<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2016-0117.R2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Hui Zou  
Qing Mai and  
Yi Yang |
| **Corresponding Author** | Hui Zou |
| **E-mail** | zouxx019@umn.edu |

Notice: Accepted version subject to English editing.
Multiclass Sparse Discriminant Analysis

Qing Mai\textsuperscript{a}, Yi Yang\textsuperscript{b} and Hui Zou\textsuperscript{c}

\textsuperscript{a} Department of Statistics, Florida State University
\textsuperscript{b} Department of Mathematics and Statistics, McGill University
\textsuperscript{c} School of Statistics, University of Minnesota

Abstract: In recent years several 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. Some 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 linear discriminant analysis methods in the literature. Thus, our proposal offers a neat unification of these seemingly unrelated proposals for binary sparse discriminant analysis. Our method can be solved by an efficient algorithm which is implemented in an open R package msda available from CRAN. We offer theoretical justification of our method by establishing a variable selection consistency result and rates of convergence under the ultrahigh dimensionality setting. We
further demonstrate the empirical performance of our method on simulated and real data.

Key words and phrases: 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 \(\pi_k = \Pr(Y = k)\). The linear discriminant analysis model states that

\[
X \mid (Y = k) \sim N(\mu_k, \Sigma), k \in \{1, 2, \ldots, K\}.
\] (1.1)

Under (1.1), the Bayes rule can be explicitly derived as follows

\[
\hat{Y} = \arg \max_k \{(X - \mu_k/2)^T \beta_k + \log \pi_k\},
\] (1.2)

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 \((\text{Michie et al. 1994; Hand 2006})\). 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 matrix will be singular. Second, \(\text{Bickel and Levina 2004}\) and \(\text{Fan and Fan 2008}\) 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 (Tibshirani et al., 2002; Trendafilov and Jolliffe, 2007; Clemmensen et al., 2011; Donoho and Jin, 2008; Fan and Wu, 2008; Wu et al., 2008; Shao et al., 2011; Cai and Liu, 2011; Witten and Tibshirani, 2011; Mai et al., 2012; Fan et al., 2012). 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 (Cai and Liu, 2011), the regularized optimal affine discriminant (Fan et al., 2012) and the direct sparse discriminant analysis (Mai et al., 2012). The linear programming discriminant finds a sparse estimate by the Dantzig selector (Candes and Tao, 2007); the regularized optimal affine discriminant (Fan et al., 2012) adds the lasso penalty (Tibshirani, 1996) to Fisher’s discriminant analysis; and the direct sparse discriminant analysis (Mai et al., 2012) 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.

As pointed out by a referee, one approach to utilize the binary classification methods in multiclass problems with $K$ classes is to break the multiclass problems into $K(K - 1)/2$ pairwise problems. In each of these $K(K - 1)/2$ problems, we apply a binary classifier. Then an observation is classified to the class that it is most often assigned to by the binary classifiers. While this approach could extend the binary classification methods to multiclass problems, in practice there could be two or more classes that an observation is most often classified to. In such cases the pairwise classification will not be able to predict the class label for the observation. Hence, it is important to have a classification method that directly targets multiclass problems and does not suffer from this type of ambiguity.

Two popular multiclass sparse discriminant analysis proposals are the $\ell_1$ penalized Fisher’s discriminant (Witten and Tibshirani 2011) and sparse optimal scoring (Clemmensen et al. 2011). However, these two methods do not have theoretical justifications. It is generally unknown whether they can
select the true variables with high probabilities, how close their estimated
discriminant directions are to the true directions, and whether the final
classifier will work similarly as the Bayes rule.

Therefore, it is desirable to have a new multiclass sparse discriminant
analysis algorithm that is conceptually intuitive, computationally efficient
and theoretically sound. To this end, we propose a new sparse discrimi-
nant 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 (Clemmensen et al., 2011) and $\ell_1$ penal-
ized Fisher’s discriminant (Witten and Tibshirani, 2011). Technical proofs
are in the supplementary materials.

2. Method

2.1 Existing proposals

The Bayes rule under a linear discriminant analysis model is

$$\hat{Y} = \arg \max_k \{(X - \mu_k^2)^T \beta_k + \log \pi_k\},$$
where $\beta_k = \Sigma^{-1}\mu_k$ for $k = 1, \ldots, K$. Let $\theta^\text{Bayes}_k = \beta_k - \beta_1$ for $k = 1, \ldots, K$. Then the Bayes rule can be written as

$$\hat{Y} = \arg \max_k \{ (\theta^\text{Bayes}_k)^T (X - \frac{\mu_1 + \mu_k}{2}) + \log \frac{\pi_k}{\pi_1} \}.$$  \hfill (2.1)

We refer to the directions $\theta^\text{Bayes} = (\theta^\text{Bayes}_2, \ldots, \theta^\text{Bayes}_K) \in \mathbb{R}^{p \times (K-1)}$ as the discriminant directions.

We briefly review two existing multiclass sparse discriminant methods: the sparse optimal scoring (Clemmensen et al., 2011) and the $\ell_1$ penalized Fisher’s discriminant (Witten and Tibshirani, 2011). Instead of estimating $\theta^\text{Bayes}$ 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 $\theta^\text{Bayes}$ and hence linear discriminant analysis on $X^T \eta$ will be equivalent to (2.1) on the population level. More specifically, these two methods look for estimates of $\eta = (\eta_1, \ldots, \eta_{K-1})$ in Fisher’s discriminant analysis:

$$\eta_k = \arg \max \eta_k^T \Sigma_b \eta_k, \text{ s.t. } \eta_k^T \Sigma \eta_k = 1, \eta_k^T \Sigma \eta_l = 0 \text{ for } l < k, \hfill (2.2)$$

where $\Sigma_b = \frac{1}{K-1} \sum_{k=1}^{K} (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^T$ with $\bar{\mu} = \frac{1}{K} \sum_k \mu_k$.

With a little abuse of terminology, we refer to $\eta$ as discriminant directions as well. To find $\eta$, define $Y^{\text{dm}}$ as an $n \times K$ matrix of dummy variables with $Y_{ik}^{\text{dm}} = 1(Y_i = k)$.

In addition to the discriminant direction $\eta_k$, sparse optimal scoring cre-
ates $K - 1$ vectors of scores $\alpha_1, \ldots, \alpha_{K-1} \in \mathbb{R}^K$. Then for $k = 1, \ldots, K - 1$, sparse optimal scoring estimates $\eta_k$ sequentially. In each step, sparse optimal scoring finds $\hat{\alpha}_k, \hat{\eta}_{SOS}$. Suppose the first $k - 1$ score vectors $\hat{\alpha}_l$, $l < k$ and discriminant directions $\hat{\eta}_{SOS}^l$, $l < k$ are available. Then sparse optimal scoring finds $\hat{\alpha}_k, \hat{\eta}_{SOS}^k$ by solving the following problem:

\[
(\hat{\alpha}_k, \hat{\eta}_{SOS}^k) = \arg \min_{\alpha_k, \eta_k} \sum_{i=1}^n (Y_{dm}^i \alpha_k - \tilde{X} \eta_k)^2 + \lambda \| \eta_k \|_1 \tag{2.3}
\]

s.t. \[
\frac{1}{n} \alpha_k^T (Y_{dm})^T Y_{dm} \alpha_k = 1, \quad \alpha_k^T (Y_{dm})^T Y_{dm} \hat{\alpha}_l = 0, \quad \text{for any } l < k,
\]

where $\tilde{X}$ is the centered data matrix, and $\lambda$ is a tuning parameter. The sparse optimal scoring is closely related to $\textbf{(2.2)}$, because when the dimension is low, the unpenalized version of $\textbf{(2.3)}$ gives the same directions (up to a scalar) as $\textbf{(2.2)}$ with the parameters $\Sigma_b$ and $\Sigma$ substituted with the sample estimates. Therefore, with the $\ell_1$ penalty, sparse optimal scoring gives sparse approximations to $\eta$.

Note that the constraint $\alpha_k^T (Y_{dm})^T Y_{dm} \alpha_l = 0, l < k$ indicates that, $\hat{\alpha}_k, \hat{\eta}_{SOS}^k$ depends on the knowledge of $\hat{\alpha}_l, \hat{\eta}_{SOS}^l, l < k$. This is why we say that 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

\[
\hat{\eta}_k = \arg \max_{\eta_k} \eta_k^T \hat{\Sigma}_b \eta_k + \lambda_k \sum_j |\hat{\sigma}_{kj}| \quad \text{s.t. } \eta_k^T \hat{\Sigma} \eta_k \leq 1,
\]
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 $\Sigma$, $\hat{\Sigma}$ is a positive definite estimate of $\Sigma$,

$$
\hat{\Sigma}_k = X^T Y^{dm} \left( (Y^{dm})^T Y^{dm} \right)^{-1/2} \Omega_k \left( (Y^{dm})^T Y^{dm} \right)^{-1/2} Y^{dm}^T \Omega_k^{-1} \left( (Y^{dm})^T Y^{dm} \right)^{-1/2} X
$$

(2.4)

and $\Omega_k$ is the identity matrix if $k = 1$ and otherwise an orthogonal projection matrix with column space orthogonal to $((Y^{dm})^T Y^{dm})^{-1/2} Y^T X \hat{\eta}_l$ for all $l < k$. Again, if the dimension is low, then unpenalized version of (2.4) is equivalent to (2.2) with the parameters replaced by the sample estimates. 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
Multiclass Sparse Discriminant Analysis

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 [Mai et al. (2012), Cai and Liu (2011) and Fan et al. (2012)].

To motivate our method, we first discuss the implication of sparsity in the multiclass problem. As explained in [Mai et al. (2012)], the right target for variable selection should be the subset of variables that influences the Bayes rule. Note that, by (2.1), the contribution from the \( j \)th variable \((X_j)\) will vanish if and only if

\[
\theta_{2j}^{\text{Bayes}} = \cdots = \theta_{Kj}^{\text{Bayes}} = 0 \tag{2.5}
\]

Let \( \mathcal{D} = \{ j : \text{condition (2.5) does not hold} \} \). Note that whether an index \( j \) belongs to \( \mathcal{D} \) depends on \( \theta_{kj} \) for all \( k \). This is because \( \theta_{kj}^{\text{Bayes}}, k = 2, \ldots, K \) are related to each other, as they are coefficients for the same predictor. In other words, \( \theta_{kj}^{\text{Bayes}}, k = 2, \ldots, K \) are naturally grouped according to \( j \). A successful multiclass sparse LDA method should correctly identify \( \mathcal{D} \), at least in theory.

The group sparsity structure under the multiclass LDA model also suggests a potential disadvantage of the sequential methods reviewed in Section 2.1. In the sequential procedures, the directions are estimated one by one, and it is less likely to estimate all the coefficients of one predictor to be zero.
Hence, the sequential methods do not utilize all the available information and are prone to loss of accuracy.

Mai et al. (2012) developed a binary sparse LDA method by taking advantage of a close link between the LDA and the ordinary least squares. By doing so, one can use any software for solving sparse penalized linear regression to fit the sparse LDA classifier proposed in Mai et al. (2012). We would like to extend such a strategy to handle the multiclass case. However, the special connection between the classical LDA and the ordinary least squares only holds for the binary case. If we do that for the multiclass case, we only obtain misleading classification results—a phenomenon called masking; see the detailed explanations given in Chapter 4.2 of Hastie et al. (2009). So the naive application of the penalized multiple responses linear regression will not give us a good multiclass sparse LDA method. On the other hand, we observe that theoretically speaking, the binary sparse LDA proposal in Mai et al. (2012) is equivalent to a sparse penalized quadratic criterion. Computationally speaking, a penalized quadratic problem is as efficient as the penalized least squares. Thus, we want to develop a multiclass sparse LDA method which can be formulated as the minimizer of a penalized quadratic objective function. We successfully accomplish this goal and present the details of our proposal in the remaining part of this
section.

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

\[
(\theta_2^{Bayes}, \ldots, \theta_K^{Bayes}) = \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^\top \Sigma \theta_k - (\mu_k - \mu_1)^\top \theta_k \right\}.
\]

In the classical low-dimension-large-sample-size setting, we can estimate \( (\theta_2^{Bayes}, \ldots, \theta_K^{Bayes}) \) via an empirical version of (2.6)

\[
(\hat{\theta}_2, \ldots, \hat{\theta}_K) = \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^\top \hat{\Sigma} \theta_k - (\hat{\mu}_k - \hat{\mu}_1)^\top \theta_k \right\},
\]

where \( \hat{\Sigma} = \frac{1}{n - K} \sum_{i=1}^{K} \sum_{Y_i = k} (X^i - \hat{\mu}_k)(X^i - \hat{\mu}_k)^\top \), \( \hat{\mu}_k = \frac{1}{n_k} \sum_{Y_i = k} X^i \) and \( n_k \) is the sample size within Class \( k \). The solution to (2.7) gives us the classical multiclass linear discriminant classifier.

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

\[
(\hat{\theta}_2, \ldots, \hat{\theta}_K) = \arg \min_{\theta_2, \ldots, \theta_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \theta_k^\top \hat{\Sigma} \theta_k - (\hat{\mu}_k - \hat{\mu}_1)^\top \theta_k \right\} + \lambda \sum_{j=1}^p \| \theta_{\cdot j} \|,
\]

where \( \lambda \) is a tuning parameter. It is clear that (2.8) is based on (2.7). In (2.8) we have used the group lasso (Yuan and Lin, 2006) to encourage the common sparsity structure. Let \( \mathcal{D} = \{ j : \hat{\theta}_{kj} \neq 0 \} \) 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 \cite{Fan and Li, 2001} or an adaptive group lasso penalty \cite{Bach, 2008} to replace the group lasso penalty in (2.8). To fix the 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.

We would like to make a remark here that our proposal is derived from a different angle than sparse optimal scoring and $\ell_1$ penalized Fisher’s discriminant analysis. Both sparse optimal scoring and $\ell_1$ penalized Fisher’s discriminant analysis penalize a formulation related to Fisher’s discriminant analysis in (2.2), while our method directly estimates the Bayes rule. This different angle leads to considerable convenience in both computation and theoretical studies. Yet we can easily recover the directions defined by Fisher’s discriminant analysis after applying our method. See Section S1 in the supplementary materials for details.

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 (2.8). 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 $K = 2$, (2.8) reduces to

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

(2.9)

Considering the Dantzig selector formulation of (2.9), we have the following constrained $\ell_1$ minimization estimator defined as

$$\hat{\theta} = \arg \min_{\theta} ||\theta||_1 \text{ s.t. } \|\hat{\Sigma}\theta - (\hat{\mu}_2 - \hat{\mu}_1)\|_\infty \leq \lambda.$$  (2.10)

The above estimator is exactly the linear programming discriminant (LPD) (Cai and Liu, 2011).

Moreover, we compare (2.9) with another two well-known sparse discriminant analysis proposals for binary classification: the regularized optimal affine discriminant (ROAD) (Fan et al., 2012) and the direct sparse discriminant analysis (DSDA) (Mai et al., 2012). Denote the estimates of the discriminant directions given by ROAD and DSDA as $\hat{\theta}^{ROAD}$ and
\( \hat{\theta}^{\text{DSDA}} \), respectively. Then we have

\[
\hat{\theta}^{\text{ROAD}}(\lambda) = \arg\min_{\theta} \theta^T \hat{\Sigma} \theta + \lambda \|\theta\|_1 \text{ s.t. } \theta^T (\hat{\mu}_2 - \hat{\mu}_1) = 1 \tag{2.11}
\]

\[
\hat{\theta}^{\text{DSDA}}(\lambda) = \arg\min_{\theta} \sum_i (Y_i - \theta_0 - (X_i^T \theta)^2 + \lambda \|\theta\|_1 \tag{2.12}
\]

We derive the following proposition to reveal the connections between our proposal \((K = 2)\) and ROAD, DSDA. Note that the proofs of this proposition and all the subsequent lemmas and theorems can be found in the appendix.

**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 \(a = \frac{2n|c_1(\lambda)|}{|c_0(\lambda)|}\). Then we have

\[
\hat{\theta}^{\text{MSDA}}(\lambda) = c_0(\lambda) \hat{\theta}^{\text{ROAD}}(2\lambda/|c_0(\lambda)|), \tag{2.13}
\]

\[
\hat{\theta}^{\text{MSDA}}(\lambda) = \frac{c_0(\lambda)}{c_1(a\lambda)} \hat{\theta}^{\text{DSDA}}(a\lambda). \tag{2.14}
\]

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 multiclass generalization of these successful binary sparse LDA methods.

We now present the efficient algorithm for solving (2.8). For convenience write \( \hat{\delta}_k = \hat{\mu}_k - \hat{\mu}_1 \). Our algorithm is based on the following lemma.

**Lemma 1.** Given \( \{\theta_{j'}, j' \neq j\} \), the solution of \( \theta_j \) to (2.8) is defined as

\[
\arg\min_{\theta_j} \sum_{k=2}^{K} \frac{1}{2}(\theta_{kj} - \tilde{\theta}_{kj})^2 + \frac{\lambda}{\tilde{\sigma}_{jj}} \|\theta_j\| \quad (2.15)
\]

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

The solution to (2.15) is given by

\[
\tilde{\theta}_j = \hat{\theta}_j \left(1 - \frac{\lambda}{\|\tilde{\theta}_j\|}\right)_+. \quad (2.16)
\]

Based on Lemma 1, we use the following blockwise-descent algorithm to implement our multiclass 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 \( \tilde{\theta}_k^{(0)} \) 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

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

4. Let \( \hat{\theta}_k \) be the solution at convergence. The output classifier is the usual linear discriminant classifier on \((X^T \hat{\theta}_2, \ldots, X^T \hat{\theta}_K)\).

We have implemented our method in an R package \texttt{msda} which is available on CRAN. Our package also handles the version of (2.8) 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 \( \Omega \in \mathbb{R}^{m \times n} \), \( \|\Omega\|_\infty = \max_j \sum_i |\omega_{ij}| \), \( \|\Omega\|_1 = \sum_j \sum_i |\omega_{ij}| \).
\[
\max_j \sum_i |\omega_{ij}|. \text{ Define }
\varphi = \max \{\|\Sigma_{D^c, D}\|_\infty, \|\Sigma_{D, D}^{-1}\|_\infty\}, \Delta = \max \{\|\mu\|_1, \|\theta_{\text{Bayes}}\|_1\};
\]
\[
\theta_{\text{Bayes}}^\text{min} = \min_{(k,j): \theta_{kj} \neq 0} |\theta_{kj}|, \theta_{\text{Bayes}}^\text{max} = \max_{(k,j)} |\theta_{kj}|;
\]
\[
\|\Sigma_{D^c, D} \Sigma_{D, D}^{-1}\|_\infty = \eta^*.
\]

Let \( d \) be the cardinality of \( D \). Also, for simplicity, we assume that \( \sigma_{jj} \) is uniformly bounded from above.

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:

\[(C0) \max_{j \in D^c} \left\{ \sum_{k=2}^{K} (\Sigma_{j,D} \Sigma_{D, D}^{-1} t_{k,D})^2 \right\}^{1/2} = \kappa < 1.\]

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 (Bach, 2008). We further note that Condition (C0) is satisfied for some most commonly used covariance structures, as shown by the following lemma.

**Lemma 2.** Assume the LDA model. Then we have the following conclusions:

1. If all elements in \( \Sigma_{D^c, D} \) are equal to 0, then Condition (C0) holds.

2. If \( D = \{1, \ldots, d\} \) and \( \Sigma \) has the autoregressive structure, i.e., there exists \(-1 < \rho < 1\) such that \( \sigma_{ij} = \rho^{i-j} \), then Condition (C0) holds.
3. If $\Sigma$ has the compound symmetry structure, i.e., there exists $0 < \rho < 1$ such that $\sigma_{ij} = \rho$ for $i \neq j$ and $\sigma_{jj} = 1$, then Condition (C0) holds.

Lemma 2 exhibits that Condition (C0) is true for some popular covariance structures. It is especially worth noting that when $\Sigma$ has the autoregressive structure or the compound symmetry structure, there is no restriction on $\rho$. In other words, the correlation can be arbitrarily strong. Therefore, Condition (C0) covers a wide range of models.

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

(C1) There exists $c_1, C_1 > 0$ such that $\frac{c_1}{K} \leq \pi_k \leq \frac{C_1}{K}$ for $k = 1, \ldots, K$ and $\frac{\theta_{\text{Bayes max}}}{\theta_{\text{Bayes min}}} < C_1$.

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

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

(C4) $\min_{k,k'} \left\{ (\theta_{k \text{Bayes}} - \theta_{k' \text{Bayes}})^\top \Sigma (\theta_{k \text{Bayes}} - \theta_{k' \text{Bayes}}) \right\}^{1/2}$ is bounded away from 0.

Condition (C1) guarantees that we will have a decent sample size for each class. Moreover, the assumption $\frac{\theta_{\text{Bayes max}}}{\theta_{\text{Bayes min}}} < C_1$ ensures that the set of important predictors is well defined, and no important predictor will
dominate other important ones. If $\frac{\theta_{\text{Bayes}}^{\text{max}}}{\theta_{\text{Bayes}}^{\text{min}}}$ and hence Condition (C1) is violated, then there are predictors with nonzero but relatively very small coefficients. In such cases, these predictors are “close to unimportant” and can be difficult to detect. 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 $d = O(n^{1/2 - \alpha})$, $0 < \alpha \leq 1/2$, 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 $\theta_{\text{Bayes}}^{\text{min}}$ 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. It can be seen that all these conditions are natural and mild. However, we make no claim that they are the weakest conditions possible.

In the following theorems, we let $C$ denote a generic positive constant that can vary from place to place.

**Theorem 1.** 1. Under conditions (C0)–(C1), there exists a generic constant $M$ such that, if $\lambda < \min\{\frac{\theta_{\text{Bayes}}^{\text{min}}}{8\varphi}, M(1 - \kappa)\}$, then with a proba-
bility greater than

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

where \(0 < \epsilon < \min\left\{ \frac{1}{2\phi}, \frac{\lambda}{1 + \phi \Delta} \right\}\), we have that \(\hat{D} = D\), and \(\|\hat{\theta}_k - \theta_k^{\text{Bayes}}\|_\infty \leq 4\varphi \lambda\) for \(k = 2, \ldots, K\).

2. If we further assume conditions (C2)–(C3), we have that if \(\left\{\frac{d^2 \log (pd)}{n}\right\}^{1/2} \ll \lambda \ll \theta_{\text{min}}^{\text{Bayes}}\), then with probability tending to 1, we have \(\hat{D} = D\), and \(\|\hat{\theta}_k - \theta_k^{\text{Bayes}}\|_\infty \to 0\) 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. For a new observation \((X, Y)\) that is not used in constructing the classifier, define

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

where \(\hat{Y}(\hat{\theta}_k, \hat{\pi}_k, k = 1, \ldots, K)\) is the prediction by our method. It can be seen that \(R_n\) is the prediction error of our estimated classifier. Also define \(R\) as the Bayes error. Then we have the following conclusions.

**Theorem 2.** 1. Under conditions (C0)–(C1), there exists a generic constant \(M_1\) such that, if \(\lambda < \min\left\{ \frac{\theta_{\text{min}}^{\text{Bayes}}}{8\varphi}, M_1(1 - \kappa) \right\}\), then with a prob-
ability greater than

\[
1 - Cpd \exp(-Cn \frac{\epsilon^2}{Kd^2}) - CK \exp(-C \frac{n}{K^2}) - Cp(K - 1) \exp(-Cn \frac{\epsilon^2}{K})
\]

(3.2)

where \(0 < \epsilon < \min\{\frac{1}{2\varphi}, \frac{\lambda}{1 + \varphi\Delta}\}\), we have

\[
|R_n - R| \leq M_1 \lambda^{1/3},
\]

(3.3)

for some generic constant \(M_1\),.

2. Under conditions (C0)–(C4), if \(\lambda \to 0\), then with probability tending to 1, we have

\[
R_n \to R.
\]

Remark 1. 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 and a testing set of size 1000 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 $\mu_k = \Sigma \beta_k$. For simplicity, we say that a matrix $\Sigma$ has the AR($\rho$) structure if $\sigma_{jk} = \rho^{|j-k|}$ for $j, k = 1, \ldots, p$; on the other hand, $\Sigma$ has the CS($\rho$) structure if $\sigma_{jk} = \rho$ for any $j \neq k$ and $\sigma_{jj} = 1$ for $j = 1, \ldots, p$.

Model 1:

$K = 4$, $\beta_{jk} = 1.6$ for $j = 2k - 1, 2k; k = 1, \ldots, K$ and $\beta_{jk} = 0$ otherwise. The covariance matrix $\Sigma$ has the AR(0.5) structure.

Model 2:

$K = 6$, $\beta_{jk} = 2.5$ for $j = 2k - 1, 2k; k = 1, \ldots, K$ and $\beta_{jk} = 0$ otherwise. The covariance matrix $\Sigma = I_5 \otimes \Omega$, where $\Omega$ has the CS(0.5) structure.

Model 3:

$K = 4$, $\beta_{jk} = k + u_{jk}$ for $j = 1, \ldots, K$, where $u_{jk}$ follows the uniform distribution over the interval $[-1/4, 1/4]$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\Sigma$ has the CS(0.5) structure.

Model 4:

$K = 4$, $\beta_{jk} = k + u_{jk}$ for $j = 1, \ldots, 4$, where $u_{jk}$ follows the uniform
distribution over the interval $[-1/4, 1/4]$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\Sigma$ has the CS(0.8) structure.

Model 5:

$K = 4$, $\beta_{2,1} = \ldots = \beta_{2,8} = 1.2$, $\beta_{3,1} = \ldots = \beta_{3,4} = -1.2$, $\beta_{3,5} = \ldots = \beta_{3,8} = 1.2$, $\beta_{4,2j-1} = -1.2$, $\beta_{4,2j} = 1.2$ for $j = 1, \ldots, 4$; $\beta_{jk} = 0$ otherwise.

The covariance matrix $\Sigma$ has the AR(0.5) structure.

Model 6:

$K = 4$, $\beta_{2,1} = \ldots = \beta_{2,8} = 1.2$, $\beta_{3,1} = \ldots = \beta_{3,4} = -1.2$, $\beta_{3,5} = \ldots = \beta_{3,8} = 1.2$, $\beta_{4,2j-1} = -1.2$, $\beta_{4,2j} = 1.2$ for $j = 1, \ldots, 4$; $\beta_{jk} = 0$ otherwise.

The covariance matrix $\Sigma$ has the AR(0.8) structure.

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 observations from Table 1. First, our method is the best across all six 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 \cite{Bach2008}. Because the other two methods do not use the adaptive lasso penalty, we do not include the results of our method using.
Table 1: Simulation results for Models 1–6. 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.

<table>
<thead>
<tr>
<th></th>
<th>Bayes</th>
<th>Our</th>
<th>Witten</th>
<th>Clemmensen</th>
<th>Bayes</th>
<th>Our</th>
<th>Witten</th>
<th>Clemmensen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error(%)</td>
<td>11.0</td>
<td>12.4</td>
<td>15.5</td>
<td>13</td>
<td>13.3</td>
<td>15.2</td>
<td>31.7</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.20)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>$C$</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>IC</td>
<td>0</td>
<td>10</td>
<td>126</td>
<td>5</td>
<td>0</td>
<td>15</td>
<td>19.5</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(4.9)</td>
<td>(4.4)</td>
<td>(0)</td>
<td>(0.7)</td>
<td>(1.5)</td>
<td>(0.3)</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error(%)</td>
<td>8.8</td>
<td>9.4</td>
<td>14.1</td>
<td>12.7</td>
<td>5.3</td>
<td>5.7</td>
<td>7</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.09)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>$C$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>IC</td>
<td>0</td>
<td>3</td>
<td>796</td>
<td>30</td>
<td>0</td>
<td>4</td>
<td>796</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>(0.4)</td>
<td>(0)</td>
<td>(0.2)</td>
<td>(0)</td>
<td>(0.5)</td>
<td>(0)</td>
<td>(2.2)</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error(%)</td>
<td>8.3</td>
<td>9.5</td>
<td>17.9</td>
<td>13.6</td>
<td>14.2</td>
<td>17.4</td>
<td>23.4</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.14)</td>
<td>(0.09)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>$C$</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.0)</td>
<td>(0)</td>
<td>(0.1)</td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>0</td>
<td>6</td>
<td>97</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(0.9)</td>
<td>(2.8)</td>
<td>(0.5)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.5)</td>
<td>(0.3)</td>
<td></td>
</tr>
</tbody>
</table>

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 (Burczynski et al., 2006). This dataset contains 22283 gene expres-
sion levels from 127 people. These 127 people are either normal people, people 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 (Fan and Fan, 2008) 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

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

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 (Fan and Fan, 2008; Fan and Song, 2010; Mai and Zou, 2013a), $d_n$ 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
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.

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

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. This again shows that our method is a very competitive classifier.

**Supplementary Materials**

All the proofs are available in the supplementary materials. Section S1 contains the connection between our method and Fisher's discriminant analysis. Section S2 contains all the other proofs.
Acknowledgements

The authors thank the editor, associate editor and referees for their helpful comments and suggestions. Zou’s research is partially supported by NSF grant DMS-1505111. Mai’s research is partly supported by CIF-1617691, National Science Foundation.

References


REFERENCES


REFERENCES


*J. R. Statist. Soc. B* 73, 753–772.

ysis for simultaneous testing for the significance of a gene set/pathway and gene selection’, 

*Bioinformatics* 25, 1145–1151.

Yuan, M. and Lin, Y. (2006), ‘Model selection and estimation in regression with grouped vari-

Qing Mai, Department of Statistics, Florida State University

E-mail: mai@stat.fsu.edu

Yi Yang, Department of Mathematics and Statistics, McGill University

E-mail: yi.yang6@mcgill.ca

Hui Zou, School of Statistics, University of Minnesota

E-mail: zouxx019@umn.edu