Statistica Sinica **29** (2019), 97-111
doi:https://doi.org/10.5705/ss.202016.0117

MULTICLASS SPARSE DISCRIMINANT ANALYSIS

Qing Mai, Yi Yang and Hui Zou

Florida State University, McGill University and University of Minnesota

Abstract: In recent years several sparse linear discriminant analysis methods have been proposed for high-dimensional classification and variable selection. Most of these proposals focus on binary classification and are not directly applicable to multiclass classification problems. Some sparse discriminant analysis methods 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 attained by two successful binary sparse linear discriminant analysis methods, providing a unification of these seemingly unrelated proposals. Our method can be solved by an efficient algorithm that 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 finding rates of convergence under the ultrahigh dimensionality setting. We further demonstrate the empirical performance of our method with simulations and data.

Key words and phrases: Discriminant analysis, high dimensional data, multiclass classification, rates of convergence, variable selection.

1. Introduction

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

$$\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), k \in \{1, 2, \ldots, K\}. \tag{1.1}$$

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

$$\hat{Y} = \arg\max_k \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_k}{2} \right)^{\mathrm{T}} \boldsymbol{\beta}_k + \log \pi_k \right\}, \tag{1.2}$$

where $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$ for $k = 1, \ldots, K$. Linear discriminant analysis has been observed to perform very well on many low-dimensional datasets (Michie, Spiegelhalter and Taylor (1994); Hand (2006)). It may not be suitable for high-dimensional datasets for at least two reasons. It cannot be applied if the dimension

p exceeds the sample size n, while Bickel and Levina (2004) and Fan and Fan (2008) have shown that, even if the true covariance matrix is an identity matrix and we know this fact, a classifier involving all the predictors is 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 Fan (2008); Wu et al. (2008); Shao et al. (2011); Cai and Liu (2011); Witten and Tibshirani (2011); Mai, Zou and Yuan (2012); Fan, Feng and Tong (2012)). In the binary case, the discriminant direction is $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. One can seek sparse estimates of $\boldsymbol{\beta}$ to generalize linear discriminant analysis to deal with high-dimensional classification. 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, Feng and Tong (2012)), and the direct sparse discriminant analysis (Mai, Zou and Yuan (2012)). The linear programming discriminant finds a sparse estimate by the Dantzig selector (Candes and Tao (2007)), the regularized optimal affine discriminant (Fan, Feng and Tong (2012)) adds the lasso penalty (Tibshirani (1996)) to Fisher's discriminant analysis, and the direct sparse discriminant analysis (Mai, Zou and Yuan (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 in the presence of ultrahigh dimensions. However, they are explicitly designed for binary classification and do not handle the multiclass case naturally.

A referee has suggested breaking the K-class problem into $K(K-1)/2$ pairwise problems, applying a binary classifier to each, and classifying according to majority vote. Tie votes complicate such an approach to the problem.

Two popular multiclass sparse discriminant analysis proposals are the ℓ_1 penalized Fisher's discriminant (Witten and Tibshirani (2011)) and sparse optimal scoring (Clemmensen et al. (2011)). These methods do not have theoretical justifications in place.

We seek a new multiclass sparse discriminant analysis algorithm that is conceptually intuitive, computationally efficient, and theoretically sound. We show that our proposal has competitive empirical performance and enjoys strong theoretical properties under ultrahigh dimensionality. In Section 2 we introduce the details of our proposal after briefly reviewing two existing proposals, and we develop an efficient algorithm for our method. Theoretical results are given in

Section 3. In Section 4 we use simulations and a data example to demonstrate the superior performance of our method over sparse optimal scoring and ℓ_1 penalized Fisher's discriminant. Proofs are in the supplementary materials.

2. Method

2.1. Existing proposals

The Bayes rule under a linear discriminant analysis model is

$$\hat{Y} = \arg\max_k \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_k}{2} \right)^{\mathrm{T}} \boldsymbol{\beta}_k + \log \pi_k \right\},$$

where $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$ for $k = 1, \ldots, K$. If $\boldsymbol{\theta}_k^{\mathrm{Bayes}} = \boldsymbol{\beta}_k - \boldsymbol{\beta}_1$ for $k = 1, \ldots, K$, the Bayes rule can be written as

$$\hat{Y} = \arg\max_k \left\{ (\boldsymbol{\theta}_k^{\mathrm{Bayes}})^{\mathrm{T}} \left(\mathbf{X} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k}{2} \right) + \log \frac{\pi_k}{\pi_1} \right\}. \tag{2.1}$$

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

Instead of estimating $\boldsymbol{\theta}^{\mathrm{Bayes}}$ directly, sparse optimal scoring and ℓ_1 penalized Fisher's discriminant estimate a set of directions $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1}) \in \mathbb{R}^{p\times(K-1)}$ such that $\boldsymbol{\eta}$ spans the same linear subspace as $\boldsymbol{\theta}^{\mathrm{Bayes}}$, and hence linear discriminant analysis on $\mathbf{X}^{\mathrm{T}}\boldsymbol{\eta}$ is equivalent to (2.1) on the population level. The methods look for estimates of $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1})$ in Fisher's discriminant analysis:

$$\boldsymbol{\eta}_k = \arg\max \boldsymbol{\eta}_k^{\mathrm{T}} \boldsymbol{\Sigma}_b \boldsymbol{\eta}_k, \quad \text{s.t. } \boldsymbol{\eta}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_k = 1, \boldsymbol{\eta}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_l = 0 \text{ for } l < k, \tag{2.2}$$

where $\boldsymbol{\Sigma}_b = \{1/(K-1)\} \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^{\mathrm{T}}$ with $\bar{\boldsymbol{\mu}} = 1/K \sum_k \boldsymbol{\mu}_k$.

We refer to $\boldsymbol{\eta}$ as discriminant directions as well. To find $\boldsymbol{\eta}$, take $\mathbf{Y}^{\mathrm{dm}}$ as an $n \times K$ matrix of dummy variables with $Y_{ik}^{\mathrm{dm}} = 1(Y_i = k)$.

Sparse optimal scoring creates $K-1$ vectors of scores $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{K-1} \in \mathbb{R}^K$. Then for $k = 1, \ldots, K-1$, sparse optimal scoring sequentially determines $\boldsymbol{\eta}_k$. Given $\hat{\boldsymbol{\alpha}}_l$ and discriminant directions $\hat{\boldsymbol{\eta}}_l^{\mathrm{SOS}}$, $l < k$, sparse optimal scoring finds $\hat{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{\eta}}_k^{\mathrm{SOS}}$ by solving

$$(\hat{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{\eta}}_k^{\mathrm{SOS}}) = \arg\min_{\boldsymbol{\alpha}_k, \boldsymbol{\eta}_k} \sum_{i=1}^{n} (\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\alpha}_k - \tilde{\mathbf{X}}\boldsymbol{\eta}_k)^2 + \lambda \|\boldsymbol{\eta}_k\|_1 \tag{2.3}$$

$$\text{s.t. } \frac{1}{n}\boldsymbol{\alpha}_k^{\mathrm{T}} (\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}} \boldsymbol{\alpha}_k = 1, \boldsymbol{\alpha}_k^{\mathrm{T}} (\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}} \hat{\boldsymbol{\alpha}}_l = 0, \text{ for any } l < k,$$

where $\tilde{\mathbf{X}}$ is the centered data matrix, and λ is a tuning parameter. Sparse optimal scoring is closely related to (2.2) because, when the dimension is low, the unpenalized version of (2.3) gives the same directions (up to a scalar) as (2.2) with

the parameters $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}$ substituted with their sample estimates. Therefore, with the ℓ_1 penalty, sparse optimal scoring gives sparse approximations to $\boldsymbol{\eta}$.

The ℓ_1 penalized Fisher's discriminant analysis estimates $\boldsymbol{\eta}_k$ by

$$\hat{\boldsymbol{\eta}}_k = \arg\max_{\boldsymbol{\eta}_k} \boldsymbol{\eta}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_b^k \boldsymbol{\eta}_k + \lambda_k \sum_j |\hat{\sigma}_j \eta_{kj}| \text{ s.t. } \boldsymbol{\eta}_k^{\mathrm{T}} \tilde{\mathbf{\Sigma}} \boldsymbol{\eta}_k \le 1,$$

for $k = 1, \ldots, K-1$, where λ_k are tuning parameters, $\hat{\sigma}_j^2$ is the (j,j)th element of the sample estimate of $\mathbf{\Sigma}$, $\tilde{\mathbf{\Sigma}}$ is a positive definite estimate of $\mathbf{\Sigma}$,

$$\hat{\mathbf{\Sigma}}_b^k = \mathbf{X}^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}} \{(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}}\}^{-1/2} \mathbf{\Omega}_k \{(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}}\}^{-1/2} (\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{X}, \tag{2.4}$$

and $\mathbf{\Omega}_k$ is the identity matrix if $k = 1$, otherwise an orthogonal projection matrix with column space orthogonal to $\{(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}} \mathbf{Y}\}^{-1/2} \mathbf{Y}^{\mathrm{T}} \mathbf{X} \hat{\boldsymbol{\eta}}_l$ for all $l < k$. Again, if the dimension is low, the unpenalized version of (2.4) is equivalent to (2.2) with the parameters replaced by the sample estimates. Since $\mathbf{\Omega}_k$ relies on $\hat{\boldsymbol{\eta}}_l$ for all $l < k$, the ℓ_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 ℓ_1 penalized Fisher's discriminant analysis and sparse optimal scoring, but it is not known whether these classifiers are consistent when more than two classes are present. While these methods estimate the discriminant directions sequentially, we believe a better multiclass sparse discriminant analysis algorithm would estimate all discriminant directions simultaneously, as in classical linear discriminant analysis. We develop a computationally efficient multiclass sparse discriminant analysis method that enjoys strong theoretical properties under ultrahigh dimensionality. It can be viewed as a natural multiclass counterpart of the three binary sparse discriminant methods in Mai, Zou and Yuan (2012), Cai and Liu (2011), and Fan, Feng and Tong (2012).

The implication of sparsity in the multiclass problem, explained in Mai, Zou and Yuan (2012), is that the right target for variable selection should be the subset of variables that influences the Bayes rule. By (2.1), the contribution from the jth variable (X_j) vanishes if and only if

$$\theta_{2j}^{\mathrm{Bayes}} = \cdots = \theta_{Kj}^{\mathrm{Bayes}} = 0. \tag{2.5}$$

Let $\mathcal{D} = \{j : (2.5) \text{ does not hold}\}$. Here whether an index j belongs to $\mathcal{D}$ depends on θ_{kj} for all k, since $\theta_{kj}^{\mathrm{Bayes}}, k = 2, \ldots, K$ are related to each other, being coefficients for the same predictor. Thus, $\theta_{kj}^{\mathrm{Bayes}}, k = 2, \ldots, K$ are naturally

grouped according to j, and successful multiclass sparse LDA method should correctly identify $\mathcal{D}$, at least in theory.

In sequential procedures, directions are estimated one by one, and it is less likely to estimate all the coefficients of one predictor to be zero. Hence, sequential methods do not utilize all the available information and are prone to loss of accuracy.

Mai, Zou and Yuan (2012) take advantage of a close link between the LDA and the ordinary least squares, so that one can use any software for solving sparse penalized linear regression to fit the sparse LDA classifier they proposed. However, such a connection only holds for the binary case. We observe that, theoretically speaking, the binary sparse LDA proposal in Mai, Zou and Yuan (2012) is equivalent to a sparse penalized quadratic criterion, and, computationally speaking, a penalized quadratic problem is as efficient as penalized least squares. Thus, we develop a multiclass sparse LDA method that can be formulated as the minimizer of a penalized quadratic objective function. This idea was also pursued in an independent work (Gaynanova, Booth and Wells (2016)).

Our proposal begins with a convex optimization formulation of the Bayes rule of the multiclass linear discriminant analysis model. With $\boldsymbol{\theta}_k^{\mathrm{Bayes}} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$ for $k = 2, \ldots, K$, on the population level,

$$(\boldsymbol{\theta}_2^{\mathrm{Bayes}}, \ldots, \boldsymbol{\theta}_K^{\mathrm{Bayes}}) = \arg\min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\theta}_k - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\theta}_k \right\}. \quad (2.6)$$

In the classical low-dimension-large-sample-size setting, we can estimate $(\boldsymbol{\theta}_2^{\mathrm{Bayes}}, \ldots, \boldsymbol{\theta}_K^{\mathrm{Bayes}})$ via an empirical version of (2.6)

$$(\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_K) = \arg\min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta}_k - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}} \boldsymbol{\theta}_k \right\}, \quad (2.7)$$

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

Write $\boldsymbol{\theta}_{\cdot j} = (\theta_{2j}, \ldots, \theta_{Kj})^{\mathrm{T}}$ and define $\|\boldsymbol{\theta}_{\cdot j}\| = (\sum_{i=2}^{K} \theta_{ij}^2)^{1/2}$. For the high-dimensional case, we propose a penalized formulation for multiclass sparse discriminant analysis,

$$(\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_K) = \arg\min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta}_k - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}} \boldsymbol{\theta}_k \right\} + \lambda \sum_{j=1}^{p} \|\boldsymbol{\theta}_{\cdot j}\|, \quad (2.8)$$

where λ 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 $\hat{\mathcal{D}} = \{j : \hat{\theta}_{kj} \neq 0\}$ denote the set of selected variables for the multiclass classification problem. We will show that with a high probability $\hat{\mathcal{D}}$ equals $\mathcal{D}$. One can also use a group version of a nonconvex penalty (Fan and Li (2001)) or an adaptive group lasso penalty (Bach (2008)) to replace the group lasso penalty in (2.8). We do not pursue this here.

After obtaining $\hat{\boldsymbol{\theta}}_k, k = 2, \ldots, K$, we fit the classical multiclass linear discriminant analysis on $(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_2, \ldots, \mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_K)$, as in sparse optimal scoring and ℓ_1 penalized Fisher's discriminant analysis. We repeat the procedure for a sequence of λ values and pick the one with the smallest cross-validation error rate.

While sparse optimal scoring and ℓ_1 penalized Fisher's discriminant analysis penalize a formulation related to Fisher's discriminant analysis in (2.2), our method directly estimates the Bayes rule. This leads to considerable convenience in both computational 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 $K = 2$ at (2.8). It turns out that the binary case of our proposal has connections with some binary sparse LDA methods in the literature. We elaborate more on this point.

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

$$\hat{\boldsymbol{\theta}}^{\mathrm{MSDA}}(\lambda) = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1 \right\}. \tag{2.9}$$

Considering the Dantzig selector formulation of (2.9), we have a constrained ℓ_1 minimization estimator,

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

This estimator is the linear programming discriminant (LPD) (Cai and Liu (2011)).

We compare (2.9) with two more sparse discriminant analysis proposals for binary classification: the regularized optimal affine discriminant (ROAD)(Fan, Feng and Tong (2012)) and the direct sparse discriminant analysis (DSDA) (Mai, Zou and Yuan (2012)). Denote the estimates of the discriminant directions given

by ROAD and DSDA as $\hat{\boldsymbol{\theta}}^{\mathrm{ROAD}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{DSDA}}$, respectively. Then we have

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

$$\hat{\boldsymbol{\theta}}^{\mathrm{DSDA}}(\lambda) = \arg\min_{\boldsymbol{\theta}} \sum_i \{Y^i - \theta_0 - (\mathbf{X}^i)^{\mathrm{T}}\boldsymbol{\theta}\}^2 + \lambda\|\boldsymbol{\theta}\|_1. \tag{2.12}$$

We can show the connections between our proposal, $K = 2$, and ROAD and DSDA. The proofs of this proposition and subsequent lemmas and theorems can be found in the appendix.

Proposition 1. *If* $c_0(\lambda) = \hat{\boldsymbol{\theta}}^{\mathrm{MSDA}}(\lambda)^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1), c_1(\lambda) = \hat{\boldsymbol{\theta}}^{\mathrm{DSDA}}(\lambda)^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, *and* $a = \{2n|c_1(\lambda)|\}/(|c_0(\lambda)|)$, *then we have*

$$\hat{\boldsymbol{\theta}}^{\mathrm{MSDA}}(\lambda) = c_0(\lambda)\hat{\boldsymbol{\theta}}^{\mathrm{ROAD}}\left(\frac{2\lambda}{|c_0(\lambda)|}\right), \tag{2.13}$$

$$\hat{\boldsymbol{\theta}}^{\mathrm{MSDA}}(\lambda) = \frac{c_0(\lambda)}{c_1(a\lambda)}\hat{\boldsymbol{\theta}}^{\mathrm{DSDA}}(a\lambda). \tag{2.14}$$

Proposition 1 shows that the classification direction by our proposal is identical to a classification direction by ROAD and a classification direction by DSDA.

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 produce an efficient algorithm for our proposed multiclass procedure. It is then a natural generalization of these binary sparse LDA methods.

In solving (2.8), write $\hat{\boldsymbol{\delta}}^k = \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1$. Our algorithm is based on the following.

Lemma 1. *Given* $\{\boldsymbol{\theta}_{\cdot j'}, j' \neq j\}$, *the solution of* $\boldsymbol{\theta}_{\cdot j}$ *to* (2.8) *is*

$$\arg\min_{\boldsymbol{\theta}_{\cdot j}} \sum_{k=2}^{K} \frac{1}{2}(\theta_{kj} - \tilde{\theta}_{kj})^2 + \frac{\lambda}{\hat{\sigma}_{jj}}\|\boldsymbol{\theta}_{\cdot j}\|, \tag{2.15}$$

where $\tilde{\theta}_{k,j} = (\hat{\delta}_j^k - \sum_{l\neq j}\hat{\sigma}_{lj}\theta_{kl})/\hat{\sigma}_{jj}$. *If* $\tilde{\boldsymbol{\theta}}_{\cdot j} = (\tilde{\theta}_{2j}, \ldots, \tilde{\theta}_{Kj})^{\mathrm{T}}$ *and* $\|\tilde{\boldsymbol{\theta}}_{\cdot j}\| = (\sum_{k=2}^{K}\tilde{\theta}_{kj}^2)^{1/2}$, *the solution to* (2.15) *is given by*

$$\hat{\boldsymbol{\theta}}_{\cdot j} = \tilde{\boldsymbol{\theta}}_{\cdot j}\left(1 - \frac{\lambda}{\|\tilde{\boldsymbol{\theta}}_{\cdot j}\|}\right)_+. \tag{2.16}$$

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

1. Compute $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}^k$, $k = 1, 2, \ldots, K$.

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

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

   (a) compute

$$\hat{\boldsymbol{\theta}}_{\cdot j}^{(m)} = \tilde{\boldsymbol{\theta}}_{\cdot j}^{(m-1)} \left(1 - \frac{\lambda}{\|\tilde{\boldsymbol{\theta}}_{\cdot j}^{(m-1)}\|}\right)_+;$$

   (b) update

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

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

We have implemented our method in an R package `msda` which is available on CRAN. Our package also handles the version of (2.8) using an adaptive group lasso penalty, because Lemma 1 and Algorithm 1 can be easily generalized to handle the adaptive group lasso penalty.

3. Theory

In this section we study the 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 $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_\infty = \max_j |\alpha_j|$, $\|\boldsymbol{\alpha}\|_1 = \sum_j |\alpha_j|$, while for a matrix $\boldsymbol{\Omega} \in \mathbb{R}^{m \times n}$, $\|\boldsymbol{\Omega}\|_\infty = \max_i \sum_j |\omega_{ij}|$, $\|\boldsymbol{\Omega}\|_1 = \max_j \sum_i |\omega_{ij}|$. Let

$$\begin{aligned} \varphi &= \max\{\|\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty, \|\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\|_\infty\}, \Delta = \max\{\|\boldsymbol{\mu}\|_1, \|\boldsymbol{\theta}^{\text{Bayes}}\|_1\}; \\ \theta_{\min}^{\text{Bayes}} &= \min_{(k,j):\theta_{kj} \neq 0} |\theta_{kj}|, \theta_{\max}^{\text{Bayes}} = \max_{(k,j)} |\theta_{kj}|; \\ \|\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\|_\infty &= \eta^*. \end{aligned}$$

Let d be the cardinality of $\mathcal{D}$. For simplicity, we assume that σ_{jj} is uniformly bounded from above.

If $\mathbf{t}_\mathcal{D} \in \mathbb{R}^{d \times (K-1)}$ is the subgradient of the group lasso penalty at the true $\boldsymbol{\theta}_\mathcal{D}$, we assume the following:

(C0) $\max_{j \in \mathcal{D}^c} \{\sum_{k=2}^K (\boldsymbol{\Sigma}_{j,\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1} \mathbf{t}_{k,\mathcal{D}})^2\}^{1/2} = \kappa < 1.$

A condition similar to (C0) has been used to study the group lasso penalized regression model (Bach (2008)). It is satisfied for many commonly used covariance structures, as shown by the following.

Lemma 2. *If the LDA model holds, then* (C0) *holds if all elements in* $\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}^c}$ *are equal to* 0*, if* $\mathcal{D} = \{1, \ldots, d\}$ *and* $\mathbf{\Sigma}$ *has an autoregressive structure, or if* $\mathbf{\Sigma}$ *has compound symmetry.*

With φ, Δ, η^* and κ fixed, we will use the following regularity conditions.

(C1) There exists $c_1, C_1 > 0$ such that $(c_1/K) \le \pi_k \le (C_1/K)$ for $k = 1, \ldots, K$ and $\left(\theta^{\text{Bayes}}_{\max}/\theta^{\text{Bayes}}_{\min}\right) < C_1$.

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

(C3) $\theta^{\text{Bayes}}_{\min} \gg \{(d^2 \log(pd))/n\}^{1/2}$;

(C4) $\min_{k,k'}\{(\boldsymbol{\theta}^{\text{Bayes}}_k - \boldsymbol{\theta}^{\text{Bayes}}_{k'})^{\mathrm{T}}\mathbf{\Sigma}(\boldsymbol{\theta}^{\text{Bayes}}_k - \boldsymbol{\theta}^{\text{Bayes}}_{k'})\}^{1/2}$ is bounded away from 0.

Condition (C1) guarantees that we will have a decent sample size for each class. The assumption $\theta^{\text{Bayes}}_{\max}/\theta^{\text{Bayes}}_{\min} < C_1$ ensures that the set of important predictors is well defined, and that no important predictor dominates others. If Condition (C1) is violated, there are predictors with nonzero but relatively small coefficients; these predictors are "close to unimportant" and can be difficult to detect. Condition (C2) requires that p not grow too fast with respect to n. It is very mild, as p can grow at a nonpolynomial rate of n. In particular, if $d = O(n^{1/2-\alpha}), 0 < \alpha \le 1/2$, (C2) is satisfied if $\log p = o(n^{2\alpha})$. Condition (C3) guarantees that the nonzero coefficients are bounded away from 0, a common assumption in the literature. The lower bound of $\theta^{\text{Bayes}}_{\min}$ tends to 0 under (C3). Condition (C4) is required so that all the classes can be separated from each other; if it is violated, even the Bayes rule cannot work well. We make no claim that these are the weakest possible conditions.

In the following, C denotes a generic positive constant that can vary from place to place.

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

$$1 - Cpd\exp\left(-Cn\frac{\epsilon^2}{Kd^2}\right) - CK\exp\left(-C\frac{n}{K^2}\right) - Cp(K-1)\exp\left(-Cn\frac{\epsilon^2}{d^2K}\right) \tag{3.1}$$

with $0 < \epsilon < \min\{1/2\varphi, \lambda/(1+\varphi\Delta)\}$*, we have that* $\hat{\mathcal{D}} = \mathcal{D}$*, and* $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^{\text{Bayes}}\|_\infty \leq 4\varphi\lambda$ *for* $k = 2, \ldots, K$*. If we further assume conditions* (C2)–(C3)*, we have that if* $\{(d^2 \log(pd))/n\}^{1/2} \ll \lambda \ll \theta_{\min}^{\text{Bayes}}$*, then with probability tending to* 1*, we have* $\hat{\mathcal{D}} = \mathcal{D}$*, and* $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^{\text{Bayes}}\|_\infty \to 0$ *for* $k = 2, \ldots, K$*.*

We now show that our proposal is a consistent estimator of the Bayes rule in terms of the misclassification error rate. For a new observation $(\mathbf{X}, Y)$, not used in constructing the classifier, let

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

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

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

$$1 - Cpd\exp\left(-Cn\frac{\epsilon^2}{Kd^2}\right) - CK\exp\left(-C\frac{n}{K^2}\right) - Cp(K-1)\exp\left(-Cn\frac{\epsilon^2}{K}\right) \quad (3.2)$$

with $0 < \epsilon < \min\{1/2\varphi, \lambda/(1+\varphi\Delta)\}$*, we have*

$$|R_n - R| \leq M_1\lambda^{1/3}, \quad (3.3)$$

for some generic constant M_1*. 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 derive the asymptotic results by letting K (the number of classes) diverge with n to infinity. This requires more cumbersome notion and bounds, but the analysis remains largely the same. For a clearer picture of the theory, we have focused on the fixed K case.

4. Numerical Studies

4.1. Simulations

We have investigated our proposal by simulation. For comparison, we have included the sparse optimal scoring and ℓ_1 penalized Fisher's discriminant analysis in the simulation study. Four simulation models were considered where the dimension $p = 800$ and the training set has a sample size $n = 75K$, K the number of classes in each model. We generated a validation set of size n to select the tuning parameters and a testing set of size 1,000 for each method. We specified $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}$ as in the following, then let $\boldsymbol{\mu}_k = \boldsymbol{\Sigma}\boldsymbol{\beta}_k$. We say that a matrix $\boldsymbol{\Sigma}$ has the AR(ρ) structure if $\sigma_{jk} = \rho^{|j-k|}$ for $j, k = 1, \ldots, p$, and that $\boldsymbol{\Sigma}$ has the CS(ρ)

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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma} = \mathbf{I}_5 \otimes \mathbf{\Omega}$, where $\mathbf{\Omega}$ has the CS(0.5) structure.
Model 3: $K = 4$, $\beta_{jk} = k + u_{jk}$ for $j = 1, \ldots, K$, where u_{jk} is uniform over the interval $[-1/4, 1/4]$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\mathbf{\Sigma}$ has the CS(0.5) structure.
Model 4: $K = 4$, $\beta_{jk} = k + u_{jk}$ for $j = 1, \ldots, 4$, and u_{jk} is uniform distribution over $[-1/4, 1/4]$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\mathbf{\Sigma}$ has the CS(0.8) structure.
Model 5: $K = 4$, $\beta_{2,1} = \cdots = \beta_{2,8} = 1.2$, $\beta_{3,1} = \cdots = \beta_{3,4} = -1.2$, $\beta_{3,5} = \cdots = \beta_{3,8} = 1.2$, $\beta_{4,2j-1} = -1.2, \beta_{4,2j} = 1.2$ for $j = 1, \ldots, 4$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\mathbf{\Sigma}$ has the AR(0.5) structure.
Model 6: $K = 4$, $\beta_{2,1} = \cdots = \beta_{2,8} = 1.2$, $\beta_{3,1} = \cdots = \beta_{3,4} = -1.2$, $\beta_{3,5} = \cdots = \beta_{3,8} = 1.2$, $\beta_{4,2j-1} = -1.2, \beta_{4,2j} = 1.2$ for $j = 1, \ldots, 4$; $\beta_{jk} = 0$ otherwise. The covariance matrix $\mathbf{\Sigma}$ has the AR(0.8) structure.

The error rates of 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. Our method shows the best across all six models, and it is a very good approximation of the Bayes rule in terms of sparsity and misclassification error rate. Although our method tends to select a few more variables aside from the true ones, this can be improved by using the adaptive group lasso penalty (Bach (2008)). Because the other two methods do not use the adaptive lasso penalty, we do not include these results.

4.2. A data example

We have demonstrated the application of our method on the IBD dataset (Burczynski et al. (2006)). This dataset contains 22,283 gene expression levels from 127 people. These people are either normal, have Crohn's disease, or have ulcerative colitis. The 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 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

Table 1. Simulation results for Models 1–6. The two competing methods are denoted by the first author of the original papers: Witten's method is the ℓ_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. Here C is the number of correctly selected variables, and IC is the number of incorrectly selected variables.

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

F-test screening. We computed the F-test statistic for each X_j,

$$f_j = \frac{\sum_{k=1}^{K} n_k(\hat{\mu}_{kj} - \hat{\bar{\mu}}_j)^2/(K-1)}{\sum_{i=1}^{n}(X_j^i - \hat{\mu}_{Y^i,j})^2/(n-K)},$$

where $\hat{\bar{\mu}}_j$ is the sample grand mean for X_j and n_g is the within-group sample size. Based on the F-test statistic, our screening kept only the predictors with F-test statistics among the d_nth largest. As widely recommended (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. We let $d_n = 127$ for the current dataset.

We estimated the rules given by sparse optimal scoring, ℓ_1 penalized Fisher's discriminant analysis and our proposal on the training set. The tuning parameters were chosen by 5-fold cross validation. We evaluated the classification errors

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 ℓ_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.

	Our	Witten	Clemmensen
Error (%)	7.32(0.972)	21.95(1.10)	9.76(0.622)
Fitted Model Size	25 (0.7)	127 (0)	27 (0.5)

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.

Supplementary Materials

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

Acknowledgment

The authors thank the Editor, an 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

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9**, 1179–1225.

Bickel, P. J. and 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**, 989–1010.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A. M., Cotreau, M. M. and Dorner, A. J. (2006). Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J. Mol. Diagn.* **8**, 51–61.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106**, 1566–1577.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics* **53**, 406–413.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **105**, 14790–14795.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–2637.

Fan, J., Feng, Y. and Tong, X. (2012), A ROAD to classification in high dimensional space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74**, 745–771.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–3604.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd Edition, Academic Press Professional, Inc.

Gaynanova, I., Booth, J. and Wells, M. (2016). Simultaneous Sparse Estimation of Canonical Vectors in the p>>N Setting. *JASA* **111**(514), 696–706.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Stat. Sci.* **21**, 1–14.

Mai, Q. and Zou, H. (2013a). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234.

Mai, Q. and Zou, H. (2013b). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics* **55**, 243–246.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99**, 29–42.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press.

Michie, D., Spiegelhalter, D. and Taylor, C. (1994). *Machine Learning, Neural and Statistical Classification*, first Edition, Ellis Horwood.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis with high dimensional data. *Ann. Statist.*

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58**, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572.

Trendafilov, N. T. and Jolliffe, I. T. (2007). DALASS: Variable selection in discriminant analysis via the lasso. *Comput. Statist. Data. Anal.* **51**, 3718–3736.

Witten, D. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73**, 753–772.

Wu, M., Zhang, L., Wang, Z., Christiani, D. and Lin, X. (2008). Sparse linear discriminant analysis 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 variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68**, 49–67.

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA.
E-mail: mai@stat.fsu.edu
Department of Mathematics and Statistics, McGill University, Montréal, QC H3A 0G4, Canada.
E-mail: yi.yang6@mcgill.ca
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: zouxx019@umn.edu

(Received March 2016; accepted June 2017)