginal density ratio of $X_j|Y = 0$ and $X_j|Y = 1$. Hence, for $Y = 0$, it is expected that $\delta_0(x_j) - \delta_1(x_j)$ should be positive, and for $Y = 1$, it should be negative, respectively. Now we define a measure to evaluate the discriminatory power of the random variable $X_j$.

**Definition 2.** We define the discriminatory power of $X_j$, $d_j$ is defined as

$$
d_j = \sum_{i=1}^{n} \left[ (1 - y_i)\{\delta_0(x_{ij}) - \delta_1(x_{ij})\} + y_i\{\delta_1(x_{ij}) - \delta_0(x_{ij})\} \right] 
$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

**Remark 1.** One may think that we define $d_j$ simply as $d_j = \sum_{i=1}^{n} |\delta_0(x_{ij}) - \delta_1(x_{ij})|$ instead of (7). However, it has a drawback. Since $\delta_0(x_{ij}) - \delta_1(x_{ij}) = \log\{f_0(x_{ij})/f_1(x_{ij})\}$, we expect that it should be positive for $Y_j = 0$, and negative for $Y_j = 1$. From the property of the logarithm function, the values of $|\log\{f_0(x_{ij})/f_1(x_{ij})\}|$ may be larger on average when $Y_j = 1$, compared to when $Y_j = 0$, i.e., it is because the derivative of $\log(x)$ is greater when evaluated at $x < 1$ against $x > 1$. For this reason, in order to obtain more elaborated values, we split the components of $d_j$’s based on the group, $Y = 0, 1$. Furthermore, if a random variable has very weak discriminatory power, it implies that $\delta_0 - \delta_1$ and $\delta_1 - \delta_0$ should be close to
around zero. Hence, these values are leveled off around zero and they have little effect on the discriminatory power, \( d_j \).

Now we propose a new classifier, \( \delta^{\text{new}} \) using individual discriminant function, which features two main aspects: (i) assessing the discriminatory power for each covariate using the individual discriminant function, and (ii) selecting a subset of noticeable variables minimizing misclassification rates and then establishing a linear discriminant analysis with selected variables.

**Algorithm for Sequential Linear Discriminant Analysis (SLDA)**

1. For each covariate, \( X_j \) for \( j = 1, \ldots, p \), calculate the individual discriminant function values evaluated at \( x_{ij} \)

   \[
   \delta_k(x_{ij}) = \left( \frac{1}{\hat{\sigma}_{jj}} x_{ij} \hat{\mu}_{jk} - \frac{1}{2 \hat{\sigma}_{jj}} \hat{\mu}_{jk}^2 + \log \hat{\pi}_k \right) I(Y_i = k),
   \]

   for \( i = 1, \ldots, n \) and \( k = 0, 1 \), where \( I(\cdot) \) is an indicator function and \( \hat{\sigma}_{jj}, \hat{\mu}_{jk} \) and \( \hat{\pi}_k \) are estimates for \( \sigma_{jj}, \mu_{jk} \) and \( \pi_k \) in (6).

2. Compute the discriminatory power for each of \( X_j \), \( d_j \) for \( j = 1, \ldots, p \), and sort \( d_j \)'s in a descending order, \( d_{(1)} \geq \cdots \geq d_{(p)} \), where \( d_{(1)} = \max_j d_j \) and \( d_{(p)} = \min_j d_j \). Let \( X_{(j)} \) be the covariate matching with \( d_{(j)} \) for \( j = 1, \ldots, p \).

3. Denote \( X_\ast_m \) to be \( (X_{(1)}, X_{(2)}, \ldots, X_{(m)})^T \), where \( m \in \{1, 2, \ldots, n - 2\} \). Start with a single variable \( X_\ast_1 = X_{(1)} \), and compute misclassification rate, which is denoted by \( \xi_1 \), by using leave-one-out cross-validation (LOOCV). Repeat this process sequentially up to \( m = n - 2 \) and compute the corresponding misclassification rates, \( \xi_1, \xi_2, \ldots, \xi_m \). We define \( m^\ast \) such that

   \[
   m^\ast = \arg\min_{m \in \{1, \ldots, n-2\}} \xi_m.
   \]

In the end, we obtain the set of covariates \( X_S \) leading to minimum classification rate, where \( S \) is the index set of \( X_j \)'s corresponding to \( X_\ast_{(m^\ast)} = \{X_{(1)}, X_{(2)}, \ldots, X_{(m^\ast)}\} \).
4. With the selected covariates from Step 3, for $k = 0, 1$, we construct a new classification rule

$$\delta_{new}(x_S) = \arg \max_k \left( x_S^T \hat{\Sigma}_S^{-1} \hat{\mu}_{S,k} - \frac{1}{2} \hat{\mu}_{S,k}^T \hat{\Sigma}_S^{-1} \hat{\mu}_{S,k} + \log \hat{\pi}_k \right),$$  

(8)

where $x_S$, $\hat{\Sigma}_S$, and $\hat{\mu}_{S,k}$ represent covariates and estimators corresponding only to the selected variables in $S$.

Our proposed classifier (8) is established with a subset of covariates which minimized the misclassification rate empirically. Hence, the SLDA, $\delta_{new}$ indeed conducts variable selection, which can be considered as a by-product in employing our proposed classifier. This point is illustrated through simulations in Section 3.2.

Remark 2. In Step 3 in the SLDA algorithm, considering the number of sample size, one may use 10-fold or 5-fold cross-validations instead of LOOCV. We have found numerically that it is not too affected by choice of cross-validation folds.

3 Simulations

In this section, we conduct a broad range of simulations to see performances of our proposed classification rule based on finite samples. Throughout all simulations, we assume that $p = 500$ and $\mu_0 = 0$, and generate $n_0 = n_1 = 100$ samples for each group. We repeat Monte Carlo experiment 500 times. In Section 3.1 we evaluate performances by comparing misclassification rates, which are computed based on 100 testing samples for each group, i.e., 200 in total. In Section 3.2 we provide performances of variable selection in terms of the true positive rate (TPR) and the false discovery rate (FDR). Moreover, we conducted one more simulation whose setting resembles the one from Guo et al. (2007), through which we have numerically showed the superiority of our proposed method with regard to variable selection.
3.1 Comparison of Misclassification Rates

We compare our proposed classifier with other existing classifiers in high dimensions such as NB, FAIR, NSC, SCRDA, and PLDA in terms of misclassification rate. Some R packages are available to implement existing classifiers: in this work “HiDimDA” for FAIR, “pamr” for NSC, “rda” for SCRDA, and “penalizedLDA” for PLDA are used to obtain misclassification rates from those classifiers. The specific settings of simulations are as follows:

- Simulation 1: The covariates are independent, i.e., the covariance matrix $\Sigma$ is an identity matrix $I_p$. The mean vector of $Y = 1$ group is $\mu_1 = (-1_{10}^T, 1_{10}^T, 0_{p-20}^T)^T$, where $1_q$ is a $q$-dimensional vector with all elements of 1, and $0_q$ a $q$-dimensional zero vector.

- Simulation 2: The mean vector for $Y = 1$ group is $\mu_1 = (-1_{10}^T, 1_{10}^T, 0_{p-20}^T)^T$. We randomly generate every component of a $p \times p$ matrix $A$ from $U(-0.12, 0.12)$, and obtain the covariance matrix by $\Sigma = A^T A$.

- Simulation 3: We assume that $\Sigma$ is a Toeplitz matrix with the $(i,j)$-th element of $1/(|i-j| + 1)$, and $\mu_1 = (1_{10}^T, 0.51_{10}^T, 0.21_{10}^T, 0_{p-20}^T)^T$.

- Simulation 4: The diagonal components of the covariance matrix $\Sigma$ is 1 and its off-diagonal elements have three cases, which are $\rho = 0.25, 0.5, 0.75$, respectively. We set $\mu_1 = (0.51_{20}^T, 0_{p-20}^T)^T$.

- Simulation 5: We assume that the $(i,j)$-th element of $\Sigma$ is $0.9^{|i-j|}$ for $i, j = 1, \ldots, p$ and $\mu_1 = (21_{10}^T, 0_{p-10}^T)^T$.

- Simulation 6: The covariance matrix has a similar form with Simulation 4 except $\rho = 0.2$. Each component of $\mu_1 = (\mu_1, \ldots, \mu_p)^T$ is generated independently from $\mu_j \sim N(0.5, 0.5^2)$ for $j = 1, \ldots, p$.

- Simulation 7: The covariance matrix is a block diagonal matrix given by

$$
\Sigma = \begin{pmatrix}
\Sigma^* & 0 \\
0 & I
\end{pmatrix},
$$

11
where $\Sigma^*$ is a $30 \times 30$ block diagonal matrix consisting of three $10 \times 10$ matrices, $A_1^T A_1$, $A_2^T A_2$, and $A_3^T A_3$, and each of $A_1$, $A_2$, and $A_3$ has the $(i,j)$-th component from $U(0,0.5)$, and $\mu_1 = (0.51^T_{10}, 0^T_{p-10})^T$.

The missclassification rates and their standard errors are provided in Table 1 and we also present boxplots in Figures 1-3. Overall, our proposed method SLDA shows better performances and at least similar results compared to other classifiers. In addition, while NSC and PLDA in Simulation 4 and FAIR in Simulations 5-6 seem more variable as seen in Figures 2 and 3, the variability of our proposed classifier looks stable throughout all simulation settings.

### 3.2 Perspective on Variable Selection

As a by-product of our proposed method, it can be used for variable selection in classification problems, i.e., in the end one is able to obtain the index set $S$ whose corresponding covariates are identified significant in sense that a set of those covariates $X_S$ resulted in the smallest misclassification rate empirically, in other words it led to largest discriminatory power.

In order to access variable selection performances, throughout all simulation settings from Section 3.1, we provide estimates of TPR and FDR which are computed by

$$\hat{TPR} = \frac{\text{# of true variables among selected ones}}{|S^*|},$$

$$\hat{FDR} = \frac{\text{# of false variables among selected ones}}{|S| \vee 1},$$

where $S^*$ is the index set of true significant variables, $| \cdot |$ is the cardinality of a set, and $a \vee b \equiv \max(a,b)$. We present two estimates in Table 2. For $\hat{TPR}$, the results seem promising except for Simulations 3 and 6, however, $\hat{FDR}$ is overall too high. In the case of Simulation 6, since every covariate is significant, it always makes the numerator of (10) be zero, hence $\hat{FDR}$ is zero. It occurred because of an intrinsic property of our proposed method, i.e., the number of selected variable cannot be greater than $m = n - 2 = n_0 + n_1 - 2$. From the point of view of variable selection, one may use our proposed method as a safeguard to obtain a set
containing most of true significant variables, even though it may include more insignificant covariates than it actually does.

The SCRDA method in [Guo et al. (2007)] can be also used for variable selection. They provided variable selection performance, e.g., for two group with independent covariance structure, which is the same setting as Simulation 1 except that $p = 10000$, $\mu_0 = 0_{10000}^T$, $\mu_1 = (0.51_{1000}^T, 0_{9900}^T)^T$, and $n_0 = n_1 = 100$. From Table 7 in [Guo et al. (2007)], $\hat{TPR} = 65.8\%$ and $\hat{FDR} = 82.0\%$ are reported. For direct comparison, we conduct an additional simulation with the same setting based on 500 iterations, and our proposed method SLDA yields $\hat{TPR} = 80.6\%$ and $\hat{FDR} = 5.4\%$. Even though $\hat{TPR}$’s close to each other, $\hat{FDR}$ of SLDA is much smaller than SCRDA, which implies that SLDA seems appealing over SCRDA in variable selection.

4 Applications

We analyze three real examples in order to illustrate the performance of our proposed classifier. Two of them comes from [Dettling (2004)]: The one is colon data set in which there are $p = 2000$ covariates that are gene expression profiles, and the response is a binary variable to indicate if a subject has a tumor or not, $n_0 = 22$ and $n_1 = 40$ patients. The other is gene expression profiles $p = 6033$ of prostate cancer patients whose observations are 102 in total, where there are $n_0 = 50$ and $n_1 = 52$. Another one is from [Wang et al. (2005)], which has 286 subjects with lymph-node-negative breast cancer. The response is whether or not the patient experienced cancer relapse in five years, and those who had relapse are $n_1 = 107$, and those who didn’t are $n_0 = 179$. [Wang et al. (2005)] selected some informative 76 genes and used them instead of the whole genes, $p = 22283$. We analyze this data with the whole genes 22283 in terms of high dimensional classification. For each data set, we employ the LOOCV and provide misclassified cases and its proportion in Table 3. As seen in the table, SLDA is not the best throughout all data sets, but it seems to yield more consistent performance.
5 Discussion

We have proposed a new classifier, SLDA, in high dimensions, which takes a different approach compared to existing methods. First, SLDA chooses a set of covariates yielding the lowest misclassification rate. In doing so, we sequentially add covariates up to $n - 2$ starting from the one having the largest discriminatory power, then for each set of covariates, we compute misclassification rates. We choose the best set of covariates having the minimum misclassification rate. Secondly, with selected variables, we establish the optimal linear classifier. Since we choose covariates at most $n - 2$, SLDA is well constructed without concerns about singularity issue and inaccurate mean vector estimates. In all numerical studies in Sections 3 and 4, SLDA yields comparable or better performances over existing methods, even in the context of variable selection. This is very noteworthy in sense that SLDA does not impose any structural assumptions and employ any regularization techniques, but it focuses more on some fundamental property—the discriminatory power of each covariate—for constructing an appropriate classifier in high dimensions. In addition, we believe that there is open area to investigate further, e.g., we may consider modifying the SLDA where common covariance assumption is violated. We leave it for future work.

References

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. & Marron, J. (2018), ‘A survey of high dimension low sample size asymptotics’, Australian & New Zealand journal of statistics 60(1), 4–19.

Bickel, P. J. & Levina, E. (2004), ‘Some theory for fisher’s linear discriminant function, naive bayes’, and some alternatives when there are many more variables than observations’, Bernoulli 10(6), 989–1010.

Dettling, M. (2004), ‘Bagboosting for tumor classification with gene expression data’, Bioinformatics 20(18), 3583–3593.
Efron, B. & Morris, C. (1973), ‘Stein’s estimation rule and its competitors—an empirical bayes approach’, *Journal of the American Statistical Association* 68(341), 117–130.

Fan, J. & Fan, Y. (2008), ‘High dimensional classification using features annealed independence rules’, *Annals of statistics* 36(6), 2605.

Fan, J., Fan, Y. & Wu, Y. (2011), High-dimensional classification, in ‘High-dimensional data analysis’, World Scientific, pp. 3–37.

Fan, J., Feng, Y. & Tong, X. (2012), ‘A road to classification in high dimensional space: the regularized optimal affine discriminant’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(4), 745–771.

Fan, J. & Li, R. (2006), Statistical challenges with high dimensionality, in ‘Proceedings of the international Congress of Mathematicians. Madrid, Spain’.

Friedman, J. H. (1989), ‘Regularized discriminant analysis’, *Journal of the American statistical association* 84(405), 165–175.

Gleser, L. J. (1986), ‘Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix’, *The Annals of Statistics* pp. 1625–1633.

Greenshtein, E. & Park, J. (2009), ‘Application of non parametric empirical Bayes estimation to high dimensional classification’, *Journal of Machine Learning Research* 10(Jul), 1687–1704.

Guo, Y., Hastie, T. & Tibshirani, R. (2007), ‘Regularized linear discriminant analysis and its application in microarrays’, *Biostatistics* 8(1), 86–100.

Hand, D. J. (2006), ‘Classifier technology and the illusion of progress’, *Statistical science* 21(1), 1–14.

Hastie, T. & Tibshirani, R. (2004), ‘Efficient quadratic regularization for expression arrays’, *Biostatistics* 5(3), 329–340.
Hoerl, A. E. & Kennard, R. W. (1970), ‘Ridge regression: Biased estimation for nonorthogonal problems’, Technometrics 12(1), 55–67.

James, W. & Stein, C. (1961), ‘Estimation with quadratic loss’, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 361–379.

Kent, J., Bibby, J. & Mardia, K. (1979), Multivariate analysis, Academic Press Amsterdam.

Kubokawa, T. & Srivastava, M. S. (2008), ‘Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data’, Journal of multivariate Analysis 99(9), 1906–1928.

Lin, P.-E. & Tsai, H.-L. (1973), ‘Generalized bayes minimax estimators of the multivariate normal mean with unknown covariance matrix’, The Annals of Statistics pp. 142–145.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2003), ‘Class prediction by nearest shrunken centroids, with applications to DNA microarrays’, Statistical Science pp. 104–117.

Tong, T., Chen, L. & Zhao, H. (2012), ‘Improved mean estimation and its application to diagonal discriminant analysis’, Bioinformatics 28(4), 531–537.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J. et al. (2005), ‘Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer’, The Lancet 365(9460), 671–679.

Witten, D. M. & Tibshirani, R. (2011), ‘Penalized classification using Fisher’s linear discriminant’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(5), 753–772.
Table 1: Simulation results with $p = 500$ and sample sizes of $n_0 = n_1 = 100$. Misclassification rates and their sample standard errors are computed based on 500 iterations, with the testing samples size of $n_0 = n_1 = 100$. “SLDA” stands for our proposed classifier, “NB” for naive Bayes, “FAIR” for features annealed independence rule in Fan & Fan (2008), “NSC” for the nearest shrunken centroid in Tibshirani et al. (2003), “SCRDA” for shrunkened centroids regularized linear discriminant analysis in Witten & Tibshirani (2011), and “PLDA” penalized linear discriminant analysis in Guo et al. (2007). The bold font is to represent the smallest misclassification rate, i.e., the best performance.

| Simulation | SLDA | NB  | FAIR | NSC  | SCRDA | PLDA |
|------------|------|-----|------|------|-------|------|
| Simulation 1 | avg 0.0265 | 0.0359 | **0.0152** | 0.0174 | 0.0204 | 0.0178 |
| Simulation 1 | sd 0.0137 | 0.0128 | 0.0088 | 0.0100 | 0.0113 | 0.0099 |
| Simulation 2 | avg 0.1020 | 0.2135 | 0.0970 | 0.0981 | **0.0941** | 0.0972 |
| Simulation 2 | sd 0.0282 | 0.0311 | 0.0247 | 0.0233 | 0.0242 | 0.0232 |
| Simulation 3 | avg 0.2212 | 0.2381 | 0.2086 | 0.2029 | 0.2142 | **0.1984** |
| Simulation 3 | sd 0.0331 | 0.0297 | 0.0322 | 0.0293 | 0.0342 | 0.0277 |
| Simulation 4 | avg **0.2252** | 0.3925 | 0.3356 | 0.2928 | 0.2392 | 0.3264 |
| Simulation 4 | $(\rho = 0.25)$ | sd 0.0355 | 0.0805 | 0.0384 | 0.0682 | 0.0385 | 0.0716 |
| Simulation 4 | $(\rho = 0.5)$ | avg 0.1690 | 0.4323 | 0.3767 | 0.3372 | **0.1584** | 0.3713 |
| Simulation 4 | $(\rho = 0.75)$ | sd 0.0318 | 0.0829 | 0.0410 | 0.0899 | 0.0316 | 0.0864 |
| Simulation 5 | avg **0.0146** | 0.0997 | 0.1404 | 0.0917 | 0.0218 | 0.0905 |
| Simulation 5 | sd 0.0137 | 0.0214 | 0.0386 | 0.0206 | 0.0148 | 0.0206 |
| Simulation 6 | avg 0.0280 | 0.1369 | 0.1575 | 0.0452 | **0.0049** | 0.1035 |
| Simulation 6 | sd 0.0158 | 0.0300 | 0.0323 | 0.0152 | 0.0073 | 0.0251 |
| Simulation 7 | avg **0.2979** | 0.3871 | 0.3673 | 0.3747 | 0.4188 | 0.3741 |
| Simulation 7 | sd 0.0436 | 0.0347 | 0.0342 | 0.0384 | 0.0498 | 0.0357 |
Table 2: Averages of estimates of TPR and FDR in all simulation settings based on 500 iterations. “Sim.4-1”, “Sim.4-2”, are “Sim.4-3” correspond to Simulation 4 settings with $\rho = 0.25$, $\rho = 0.5$, and $\rho = 0.75$, respectively.

|        | Sim.1 | Sim.2 | Sim.3 | Sim.4-1 | Sim.4-2 | Sim.4-3 | Sim.5 | Sim.6 | Sim.7 |
|--------|-------|-------|-------|----------|----------|----------|-------|-------|-------|
| TPR    | 0.9003| 0.9586| 0.5128| 0.9815   | 0.9825   | 0.9826   | 0.9076| 0.1663| 0.9118|
| FDR    | 0.1901| 0.2336| 0.1574| 0.6233   | 0.6437   | 0.6512   | 0.2498| 0.0000| 0.3449|
Table 3: Misclassified cases based on LOOCV. The “case” has a form of “m/n”, where m is the number of misclassified cases and n is the number of samples, and “ratio” is the proportion of misclassified cases. The bold font is to represent the best performance among all classifiers.

|       | SLDA | NB | FAIR | NSC | SCRDA | PLDA |
|-------|------|----|------|-----|-------|------|
| Colon | case | 7/62 | 9/62 | 7/62 | 7/62  | 10/62 | 9/62 |
|       | ratio| 0.1129 | 0.1452 | 0.1129 | 0.1129 | 0.1613 | 0.1452 |
| Prostate | case | 11/102 | 40/102 | 8/102 | 8/102 | 5/102 | 40/102 |
|       | ratio| 0.1078 | 0.3922 | 0.0784 | 0.0784 | 0.0490 | 0.3922 |
| Breast | case | 97/286 | 110/286 | 96/286 | 105/286 | 103/286 | 109/286 |
|       | ratio| 0.3392 | 0.3846 | 0.3357 | 0.3671 | 0.3601 | 0.3811 |
Figure 1: Boxplots of misclassification rates in Simulations 1-3
Figure 2: Boxplots of misclassification rates in Simulation 4
Figure 3: Boxplots of misclassification rates in Simulations 5-7