Abstract

In recent years many sparse linear discriminant analysis methods have been proposed for high-dimensional classification and variable selection. However, most of these proposals focus on binary classification and they are not directly applicable to multiclass classification problems. There are two sparse discriminant analysis methods that can handle multiclass classification problems, but their theoretical justifications remain unknown. In this paper, we propose a new multiclass sparse discriminant analysis method that estimates all discriminant directions simultaneously. We show that when applied to the binary case our proposal yields a classification direction that is equivalent to those by two successful binary sparse LDA methods in the literature. An efficient algorithm is developed for computing our method with high-dimensional data. Variable selection consistency and rates of convergence are established under the ultrahigh dimensionality setting. We further demonstrate the superior performance of our proposal over the existing methods on simulated and real data.

*Department of Statistics, Florida State University (mai@stat.fsu.edu)
†School of Statistics, University of Minnesota (yiyang@umn.edu). Yi and Mai are joint first authors.
‡Corresponding author, School of Statistics, University of Minnesota (zouxx019@umn.edu)
Keywords: Discriminant analysis; High dimensional data; Variable selection; Multiclass classification; Rates of convergence.

1 Introduction

In multiclass classification we have a pair of random variables \((Y, X)\), where \(X \in \mathbb{R}^p\) and \(Y \in \{1, \ldots, K\}\). We need to predict \(Y\) based on \(X\). Define \(\pi_k = \text{pr}(Y = k)\). The linear discriminant analysis model states that

\[
X \mid (Y = k) \sim N(\mu_k, \Sigma), k \in \{1, 2, \ldots, K\}.
\]

(1)

Under (1), the Bayes rule can be explicitly derived as follows

\[
\hat{Y} = \arg \max_k \{(X - \mu_k)^T \beta_k + \log \pi_k\},
\]

(2)

where \(\beta_k = \Sigma^{-1} \mu_k\) for \(k = 1, \ldots, K\). Linear discriminant analysis has been observed to perform very well on many low-dimensional datasets [14, 10]. However, it may not be suitable for high-dimensional datasets for at least two reasons. First, it is obvious that linear discriminant analysis cannot be applied if the dimension \(p\) exceeds the sample size \(n\), because the sample covariance will be singular. Second, [6] showed that even if the true covariance matrix is an identity matrix and we know this fact, a classifier involving all the predictors will be no better than random guessing.

In recent years, many high-dimensional generalizations of linear discriminant analysis have been proposed [17, 18, 5, 6, 20, 15, 3, 19, 13, 9]. In the binary case, the discriminant direction is \(\beta = \Sigma^{-1}(\mu_2 - \mu_1)\). One can seek sparse estimates of \(\beta\) to generalize linear discriminant analysis to deal with high dimensional classification. Indeed, this is the common feature of three popular sparse discriminant analysis methods: the linear programming discriminant [3], the regularized
optimal affine discriminant [9] and the direct sparse discriminant analysis [13]. The linear programming discriminant finds a sparse estimate by the Dantzig selector [4]; the regularized optimal affine discriminant [9] adds the lasso penalty [16] to Fisher’s discriminant analysis; and the direct sparse discriminant analysis [13] derives the sparse discriminant direction via a sparse penalized least squares formulation. The three methods can detect the important predictors and consistently estimate the classification rule with overwhelming probabilities with the presence of ultrahigh dimensions. However, they are explicitly designed for binary classification and do not handle the multiclass case naturally.

Two popular multiclass sparse discriminant analysis proposals are the $\ell_1$ penalized Fisher’s discriminant [19] and sparse optimal scoring [5]. However, these two methods do not have theoretical justifications. It is desirable to have a multiclass sparse discriminant analysis algorithm that is conceptually intuitive, computationally efficient and theoretically sound. To this end, we propose a new sparse discriminant method for high-dimensional multiclass problems. We show that our proposal not only has competitive empirical performance but also enjoys strong theoretical properties under ultrahigh dimensionality. In Section 2 we introduce the details of our proposal after briefly reviewing the existing two proposals. We also develop an efficient algorithm for our method. Theoretical results are given in Section 3. In Section 4 we use simulations and a real data example to demonstrate the superior performance of our method over sparse optimal scoring [5] and $\ell_1$ penalized Fisher’s discriminant [19]. Technical proofs are in an Appendix.
2 Method

2.1 Existing proposals

The Bayes rule under a linear discriminant analysis model is

\[ \hat{Y} = \arg \max_k \{(X - \mu_k)^T \beta_k + \log \pi_k\} \]

where \( \beta_k = \Sigma^{-1}_k \mu_k \) for \( k = 1, \ldots, K \). Let \( \theta_k = \beta_k - \beta_1 \) for \( k = 1, \ldots, K \). Then the Bayes rule can be written as

\[ \hat{Y} = \arg \max_k \{\theta_k^T (X - \mu_k) + \log \pi_k\} \] (3)

We refer to the directions \( \theta = (\theta_2, \ldots, \theta_K) \in \mathbb{R}^{p \times (K-1)} \) as the discriminant directions.

We briefly review two existing multiclass sparse discriminant methods: the sparse optimal scoring [5] and the \( \ell_1 \) penalized Fisher’s discriminant [19]. Instead of estimating \( \theta \) directly, these two methods estimate a set of directions \( \eta = (\eta_1, \ldots, \eta_{K-1}) \in \mathbb{R}^{p \times (K-1)} \) such that \( \eta \) spans the same linear subspace as \( \theta \) and hence linear discriminant analysis on \( X^T \eta \) will be equivalent to (3) on the population level. With a little abuse of terminology, we refer to \( \eta \) as discriminant directions as well. To find \( \eta \), define \( Y_{dm} \) as an \( n \times K \) matrix of dummy variables with \( Y_{dm}^{ik} = 1(Y_i = k) \).

For \( k = 1, \ldots, K - 1 \), sparse optimal scoring estimates \( \eta_k \) by solving the following problem:

\[
(\hat{\alpha}_k, \hat{\eta}_k^{SOS}) = \arg \min_{\alpha_k, \eta_k} \sum_{i=1}^{n} (Y_{dm}^{i} \alpha_k - \tilde{X} \eta_k)^2 + \lambda \| \eta_k \|_1 \\
\text{s.t. } \frac{1}{n} \alpha_k^T (Y_{dm}^{i})^T Y_{dm} \alpha_k = 1, \alpha_k^T (Y_{dm}^{i})^T Y_{dm} \alpha_l = 0, l < k,
\]

where \( \alpha_k \in \mathbb{R}^K \) and \( \tilde{X} \) is the centered data matrix. Note that the constraint \( \alpha_k^T (Y_{dm}^{i})^T Y_{dm} \alpha_l = 0, l < k \) indicates that, \( (\hat{\alpha}_k, \hat{\eta}_k^{SOS}) \) depends on the knowledge of \( (\hat{\alpha}_k, \hat{\eta}_k^{SOS}), l < k \). Hence, the sparse optimal scoring adopts a sequential approach to estimate the discriminant directions.
The $\ell_1$ penalized Fisher’s discriminant analysis estimates $\eta_k$ by

$$
\hat{\eta}_k = \arg \max_{\eta_k} \eta_k^T \hat{\Sigma}_k^k \eta_k + \lambda_k \sum_j |\hat{\sigma}_{jk}| \text{ s.t. } \eta_k^T \tilde{\Sigma} \eta_k \leq 1,
$$

for $k = 1, \ldots, K - 1$, where $\lambda_k$ are tuning parameters, $\hat{\sigma}_{j}^2$ is the $(j,j)$th element of the sample estimate of $\Sigma$, $\tilde{\Sigma}$ is a positive definite estimate of $\Sigma$, $\hat{\Sigma}_k = X^T Y^{dm}((Y^{dm})^T Y^{dm})^{-1/2} \Omega_k ((Y^{dm})^T Y^{dm})^{-1/2} (Y^{dm})^T \hat{X}$ and $\Omega_k$ is the identity matrix if $k = 1$ and otherwise an orthogonal projection matrix with column space orthogonal to $((Y^{dm})^T Y)^{-1/2} Y^T \hat{X}$ for all $l < k$. Since $\Omega_k$ relies on $\hat{\eta}_l$ for all $l < k$, the $\ell_1$ penalized Fisher’s discriminant analysis also finds the discriminant directions sequentially.

### 2.2 Our proposal

Good empirical results have been reported for supporting the $\ell_1$ penalized Fisher’s discriminant analysis and the sparse optimal scoring. However, it is unknown whether either of these two classifiers is consistent when more than two classes are present. Moreover, both sparse optimal scoring and $\ell_1$ penalized Fisher’s discriminant analysis estimate the discriminant directions sequentially. We believe a better multiclass sparse discriminant analysis algorithm should be able to estimate all discriminant directions simultaneously, just like the classical linear discriminant analysis. We aim to develop a new computationally efficient multiclass sparse discriminant analysis method that enjoy strong theoretical properties under ultrahigh dimensionality. Such a method can be viewed as a natural multiclass counterpart of the three binary sparse discriminant methods in [13], [3] and [9].

To motivate our method, we first discuss the implication of sparsity in the multiclass problem.
Note that, by (3), the contribution from the $j$th variable ($X_j$) will vanish if and only if

$$\theta_{2j} = \cdots = \theta_{Kj} = 0$$

(4)

Let $D = \{ j : \text{condition (4) does not hold} \}$. Note that whether an index $j$ belongs to $D$ depends on $\theta_{kj}$ for all $k$. This is because $\theta_{kj}$, $k = 2, \ldots, K$ are related to each other, as they are coefficients for the same predictor. In other words, $\theta_{kj}$, $k = 2, \ldots, K$ are naturally grouped according to $j$. Then the sparsity assumption states that $|D| \ll p$, which is referred to as the common sparsity structure.

Our proposal begins with a convex optimization formulation of the Bayes rule of the multiclass linear discriminant analysis model. Recall that $\theta_k = \Sigma^{-1}(\mu_k - \mu_1)$ for $k = 2, \ldots, K$. On the population level, we have

$$\begin{aligned}
(\theta_2, \ldots, \theta_K) &= \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^T \Sigma \theta_k - (\mu_k - \mu_1)^T \theta_k \right\}.
\end{aligned}$$

(5)

In the classical low-dimension-large-sample-size setting, we can estimate $(\theta_2, \ldots, \theta_K)$ via an empirical version of (5)

$$\begin{aligned}
(\hat{\theta}_2, \ldots, \hat{\theta}_K) &= \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^T \hat{\Sigma} \theta_k - (\hat{\mu}_k - \hat{\mu}_1)^T \theta_k \right\},
\end{aligned}$$

(6)

where $\hat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{y^i = k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T$, $\hat{\mu}_k = \frac{1}{n_k} \sum_{y^i = k} X_i$ and $n_k$ is the sample size within Class $k$. The solution to (6) gives us the classical multiclass linear discriminant classifier.

For presentation purpose, write $\theta_{j} = (\theta_{2j}, \ldots, \theta_{Kj})^T$ and define $\|\theta_{j}\| = (\sum_{i=2}^{K} \theta_{ij}^2)^{1/2}$. For the high-dimensional case, we propose the following penalized formulation for multiclass sparse discriminant analysis.

$$\begin{aligned}
(\hat{\theta}_2, \ldots, \hat{\theta}_K) &= \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^T \Sigma \theta_k - (\hat{\mu}_k - \hat{\mu}_1)^T \theta_k \right\} + \lambda \sum_{j=1}^{p} \|\theta_{j}\|,
\end{aligned}$$

(7)
where $\lambda$ is a tuning parameter. It is clear that (7) is based on (6). In (7) we have used the group lasso [21] to encourage the common sparsity structure. Let $\hat{D} = \{j : \hat{\theta}_{kj} \neq 0\}$ which denotes the set of selected variables for the multiclass classification problem. We will show later that with a high probability $\hat{D}$ equals $D$. One can also use a group version of a nonconvex penalty [7] or an adaptive group lasso penalty [1] to replace the group lasso penalty in (7). To fix main idea, we do not pursue this direction here.

After obtaining $\hat{\theta}_k, k = 2, \ldots, K$, we fit the classical multiclass linear discriminant analysis on $(X^T\hat{\theta}_2, \ldots, X^T\hat{\theta}_K)$, as in sparse optimal scoring and $\ell_1$ penalized Fisher’s discriminant analysis.

We repeat the procedure for a sequence of $\lambda$ values and pick the one with the smallest cross-validation error rate.

### 2.3 Connections with existing binary sparse LDA methods

Although our proposal is primarily motivated by the multiclass classification problem, it can be directly applied to the binary classification problem as well by simply letting $K = 2$ in the formulation (7). It turns out that the binary special case of our proposal has very intimate connections with some proven successful binary sparse LDA methods in the literature. We elaborate more on this point in what follows.

When $K = 2$, (7) reduces to

$$\hat{\theta}^{\text{MSDA}}(\lambda) = \arg\min_{\theta} \frac{1}{2} \theta^T \Sigma \theta - (\hat{\mu}_2 - \hat{\mu}_1) \theta^T + \lambda \|\theta\|_1$$

(8)

Considering the Dantzig selector formulation of (8), we have the following constrained $\ell_1$ minimization estimator defined as

$$\hat{\theta} = \arg\min_{\theta} \|\theta\|_1 \text{ s.t. } \|\Sigma \theta - (\hat{\mu}_2 - \hat{\mu}_1)\|_\infty \leq \lambda.$$ (9)
The above estimator is exactly the linear programming discriminant (LPD) [3].

Moreover, we compare (8) with another two well-known sparse discriminant analysis proposals for binary classification: the regularized optimal affine discriminant (ROAD) [9] and the direct sparse discriminant analysis (DSDA) [13]. Denote the estimates of the discriminant directions given by ROAD and DSDA as $\hat{\theta}^{\text{ROAD}}$ and $\hat{\theta}^{\text{DSDA}}$, respectively. Then we have

$$\hat{\theta}^{\text{ROAD}}(\lambda) = \arg \min_{\theta} \theta^T \hat{\Sigma} \theta + \lambda \|\theta\|_1 \text{ s.t. } \theta^T (\hat{\mu}_2 - \hat{\mu}_1) = 1 \quad (10)$$

$$\hat{\theta}^{\text{DSDA}}(\lambda) = \arg \min_{\theta} \sum_i (Y^i - \theta_0 - (X^i)^T \theta)^2 + \lambda \|\theta\|_1 \quad (11)$$

We derive the following proposition to reveal the connections between our proposal ($K = 2$) and ROAD, DSDA.

**Proposition 1.** Define $c_0(\lambda) = \hat{\theta}^{\text{MSDA}}(\lambda)^T (\hat{\mu}_2 - \hat{\mu}_1)$, $c_1(\lambda) = \hat{\theta}^{\text{DSDA}}(\lambda)^T (\hat{\mu}_2 - \hat{\mu}_1)$ and $a = \frac{n |c_1(\lambda)|}{|c_0(\lambda)|}$. Then we have

$$\hat{\theta}^{\text{MSDA}}(\lambda) = c_0(\lambda) \frac{\hat{\theta}^{\text{ROAD}}(\lambda/|c_0(\lambda)|)}{|c_0(\lambda)|}, \quad (12)$$

$$\hat{\theta}^{\text{MSDA}}(\lambda) = \frac{c_0(\lambda)}{c_1(a\lambda)} \hat{\theta}^{\text{DSDA}}(a\lambda). \quad (13)$$

Proposition 1 shows that the classification direction by our proposal is identical to a classification direction by ROAD and a classification direction by DSDA. Consequently, our proposal ($K = 2$) has the same solution path as ROAD and DSDA.

### 2.4 Algorithm

Besides their solid theoretical foundation, LPD, ROAD and DSDA all enjoy computational efficiency. In particular, DSDA’s computational complexity is the same as fitting a lasso linear regression model. In this section we show that our proposal for the multiclass problem can be solved by
a very efficient algorithm. In light of this and Proposition 1, our proposal is regarded as the natural
class generalization of these successful binary sparse LDA methods.

We now present the efficient algorithm for solving (7). For convenience write \( \hat{\delta}^k = \hat{\mu}_k - \hat{\mu}_1 \).

Our algorithm is based on the following lemma.

**Lemma 1.** Given \( \{\theta_{j'j}, j' \neq j\} \), the solution of \( \theta_j \) to (7) is defined as

\[
\arg\min_{\theta_j} \sum_{k=2}^{K} \frac{1}{2} (\theta_{kj} - \hat{\theta}_{kj})^2 + \frac{\lambda}{\hat{\sigma}_{jj}} \| \theta_j \|
\]

(14)

where \( \hat{\theta}_{k,j} = \frac{\delta_k - \sum_{l \neq j} \hat{\delta}_{lj} \theta_{kl}}{\hat{\sigma}_{jj}} \). Let \( \hat{\theta}_j = (\hat{\theta}_{2j}, \ldots, \hat{\theta}_{Kj})^T \) and \( \| \hat{\theta}_j \| = (\sum_{k=2}^{K} \hat{\theta}_{kj}^2)^{1/2} \). The solution to (14) is given by

\[
\hat{\theta}_j = \hat{\theta}_j \left( 1 - \frac{\lambda}{\| \hat{\theta}_j \|} \right)_+.
\]

(15)

Based on Lemma 1 we use the following blockwise-descent algorithm to implement our mul-
ticlass sparse discriminant analysis.

**Algorithm 1** (Multiclass sparse discriminant analysis for a given penalization parameter).

1. Compute \( \hat{\Sigma} \) and \( \hat{\delta}^k \), \( k = 1, 2, \ldots, K \);

2. Initialize \( \hat{\theta}_k^{(0)} \) and compute \( \hat{\theta}_k^{(0)} \) accordingly;

3. For \( m = 1, \ldots \), do the following loop until convergence: for \( j = 1, \ldots, p \),

   (a) compute

   \[
   \hat{\theta}_j^{(m)} = \hat{\theta}_j^{(m-1)} \left( 1 - \frac{\lambda}{\| \hat{\theta}_j^{(m-1)} \|} \right)_+;
   \]
(b) **update**

\[ \tilde{\theta}_{kj} = \frac{\delta_k^l - \sum_{l \neq j} \hat{\sigma}_{lj} \hat{\theta}_{kl}^{(m)}}{\hat{\sigma}_{jj}}. \]

4. Let \( \hat{\Theta} \) as the solution at convergence. The output classifier is the usual linear discriminant classifier on \((X^\top \hat{\theta}_{2}, \ldots, X^\top \hat{\theta}_{K})\).

We have implemented our method in an R package `msda` which is available upon request. Our package also handles the version of (7) using an adaptive group lasso penalty, because both Lemma 1 and Algorithm 1 can be easily generalized to handle the adaptive group lasso penalty.

3 **Theory**

In this section we study theoretical properties of our proposal under the setting where \( p \) can be much larger than \( n \). Under regularity conditions we show that our method can consistently select the true subset of variables and at the same time consistently estimate the Bayes rule.

We begin with some useful notation. For a vector \( \alpha \), \( \| \alpha \|_{\infty} = \max_j |\alpha_j| \), \( \| \alpha \|_1 = \sum_j |\alpha_j| \),

while, for a matrix \( \Omega \in \mathbb{R}^{m \times n} \), \( \| \Omega \|_{\infty} = \max_i \sum_j |\omega_{ij}| \), \( \| \Omega \|_1 = \max_j \sum_i |\omega_{ij}| \). Define

\[ \varphi = \max\{\| \Sigma_{D^c, D} \|_{\infty}, \| \Sigma_{D, D}^{-1} \|_{\infty} \}, \Delta = \max\{\| \mu \|_1, \| \theta \|_1 \}; \]

\[ \theta_{\min} = \min_{(k,j): \theta_{kj} \neq 0} |\theta_{kj}|, \theta_{\max} = \max_{(k,j)} |\theta_{kj}|; \]

\[ \| \Sigma_{D^c, D} \Sigma_{D, D}^{-1} \|_{\infty} = \eta^*. \]

Let \( d \) be the cardinality of \( D \).

Define \( t_D \in \mathbb{R}^{d \times (K-1)} \) as the subgradient of the group lasso penalty at the true \( \theta_D \) and we assume the following condition:
max_{j \in \mathcal{D}^c} \left\{ \sum_{k=2}^{K} (\boldsymbol{\Sigma}_{j,\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1} t_{k,\mathcal{D}})^2 \right\}^{1/2} = \kappa < 1.

Condition (C0) is required to guarantee the selection consistency. A condition similar to condition (C0) has been used to study the group lasso penalized regression model [1].

For the asymptotic results, we let $\varphi, \Delta, \eta^*, \kappa$ be fixed and further assume the following regularity conditions:

(C1) There exists $c_1, C_1$ such that $\frac{c_1}{K} \leq \pi_k \leq \frac{C_1}{K}$ for $k = 1, \ldots, K$ and $\frac{\theta_{\max}}{\theta_{\min}} < C_1$.

(C2) $n, p, \to \infty$ and $\frac{d^2 \log (pd)}{n} \to 0$;

(C3) $\theta_{\min} \gg \left\{ \frac{d^2 \log (pd)}{n} \right\}^{1/2}$;

(C4) $\min_{k, k'} \left\{ (\theta_k - \theta_{k'})^T \Sigma (\theta_k - \theta_{k'}) \right\}^{1/2}$ is bounded away from 0.

Condition (C1) guarantees that we will have a decent sample size for each class. Condition (C2) requires that $p$ cannot grow too fast with respect to $n$. This condition is very mild, because it can allow $p$ to grow at a nonpolynomial rate of $n$. In particular, if $d = O(n^{1/2-\alpha})$ and $K = \log^M n$ for some $M > 0$, then condition (C2) is satisfied if $\log p = o(n^{2\alpha})$. Condition (C3) guarantees that the nonzero coefficients are bounded away from 0, which is a common assumption in the literature. The lower bound of $\theta_{\min}$ tends to 0 under condition (C3). Condition (C4) is required such that all the classes can be separated from each other. If condition (C4) is violated, even the Bayes rule cannot work well.

**Theorem 1.** Under conditions (C0)–(C4), if $\left\{ \frac{d^2 \log (pd)}{n} \right\}^{1/2} \ll \lambda \ll \theta_{\min}$, then with probability tending to 1, we have $\hat{\mathcal{D}} = \mathcal{D}$, and $\|\hat{\theta}_k - \theta_k\|_{\infty} \leq 4\varphi \lambda$ for $k = 2, \ldots, K$.

Next, we show that our proposal is a consistent estimator of the Bayes rule in terms of the misclassification error rate.
Theorem 2. The prediction by our method is \( \hat{Y}(\hat{\theta}_k, \hat{\pi}_k, k = 1, \ldots, K) \). Define

\[
R_n = \text{pr}(\hat{Y}(\hat{\theta}_k, \hat{\pi}_k, k = 1, \ldots, K) \neq Y \mid \text{observed data}).
\]

Under conditions (C0)–(C4), with probability tending to 1, we have

\[ R_n \to \text{the Bayes error rate}. \]

Remark 2. Based on our proof we can further derive the asymptotic results by letting \( K \) (the number of classes) diverge with \( n \) to infinity. We only need to use more cumbersome notion and bounds, but the analysis remains pretty much the same. To show a clearer picture of the theory, we have focused on the fixed \( K \) case.

4 Numerical Studies

4.1 Simulations

We demonstrate our proposal by simulation. For comparison, we include the sparse optimal scoring and \( \ell_1 \) penalized Fisher’s discriminant analysis in the simulation study. Four simulation models are considered where the dimension \( p = 800 \) and the training set has a sample size \( n = 75K \), where \( K \) is the number of classes in each model. We generate a validation set of size \( n \) to select the tuning parameters for each method. Recall that \( \beta_k = \Sigma^{-1} \mu_k \). We specify \( \beta_k \) and \( \Sigma \) as in the following four models and then let \( \mu_k = \Sigma \beta_k \).

Model 1:

\( K = 4, \Sigma \) has the autoregressive structure with a parameter of 0.5, \( \beta_{ij} = 1.6 \) for \( i = 2j - 1, 2j; j = 1, \ldots, K \).
Model 2:

\[ K = 6, \Sigma = I_5 \otimes \Omega, \] where \( \Omega \) has the compound symmetry structure with a parameter of 0.5, \( \beta_{ij} = 2.5 \) for \( i = 2j - 1, 2j; j = 1, \ldots, K \).

Model 3:

\[ K = 4, \Sigma \] has the compound symmetry structure with a parameter of 0.5, \( \beta_{jk} = k + u_{jk} \) for \( j = 1, \ldots, 4 \), where \( u_{jk} \) follows the uniform distribution over the interval \([-1/4, 1/4]\).

Model 4:

\[ K = 4, \Sigma \] has the autoregressive structure with a parameter of 0.5, \( \beta_{2,1} = \ldots = \beta_{2,8} = 1.2, \beta_{3,1} = \ldots = \beta_{3,4} = -1.2, \beta_{3,5} = \ldots = \beta_{3,8} = 1.2, \beta_{4,2j-1} = -1.2, \beta_{4,2j} = 1.2 \) for \( j = 1, \ldots, 4 \).

The error rates of these methods are listed in Table 1. To compare variable selection performance, we report the number of correctly selected variables (C) and the number of incorrectly selected variables (IC) by each method. We want to highlight two two observations from Table 1. First, our method is the best across all four models. Second, our method is a very good approximation of the Bayes rule in terms of both sparsity and misclassification error rate. Although our method tends to select a few more variables besides the true ones, this can be improved by using the adaptive group lasso penalty [1]. Because the other two methods do not use the adaptive lasso penalty, we do not include the results of our method using the adaptive group lasso penalty for a fair comparison.

### 4.2 A real data example

We further demonstrate the application of our method on the IBD dataset [2]. This dataset contains 22283 gene expression levels from 127 people. These 127 people are either normal people, people
Table 1: Simulation results for Models 1–4. The two competing methods are denoted by the first author of the original papers. In particular, Witten’s method is the $\ell_1$ penalized Fisher’s discriminant analysis, and Clemmensen’s method is the sparse optimal scoring method. The reported numbers are medians based on 500 replicates. Standard errors are in parentheses. The quantity C is the number of correctly selected variables, and IC is the number of incorrectly selected variables.
with Crohn’s disease or people with ulcerative colitis. This dataset can be downloaded from Gene Expression Omnibus with accession number GDS1615. We randomly split the datasets with a 2:1 ratio in a balanced manner to form the training set and the testing set.

It is known that the marginal \( t \)-test screening [6] can greatly speed up the computation for linear discriminant analysis in binary problems. For a multiclass problem the natural generalization of \( t \)-test screening is the \( F \)-test screening. Compute the \( F \)-test statistic for each \( X_j \) defined as

\[
f_j = \frac{\sum_{k=1}^{K} n_k (\hat{\mu}_{kj} - \hat{\mu}_j)^2 / (G - 1)}{\sum_{i=1}^{n} (X_{ij} - \hat{\mu}_{Y_i,j})^2 / (n - G)},
\]

where \( \hat{\mu}_j \) is the sample grand mean for \( X_j \) and \( n_g \) is the within-group sample size. Based on the \( F \)-test statistic, we define the \( F \)-test screening by only keeping the predictors with \( F \)-test statistics among the \( d_n \)th largest. As recommended by many researchers [6, 8, 11], \( d_n \) can be the same as the sample size, if we believe that the number of truly important variables is much smaller than the sample size. Therefore, we let \( d_n = 127 \) for the current dataset.

We estimate the rules given by sparse optimal scoring, \( \ell_1 \) penalized Fisher’s discriminant analysis and our proposal on the training set. The tuning parameters are chosen by 5 fold cross validation. Then we evaluate the classification errors on the testing set. The results based on 100 replicates are listed in Table 2. It can be seen that our proposal achieves the highest accuracy with the sparsest classification rule.

5 Summary

In this paper we have proposed a new formulation to derive sparse multiclass discriminant classifiers. We have shown that our proposal has a solid theoretical foundation and can be solved by
Our Witten Clemmensen

<table>
<thead>
<tr>
<th>Error(%)</th>
<th>7.32(0.972)</th>
<th>21.95(1.10)</th>
<th>9.76(0.622)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted Model Size</td>
<td>25(0.7)</td>
<td>127(0)</td>
<td>27(0.5)</td>
</tr>
</tbody>
</table>

Table 2: Classification and variable selection results on the real dataset. The two competing methods are denoted by the first author of the original papers. In particular, Witten’s method is the $\ell_1$ penalized Fisher’s discriminant analysis, and Clemmensen’s method is the sparse optimal scoring method. All numbers are medians based on 100 random splits. Standard errors are in parentheses.

a very efficient computational algorithm. Our proposal actually gives a unified treatment of the multiclass and binary classification problems. We have shown that the solution path of the binary version of our proposal is equivalent to that by ROAD and DSDA. Moreover, LPD is identical to the Dantzig selector formulation of our proposal for the binary case. In light of these evidence, our proposal is regarded as the natural multiclass generalization of those proven successful binary sparse LDA methods.

**Appendix: Technical Proofs**

*Proof of Proposition 1.* We first show (12).

For a vector $\theta \in \mathbb{R}^p$, Define

\[
L_{\text{MSDA}}(\theta, \lambda) = \frac{1}{2} \theta^T \hat{\Sigma} \theta - (\hat{\mu}_2 - \hat{\mu}_1)^T \theta + \lambda \| \theta \|_1, \tag{16}
\]

\[
L_{\text{ROAD}}(\theta, \lambda) = \theta^T \hat{\Sigma} \theta + \lambda \| \theta \|_1 \tag{17}
\]

Set $\tilde{\theta} = c_0(\lambda)^{-1} \hat{\theta}_{\text{MSDA}}(\lambda)$. Since $\tilde{\theta}^T (\hat{\mu}_2 - \hat{\mu}_1) = 1$, it suffices to check that, for any $\tilde{\theta}'$ such
that \((\tilde{\theta}^T (\tilde{\mu}_2 - \mu_1) = 1,\) we have \(L^{\text{ROAD}}(\tilde{\theta}, \frac{\lambda}{|c_0(\lambda)|}) \leq L^{\text{ROAD}}(\tilde{\theta}', \frac{\lambda}{|c_0(\lambda)|}).\) Now for any such \(\tilde{\theta}',\)

\[
L^{\text{MSDA}}(c_0(\lambda)\tilde{\theta}', \lambda) = c_0(\lambda)^2L^{\text{ROAD}}(\tilde{\theta}', \frac{\lambda}{|c_0(\lambda)|}) - c_0(\lambda) \quad (18)
\]

Similarly,

\[
L^{\text{MSDA}}(c_0(\lambda)\tilde{\theta}, \lambda) = c_0(\lambda)^2L^{\text{ROAD}}(\tilde{\theta}, \frac{\lambda}{|c_0(\lambda)|}) - c_0(\lambda). \quad (19)
\]

Since \(L^{\text{MSDA}}(c_0(\lambda)\tilde{\theta}, \lambda) \leq L^{\text{MSDA}}(c_0(\lambda)\tilde{\theta}', \lambda),\) we have (12).

On the other hand, by Theorem 1 in [12], we have

\[
\hat{\theta}^{\text{DSDA}}(\lambda) = c_1(\lambda)\hat{\theta}^{\text{ROAD}}\left(\frac{\lambda}{n|c_1(\lambda)|}\right) \quad (20)
\]

Therefore,

\[
\hat{\theta}^{\text{ROAD}}(\frac{\lambda}{|c_0(\lambda)|}) = \hat{\theta}^{\text{ROAD}}\left(\frac{\lambda}{n|c_1(\lambda)|}\right) \quad (21)
\]

\[
= \left(c_1(\lambda)\frac{\lambda}{|c_0(\lambda)|}\right)^{-1}\hat{\theta}^{\text{DSDA}}\left(\frac{\lambda}{|c_0(\lambda)|}\right) \quad (22)
\]

\[
= (c_1(\lambda))^{-1}\hat{\theta}^{\text{DSDA}}(a\lambda) \quad (23)
\]

Combine (23) with (12) and we have (13). \(\square\)

**Proof of Lemma 1.** By straightforward calculation, we can write

\[
\frac{1}{2} \theta_k^T \Sigma_k - (\tilde{\mu}_k - \mu_1)^T \theta_k
\]

\[
= \frac{1}{2} \theta_k^2\hat{\sigma}_{jj} + \left(\sum_{l \neq j} \hat{\sigma}_{lj} \theta_{kl} - \hat{\delta}_{j}^k\right) \theta_{kj} + \frac{1}{2} \sum_{m \neq j, l \neq j} \theta_{kl} \theta_{km} \hat{\sigma}_{lm} - \sum_{l \neq j} \hat{\delta}_{k}^l \theta_{kl}.
\]

Therefore, given \(\{\theta_j, j' \neq j\},\) the solution of \(\theta_j\) is defined as

\[
\arg \min_{\theta_{2,j},...,\theta_{K,j}} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_{kj}^2\hat{\sigma}_{jj} + \left(\sum_{l \neq j} \hat{\sigma}_{lj} \theta_{kl} - \hat{\delta}_{j}^k\right) \theta_{kj} \right\} + \lambda\|\theta_j\|,
\]

which is equivalent to (14). It is easy to get (15) from (14) [21]. \(\square\)
In what follows we use \( C \) to denote a generic constant for convenience.

Now we define an oracle “estimator” that relies on the knowledge of \( D \) for a specific tuning parameter \( \lambda \):

\[
\hat{\theta}_D^{\text{oracle}} = \arg \min_{\theta_2, \ldots, \theta_K, D} \sum_{k=2}^{K} \{ \frac{1}{2} \theta_k^T \hat{\Sigma}_{D,D} \theta_k - (\hat{\mu}_{k,D} - \hat{\mu}_{1,D})^T \theta_k \} + \lambda \sum_{j \in D} \| \theta_j \|. \tag{24}
\]

The proof of Theorem 1 is based on a series of technical lemmas.

**Lemma 2.** Define \( \hat{\theta}_D^{\text{oracle}}(\lambda) \) as in (24). Then \( \hat{\theta}_k = (\hat{\theta}_D^{\text{oracle}},0) \), \( k = 2, \ldots, K \) is the solution to (7) if

\[
\max_{j \in D^c} \left\{ \sum_{k=2}^{K} \left\{ \left( \hat{\Sigma}_{D,D}^{\text{oracle}} \right)_j - (\hat{\mu}_{k,j} - \hat{\mu}_{1,j}) \right\}^2 \right\}^{1/2} < \lambda. \tag{25}
\]

**Proof of Lemma 2.** The proof is completed by checking that \( \hat{\theta}_k = (\hat{\theta}_D^{\text{oracle}},0) \) satisfies the KKT condition of (7). \( \square \)

**Lemma 3.** For each \( k \), \( \Sigma_{D,D}^{\text{oracle}}(\mu_{k,D} - \mu_{1,D}) = \mu_{k,D^c} - \mu_{1,D^c} \).

**Proof of Lemma 3.** For each \( k \), we have \( \theta_{k,D^c} = 0 \). By definition, \( \theta_{D^c} = (\Sigma^{-1}(\mu_k - \mu_1))_{D^c} \).

Then by block inversion, we have that

\[
\theta_{k,D^c} = -(\Sigma_{D^c,D^c} - \Sigma_{D^c,D} \Sigma_{D,D} \Sigma_{D,D^c})^{-1}(\Sigma_{D^c,D} \Sigma_{D,D^c}^{-1}(\mu_{k,D} - \mu_{1,D}) - (\mu_{k,D^c} - \mu_{1,D^c}))
\]

and the conclusion follows. \( \square \)

**Proposition 2.** There exist a constant \( \epsilon_0 \) such that for any \( \epsilon \leq \epsilon_0 \) we have

\[
\text{pr}\{|(\hat{\mu}_{k,j} - \hat{\mu}_{1,j}) - (\mu_{k,j} - \mu_{ij})| \geq \epsilon\} \leq C \exp(-C \frac{n \epsilon^2}{K}) + C \exp(-\frac{C n}{K^2}), \tag{26}
\]

\( k = 2, \ldots, K, \ j = 1, \ldots, p; \)

\[
\text{pr}(|\sigma_{ij} - \sigma_{ij}| \geq \epsilon) \leq 2 \exp(-C \frac{n \epsilon^2}{K}) + 2 \exp(-\frac{C n}{K^2}), \ i, j = 1, \ldots, p. \tag{27}
\]
Proof of Proposition 2. We first show (26). Note that, by Chernoff bound
\[
\Pr(\hat{\mu}_{kj} - \mu_{kj} \geq \epsilon) \leq E(\Pr(\hat{\mu}_{kj} - \mu_{kj} \geq \epsilon \mid Y)) \leq E(\exp(-Cn_k\epsilon^2))
\]
\[
\leq 2 \exp(-C'\frac{n\epsilon^2}{K}) + 2 \exp(-\frac{Cn}{K^2}).
\]
A similar inequality holds for \(\hat{\sigma}_{ij}\), and (26) follows.

For (27), note that
\[
\hat{\sigma}_{ij} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{Y^m=k} (X_i^m - \hat{\mu}_{ki})(X_j^m - \hat{\mu}_{kj})
\]
\[
= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{Y^m=k} (X_i^m - \mu_i^m)(X_j^m - \mu_j^m) + \frac{1}{n-K} \sum_{k=1}^{K} \sum_{Y^m=k} n_k(\hat{\mu}_{ki} - \mu_{ki})(\hat{\mu}_{kj} - \mu_{kj})
\]
\[
= \hat{\sigma}_{ij}^{(0)} + \frac{1}{n-K} \sum_{k=1}^{K} n_k(\hat{\mu}_{ki} - \mu_{ki})(\hat{\mu}_{kj} - \mu_{kj}).
\]
Now by Chernoff bound, \(\Pr(|\hat{\sigma}_{ij}^{(0)} - \sigma_{ij}| \geq \epsilon) \leq C \exp(-Cn\epsilon^2)\). Combining this fact with (26), we have the desired result.

Now we consider two events depending on a small \(\epsilon > 0\):
\[
A(\epsilon) = \{|\hat{\sigma}_{ij} - \sigma_{ij}| < \frac{\epsilon}{d} \text{ for any } i = 1, \cdots, p \text{ and } j \in \mathcal{D}\},
\]
\[
B(\epsilon) = \{|(\hat{\mu}_{kj} - \hat{\mu}_{1j}) - (\mu_{kj} - \mu_{1j})| < \epsilon \text{ for any } k \text{ and } j\}.
\]

By simple union bounds, we can derive Lemma 4 and Lemma 5.

**Lemma 4.** There exist a constant \(\epsilon_0\) such that for any \(\epsilon \leq \epsilon_0\) we have

1. \(\Pr(A(\epsilon)) \geq 1 - Cpd \exp(-Cn \frac{\epsilon^2}{Kd^2}) - CK \exp(-\frac{Cn}{K^2})\);

2. \(\Pr(B(\epsilon)) \geq 1 - Cp(K-1) \exp(-Cn \frac{\epsilon^2}{K}) - CK \exp(-\frac{Cn}{K^2})\);
3. \( \Pr(A(\epsilon) \cap B(\epsilon)) \geq 1 - \gamma(\epsilon) \), where

\[
\gamma(\epsilon) = Cpd \exp\left(-C\frac{n\epsilon^2}{d^2}\right) + Cp(K - 1) \exp\left(-C\frac{n\epsilon^2}{K}\right) + 2CK \exp\left(-\frac{Cn}{K^2}\right).
\]

**Lemma 5.** Assume that both \( A(\epsilon) \) and \( B(\epsilon) \) have occurred. We have the following conclusions:

\[
\|\hat{\Sigma}_{D,D} - \Sigma_{D,D}\|_\infty < \epsilon;
\]

\[
\|\hat{\Sigma}_{Dc,D} - \Sigma_{Dc,D}\|_\infty < \epsilon;
\]

\[
\|((\hat{\mu}_k - \hat{\mu}_1) - (\mu_k - \mu_1))\|_\infty < \epsilon;
\]

\[
\|((\hat{\mu}_{k,D} - \hat{\mu}_{1,D}) - (\mu_{k,D} - \mu_{1,D}))\|_1 < \epsilon.
\]

**Lemma 6.** If both \( A(\epsilon) \) and \( B(\epsilon) \) have occurred for \( \epsilon < \frac{1}{\phi} \), we have

\[
\|\hat{\Sigma}_{D,D}^{-1} - \Sigma_{D,D}^{-1}\|_1 < \epsilon \phi^2 (1 - \phi\epsilon)^{-1},
\]

\[
\|\hat{\Sigma}_{Dc,D}(\hat{\Sigma}_{D,D})^{-1} - \Sigma_{Dc,D}(\Sigma_{D,D})^{-1}\|_\infty < \frac{\phi\epsilon}{1 - \phi\epsilon}.
\]

**Proof of Lemma 6.** Let \( \eta_1 = \|\hat{\Sigma}_{D,D} - \Sigma_{D,D}\|_\infty \), \( \eta_2 = \|\hat{\Sigma}_{Dc,D} - \Sigma_{Dc,D}\|_\infty \) and \( \eta_3 = \|\hat{\Sigma}_{D,D}^{-1} - \Sigma_{D,D}^{-1}\|_\infty \). First we have

\[
\eta_3 \leq \|(\hat{\Sigma}_{D,D})^{-1}\|_\infty \times \|\hat{\Sigma}_{D,D} - \Sigma_{D,D}\|_\infty \times \|(\Sigma_{D,D})^{-1}\|_\infty = (\phi + \eta_3)\phi\eta_1.
\]

On the other hand,

\[
\|\hat{\Sigma}_{Dc,D}(\hat{\Sigma}_{D,D})^{-1} - \Sigma_{Dc,D}(\Sigma_{D,D})^{-1}\|_\infty \leq \|\hat{\Sigma}_{Dc,D} - \Sigma_{Dc,D}\|_\infty \times \|(\hat{\Sigma}_{D,D})^{-1} - (\Sigma_{D,D})^{-1}\|_\infty \\
\quad + \|\hat{\Sigma}_{Dc,D} - \Sigma_{Dc,D}\|_\infty \times \|(\Sigma_{D,D})^{-1}\|_\infty \\
\quad + \|\Sigma_{Dc,D}\|_\infty \times \|(\Sigma_{D,D})^{-1} - (\Sigma_{D,D})^{-1}\|_\infty \\
\quad \leq \eta_2\eta_3 + \eta_2\phi + \phi\eta_3.
\]
By $\varphi \eta_1 < 1$ we have $\eta_3 \leq \varphi^2 \eta_1 (1 - \varphi \eta_1)^{-1}$ and hence

$$
\|\hat{\Sigma}^{-1}_{D,D}(\hat{\Sigma}^{-1}_{D,D} - \Sigma^{-1}_{D,D})\|_\infty < \frac{\varphi \epsilon}{1 - \varphi \epsilon}.
$$

\[\Box\]

**Lemma 7.** Define

$$
\hat{\theta}^0_{k,D} = \hat{\Sigma}^{-1}_{D,D}(\hat{\mu}_{k,D} - \hat{\mu}_{1,D}).
$$

(28)

Then $\|\hat{\theta}^0_{k,D} - \theta_{k,D}\|_1 \leq \frac{\varphi \epsilon (1 + \varphi \Delta)}{1 - \varphi \epsilon}$.

**Proof of Lemma 7.** By definition, we have

$$
\|\hat{\Sigma}^{-1}_{D,D}(\hat{\mu}_{k,D} - \hat{\mu}_{1,D}) - \Sigma^{-1}_{D,D}(\mu_{k,D} - \mu_{1,D})\|_1 \\
\leq \|\hat{\Sigma}^{-1}_{D,D} - \Sigma^{-1}_{D,D}\|_1 \|\hat{\mu}_{k,D} - \hat{\mu}_{1,D} - (\mu_{k,D} - \mu_{1,D})\|_1 \\
+ \|\Sigma^{-1}_{D,D}\|_1 \|\hat{\mu}_{k,D} - \hat{\mu}_{1,D} - (\mu_{k,D} - \mu_{1,D})\|_1 + \|\hat{\Sigma}^{-1}_{D,D} - \Sigma^{-1}_{D,D}\|_1 \|\mu_{k,D} - \mu_{1,D}\|_1 \\
\leq \frac{\varphi \epsilon (1 + \varphi \Delta)}{1 - \varphi \epsilon}.
$$

\[\Box\]

**Lemma 8.** If $A(\epsilon)$ and $B(\epsilon)$ have occurred for $\epsilon < \min\{\frac{1}{2\varphi}, \frac{\lambda}{1 + \varphi \Delta}\}$, then for all $k$

$$
\|\hat{\theta}^{\text{oracle}}(\lambda) - \theta_{k,D}\|_\infty \leq 4\varphi \lambda.
$$

**Proof of Lemma 8.** Observe $\hat{\theta}^{\text{oracle}} = \hat{\Sigma}^{-1}_{D,D}(\hat{\mu}_{k,D} - \hat{\mu}_{1,D}) - \lambda \hat{\Sigma}^{-1}_{D,D} \hat{t}_{k,D}$. Therefore,

$$
\|\hat{\theta}^{\text{oracle}} - \theta_{k,D}\|_\infty \\
\leq \|\hat{\theta}^0_{k,D} - \theta_{k,D}\|_\infty + \lambda \|\hat{\Sigma}^{-1}_{D,D} - \Sigma^{-1}_{D,D}\|_1 \|\hat{t}_{k,D}\|_\infty + \lambda \|\Sigma^{-1}_{D,D}\|_1 \|\hat{t}_{k,D}\|_\infty
$$

where $\hat{\theta}^0_{k,D}$ is defined as in (28). Now $\|\hat{t}_{k,D}\|_\infty \leq 1$ and we have

$$
\|\hat{\theta}^{\text{oracle}} - \theta_{k,D}\|_\infty \leq \frac{\varphi \epsilon (1 + \varphi \Delta) + \lambda \varphi}{1 - \varphi \epsilon} < 4\varphi \lambda.
$$

21
We are ready to complete the proof of Theorem 1.

Proof of Theorem 1. For any \( \lambda < \frac{\theta_{\min}}{\kappa \varphi} \) and \( \epsilon < \min\left\{ \frac{1}{2\kappa}, \frac{\lambda}{1 + \varphi \Delta} \right\} \). Consider the event \( A(\epsilon) \cap B(\epsilon) \).

By Lemmas 2,4,8 it suffices to verify (11).

For any \( j \in D^c \), by Lemma 3 we have

\[
\left| (\hat{\Sigma}_{D^c,D} \hat{\theta}_{k,D}^{(\text{oracle})})_j - (\hat{\mu}_{kj} - \hat{\mu}_{1j}) \right|
\]

\[
\leq \left| (\hat{\Sigma}_{D^c,D} \hat{\theta}_{k,D}^{(\text{oracle})})_j - (\Sigma_{D^c,D} \theta_{k,D})_j \right| + \left| (\hat{\mu}_{kj} - \hat{\mu}_{1j}) - (\mu_{kj} - \mu_{1j}) \right|
\]

\[
\leq \left| (\hat{\Sigma}_{D^c,D} \hat{\theta}_{k,D}^{(\text{oracle})})_j - (\Sigma_{D^c,D} \theta_{k,D})_j \right| + \epsilon
\]

\[
\leq \left| (\hat{\Sigma}_{D^c,D} \hat{\theta}_{k,D}^{(0)})_j - (\Sigma_{D^c,D} \theta_{k,D})_j \right| + \epsilon + \lambda \left| (\hat{\Sigma}_{D^c,D} \Sigma_{D^c,D}^{-1} \hat{t}_{k,D})_j \right|
\]

\[
(\hat{\Sigma}_{D^c,D} \hat{\theta}_{k,D}^{(\text{oracle})})_j - (\Sigma_{D^c,D} \theta_{k,D})_j \leq \epsilon
\]

\[
\leq \| (\hat{\Sigma}_{D^c,D})_j - (\Sigma_{D^c,D})_j \|_1 \| \hat{\theta}^0_{k,D} - \theta_{k,D} \|_\infty + \| \theta_{k,D} \|_\infty \| (\hat{\Sigma}_{D^c,D})_j - (\Sigma_{D^c,D})_j \|_1
\]

\[
+ \| (\Sigma_{D^c,D})_j \|_1 \| \hat{\theta}^0_{k,D} - \theta_{k,D} \|_\infty + \epsilon
\]

\[
\leq C \epsilon.
\]  

(29)

\[
\left| (\hat{\Sigma}_{D^c,D} \Sigma_{D^c,D}^{-1} \hat{t}_{k,D})_j - (\Sigma_{D^c,D} \Sigma_{D^c,D}^{-1} t_{k,D})_j \right|
\]

\[
\leq \| \hat{\Sigma}_{D^c,D} \Sigma_{D^c,D}^{-1} - \Sigma_{D^c,D} \Sigma_{D^c,D}^{-1} \|_\infty \| \hat{t}_{k,D} - t_{k,D} \|_\infty
\]

\[
+ \| \Sigma_{D^c,D} \Sigma_{D^c,D}^{-1} \|_\infty \| \hat{t}_{k,D} - t_{k,D} \|_\infty + \| \Sigma_{D^c,D} \Sigma_{D^c,D}^{-1} - \Sigma_{D^c,D} \Sigma_{D^c,D}^{-1} \|_\infty \| (t_{k,D})_j \|
\]
\[ |\hat{t}_{kj} - t_{kj}| = \frac{|\hat{\theta}_{kj}||\theta_j| - \theta_{kj}||\hat{\theta}_j|}{||\theta_j||^2||\hat{\theta}_j||} \leq \frac{|\hat{\theta}_{kj} - \theta_{kj}||\theta_j| + \theta_{\text{max}}||\theta_j - \hat{\theta}_j||}{||\theta_j||^2||\hat{\theta}_j||} \leq \frac{C\varphi}{\theta_{\text{min}}\sqrt{(K - 1)}}\lambda. \]

Therefore,

\[
\lambda|\Sigma_{Dc,D}^{-1}\Sigma_{D,D}^{-1}t_{k,D})_j| \leq \lambda|\Sigma_{Dc,D}^{-1}\Sigma_{D,D}^{-1}t_{k,D})_j| + \frac{C\varphi}{\theta_{\text{min}}\sqrt{(K - 1)}}\lambda^2(\frac{\varphi\epsilon}{1 - \varphi\epsilon} + H_\infty) + \frac{\varphi\epsilon\lambda}{1 - \varphi\epsilon}|(t_{k,D})_j|. \quad (30)
\]

Combining (29) and (30) we have (11).

\[ R_n - R \leq \text{pr}(\hat{Y}(\theta_2, \ldots, \theta_K) \neq \hat{Y}(\hat{\theta}_2, \ldots, \hat{\theta}_K)) \leq 1 - \text{pr}(|\hat{t}_k - l_k| < \epsilon/2, |l_k - l_{k'}| > \epsilon) \leq \text{pr}(|\hat{t}_k - l_k| \geq \epsilon/2) + \text{pr}(|l_k - l_{k'}| \leq \epsilon). \]

Now, for \( X \) in each class, \( l_k - l_{k'} \) is normal with variance \( (\theta_k - \theta_{k'})^\text{T}\Sigma(\theta_k - \theta_{k'}) \). Therefore,

\[
\text{pr}(|l_k - l_{k'}| \leq \epsilon) = \sum_{k''} \text{pr}(|l_k - l_{k'}| \leq \epsilon | Y = k'')\pi_{k''} \leq K^2 \sum_{k''} \pi_{k''} \{((\theta_k - \theta_{k'})^\text{T}\Sigma(\theta_k - \theta_{k'}))^2\}^{1/2} \leq CK^2\epsilon.
\]

Proof of Theorem 2. Define \( \hat{Y}(\theta_2, \ldots, \theta_K) \) as the prediction by the Bayes rule and \( \hat{Y}(\hat{\theta}_2, \ldots, \hat{\theta}_K) \) as the prediction by the estimated classification rule. Also define \( l_k = (X - \frac{\mu_k}{2})^\text{T}\theta_k + \log(\pi_k) \) and \( \hat{l}_k = (X - \frac{\mu_k}{2})^\text{T}\hat{\theta}_k + \log(\hat{\pi}_k) \).

For any \( \epsilon > 0 \)

\[ R_n - R \leq \text{pr}(\hat{Y}(\theta_2, \ldots, \theta_K) \neq \hat{Y}(\hat{\theta}_2, \ldots, \hat{\theta}_K)) \leq 1 - \text{pr}(|\hat{t}_k - l_k| < \epsilon/2, |l_k - l_{k'}| > \epsilon) \leq \text{pr}(|\hat{t}_k - l_k| \geq \epsilon/2) + \text{pr}(|l_k - l_{k'}| \leq \epsilon). \]
On the other hand, conditional on training data, $\hat{l}_k - l_k$ is normal with mean $u(k, k') = \mu_k^T (\hat{\mu}_k - \mu_k) + \frac{1}{2} (\mu_k^T \theta_k - \hat{\mu}_k^T \hat{\theta}_k) + \log \hat{\pi}_k - \log \pi_k$ and variance $(\hat{\theta}_k - \theta_k)^T \Sigma (\hat{\theta}_k - \theta_k)$ within class $k'$. By Markov’s inequality, we have

$$\Pr(|\hat{l}_k - l_k| \geq \epsilon/2) = \sum_{k'} \pi_{kk'} \Pr(|\hat{l}_k - l_k| \geq \epsilon/2 | Y = k') \leq C \max_{k} (\hat{\theta}_k - \theta_k)^T \Sigma (\hat{\theta}_k - \theta_k) \left( \epsilon - u(k, k') \right)^2.$$ 

Moreover,

$$|u(k, k')| \leq |\mu_{kk'} (\hat{\theta}_k - \theta_k)| + \frac{1}{2} |\mu_k^T (\hat{\mu}_k - \mu_k)| + \frac{1}{2} |(\mu_k - \hat{\mu}_k)^T \hat{\theta}_k| + |\log \hat{\pi}_k - \log \pi_k| \leq O_P (\log (pd) \frac{d}{\sqrt{n}}) + K O_P (\frac{1}{\sqrt{n}});$$

$$\quad \quad (\hat{\theta}_k - \theta_k)^T \Sigma (\hat{\theta}_k - \theta_k) \leq O_P (\log (pd) \frac{d}{\sqrt{n}}).$$

Hence, pick $\epsilon = o(1/K^2)$, then $R_n \to R$ with a probability tending to 1.

References


