Biometrika (2012), **99**, 1, *pp.* 29–42
doi: 10.1093/biomet/asr066
Advance Access publication 5 January 2012

A direct approach to sparse discriminant analysis in ultra-high dimensions

BY QING MAI, HUI ZOU

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
maixx034@umn.edu hzou@stat.umn.edu

AND MING YUAN

M. Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, U.S.A.
myuan@isye.gatech.edu

SUMMARY

Sparse discriminant methods based on independence rules, such as the nearest shrunken centroids classifier (Tibshirani et al., 2002) and features annealed independence rules (Fan & Fan, 2008), have been proposed as computationally attractive tools for feature selection and classification with high-dimensional data. A fundamental drawback of these rules is that they ignore correlations among features and thus could produce misleading feature selection and inferior classification. We propose a new procedure for sparse discriminant analysis, motivated by the least squares formulation of linear discriminant analysis. To demonstrate our proposal, we study the numerical and theoretical properties of discriminant analysis constructed via lasso penalized least squares. Our theory shows that the method proposed can consistently identify the subset of discriminative features contributing to the Bayes rule and at the same time consistently estimate the Bayes classification direction, even when the dimension can grow faster than any polynomial order of the sample size. The theory allows for general dependence among features. Simulated and real data examples show that lassoed discriminant analysis compares favourably with other popular sparse discriminant proposals.

Some key words: Discriminant analysis; Features annealed independence rule; Lasso; Nearest shrunken centroids classifier; Nonpolynomial-dimension asymptotics.

1. INTRODUCTION

Consider a binary classification problem where $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ represents the predictor vector and $G = 1, 2$ denotes the class label. Linear discriminant analysis is perhaps the oldest classification technique that is still being used routinely in applications. The linear discriminant analysis model assumes $x \mid G = g \sim N(\mu_g, \Sigma)$, $\mathrm{pr}(G = 1) = \pi_1$, $\mathrm{pr}(G = 2) = \pi_2$. Then, the Bayes rule, which is the theoretically optimal classifier minimizing the 0–1 loss, classifies a data point to class 2 if and only if

$$\{x - (\mu_1 + \mu_2)/2\}^{\mathrm{T}} \Sigma^{-1} (\mu_2 - \mu_1) + \log(\pi_2/\pi_1) > 0. \tag{1}$$

Let $\hat{\mu}_1, n_1$ and $\hat{\mu}_2, n_2$ be the sample mean vector and sample size within classes 1 and 2, respectively. Let $\hat{\Sigma}$ be the pooled sample covariance estimate of Σ. Linear discriminant analysis sets

$\mu_1 = \hat{\mu}_1$, $\mu_2 = \hat{\mu}_2$, $\Sigma = \hat{\Sigma}$, $\pi_1 = n_1/n$, $\pi_2 = n_2/n$ in (1). Despite its simplicity, it has proved to be a reasonably good classifier in many applications. For example, Michie et al. (1994) and Hand (2006) have shown that linear discriminant analysis has very competitive performance for many real-world benchmark datasets.

With the rapid advance of technology, high-dimensional data appear more and more frequently. In such data, the dimension, p, can be much larger than the sample size, n. It has been empirically observed that, for classification problems with high-dimension-and-low-sample-size data, some simple linear classifiers perform as well as more sophisticated classification algorithms such as the support vector machine and boosting. See, e.g., the comparison study by Dettling (2004). Hall et al. (2005) provided some geometric insight into this phenomenon. In recent years, many papers have considered ways to modify linear discriminant analysis to be suitable for high-dimensional classification. One approach is to use more sophisticated estimates of the inverse covariance matrix Σ^{-1} to replace the naive sample estimate. Under sparsity assumptions, one can obtain good estimators of Σ and Σ^{-1} even when p is much larger than n (Bickel & Levina, 2008; Cai et al., 2010; Rothman et al., 2008). However, a better estimate of Σ^{-1} does not necessarily lead to a better classifier. In an ideal scenario where we know that Σ is an identity matrix and $\pi_1 = \pi_2 = 0{\cdot}5$, then we could classify x to class 2 if $\{x - (\hat{\mu}_1 + \hat{\mu}_2)/2\}^{\mathrm{T}}(\hat{\mu}_2 - \hat{\mu}_1) > 0$. Although this classifier does not suffer from the difficulty of estimating a large covariance matrix, Fan & Fan (2008) showed that this classifier performs no better than random guessing when p is sufficiently large, due to noise accumulation in estimating μ_1 and μ_2. Therefore, effectively exploiting sparsity is critically important for high-dimensional classification.

Tibshirani et al. (2002) proposed the nearest shrunken centroids classifier for tumour classification and gene selection using microarray data. This classifier is defined as follows. For each variable x_j, we compute $d_{jg} = (1/n_g + 1/n)^{-1/2}(\hat{\mu}_{gj} - \bar{x}_j)(s_j + s_0)^{-1}$ $(g = 1, 2)$, where $\hat{\mu}_{gj}$ is the within-class sample mean, s_j^2 is the sample estimate of Σ_{jj} and s_0 is a small positive constant added for robustness. For simplicity, we set $s_0 = 0$. Define the shrunken centroids mean by

$$\hat{\mu}'_{gj} = \bar{x}_j + (n_g^{-1} + n^{-1})^{1/2} s_j d_{jg}^{\lambda} \quad (g = 1, 2),$$

where $\bar{x}_j = (n_1\hat{\mu}_1 + n_2\hat{\mu}_2)/n$ is the marginal sample mean of x_j, λ is a pre-chosen positive constant and d_{jg}^{λ} is computed by soft-thresholding d_{jg}: $d_{jg}^{\lambda} = \mathrm{sign}(d_{jg})(|d_{jg}| - \lambda)_+$ $(g = 1, 2)$. The nearest shrunken centroids classifier classifies x to class 2 if

$$\sum_{j=1}^{p} \{x_j - (\hat{\mu}'_{2j} + \hat{\mu}'_{1j})/2\} s_j^{-2} (\hat{\mu}'_{2j} - \hat{\mu}'_{1j}) + \log(n_2/n_1) > 0. \tag{2}$$

Comparing (2) and (1), we see that this classifier modifies the usual linear discriminant analysis in two ways. First, it uses only the diagonal of the sample covariance matrix to estimate Σ. If $\lambda = 0$, this classifier reduces to diagonal linear discriminant analysis, which has been shown in Bickel & Levina (2004) to work much better than linear discriminant analysis in high dimensions. Secondly, this classifier uses the shrunken centroids means to estimate μ_1, μ_2 for feature selection. If we use a sufficiently large λ, then the soft-thresholding operation will force $\hat{\mu}'_{j1} = \hat{\mu}'_{j2} = \bar{x}_j$ for some variables, which then make no contribution to the classifier defined in (2). Many experiments have shown that the nearest shrunken centroids classifier is very competitive for high-dimensional classification. More recently, Fan & Fan (2008) proposed features annealed independence rules, in which feature selection is done by hard-thresholding marginal t-statistics for testing whether $\mu_{1j} = \mu_{2j}$.

Since the goal of sparse discriminant analysis is to find those features that contribute most to classification, the target of an ideal feature selection should be the discriminative set that contains all the discriminative features that contribute to the Bayes rule, because we would use the Bayes rule for classification if it was available. Feature selection is needed when the cardinality of the discriminative set is much smaller than the total number of features. The performance of feature selection by a sparse method is measured by its ability to discover the discriminative set. There is little theoretical work for justifying the nearest shrunken centroids classifier and its variants. To our knowledge, only Fan & Fan (2008) provided detailed theoretical analysis of features annealed independence rules, under the fundamental assumption that Σ is a diagonal matrix. However, such an assumption is too restrictive to hold in applications, because strong correlations can exist in high-dimensional data, and ignoring them may lead to misleading feature selection. We argue that independence rules aim to discover the so-called signal set, whose definition is given explicitly in § 2. We further provide a necessary and sufficient condition under which this set is identical to the discriminative set. This condition can easily be violated and hence independence rules could be problematic.

In this work, we propose a new procedure for sparse discriminant analysis in high dimensions. Our proposal is motivated by the fact that classical linear discriminant analysis can be reconstructed exactly via least squares (Hastie et al., 2008). We suggest using penalized sparse least squares to derive sparse discriminant methods. Our proposal is computationally efficient in high dimensions owing to efficient algorithms for computing penalized least squares. We further provide theoretical justifications for our proposal. If the Bayes rule has a sparse representation, our theoretical results show that the proposed sparse method can simultaneously identify the discriminative set and estimate the Bayes classification direction consistently. The theory is valid even when the dimension can grow faster than any polynomial order of the sample size and does not impose strong assumptions on the correlation structure among predictors.

2. Signal set and the discriminative set

Consider the problem of tumour classification with gene expression arrays. It is an intuitively sound claim that differentially expressed genes should be responsible for the tumour classification and equally expressed genes can safely be discarded. However, we show in this section that a differentially expressed gene can have no role in classification and an equally expressed gene can significantly influence classification.

By definition, the discriminative set is equal to $A = \{j : \{\Sigma^{-1}(\mu_2 - \mu_1)\}_j \neq 0\}$, since the Bayes classification direction is $\Sigma^{-1}(\mu_2 - \mu_1)$. Variables in A are called discriminative variables. Define the signal set $\tilde{A} = \{j : \mu_{1j} \neq \mu_{2j}\}$; variables in $\tilde{A}$ are called signals. Ideally, $\tilde{A}$ is the variable selection outcome of an independence rule. Practically, independence rules pick the strongest signals indicated by the data. When Σ is diagonal, $A = \tilde{A}$. For a general covariance matrix, however, the discriminative and the signal sets can be very different, as shown in the following proposition.

Proposition 1. *Let*

$$\Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,A^c} \\ \Sigma_{A^c,A} & \Sigma_{A^c,A^c} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{\tilde{A},\tilde{A}} & \Sigma_{\tilde{A},\tilde{A}^c} \\ \Sigma_{\tilde{A}^c,\tilde{A}} & \Sigma_{\tilde{A}^c,\tilde{A}^c} \end{pmatrix}.$$

1. *If and only if* $\Sigma_{\tilde{A}^c,\tilde{A}}\Sigma_{\tilde{A},\tilde{A}}^{-1}(\mu_{2,\tilde{A}} - \mu_{1,\tilde{A}}) = 0$, *we have* $A \subseteq \tilde{A}$.
2. *If and only if* $\mu_{2,A^c} = \mu_{1,A^c}$ *or* $\Sigma_{A^c,A}\Sigma_{A,A}^{-1}(\mu_{2,A} - \mu_{1,A}) = 0$, *we have* $\tilde{A} \subseteq A$.

Based on Proposition 1 we can construct examples that a non-signal can be discriminative and a nondiscriminative feature can be a signal. Consider a linear discriminant analysis model with $p = 25$, $\mu_1 = 0_p$, $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0{\cdot}5$ $(i, j = 1, \ldots, 25; i \neq j)$. If $\mu_2 = (1_5^{\mathrm{T}}, 0_{20}^{\mathrm{T}})^{\mathrm{T}}$, then $\tilde{A} = \{1, 2, 3, 4, 5\}$ and $A = \{j : j = 1, \ldots, 25\}$, since $\Sigma^{-1}(\mu_2 - \mu_1) = (1{\cdot}62 \times 1_5^{\mathrm{T}}, -0{\cdot}38 \times 1_{20}^{\mathrm{T}})^{\mathrm{T}}$. Similarly, if we let $\mu_2 = (3 \times 1_5^{\mathrm{T}}, 2{\cdot}5 \times 1_5^{\mathrm{T}})^{\mathrm{T}}$, then all variables are signals but $A = \{1, 2, 3, 4, 5\}$, because $\Sigma^{-1}(\mu_2 - \mu_1) = (1_5^{\mathrm{T}}, 0_{20}^{\mathrm{T}})^{\mathrm{T}}$.

The above arguments warn us that independence rules could select a wrong set of features. Different sparse discriminant analysis methods have been proposed based on Fisher's view of linear discriminant analysis: the discriminant direction is obtained by maximizing $\beta^{\mathrm{T}}\hat{B}\beta/\beta^{\mathrm{T}}\hat{\Sigma}\beta$, where $\hat{B} = (\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}(\hat{\mu}_2 - \hat{\mu}_1)$. Wu et al. (2009) proposed the ℓ_1-constrained Fisher discriminant:

$$\min_{\beta} \quad \beta^{\mathrm{T}}\hat{\Sigma}\beta, \quad \text{subject to} \quad (\beta^{\mathrm{T}}\hat{B}\beta)^{1/2} = 1, \ \|\beta\|_1 \leqslant \tau. \tag{3}$$

A referee pointed out that a similar estimator was proposed by Trendafilov & Jolliffe (2007). When revising this paper, it came to our attention that Witten & Tibshirani (2011) proposed another ℓ_1-penalized linear discriminant:

$$\max_{\beta} \quad \left\{ \beta^{\mathrm{T}}\hat{B}\beta - \lambda \sum_{j=1}^{p} |s_j\beta_j| \right\}, \quad \text{subject to} \quad \beta^{\mathrm{T}}\hat{\Sigma}\beta \leqslant 1. \tag{4}$$

Little is known about the theoretical properties of the estimators defined in (3) and (4), but we include them in our numerical experiments.

3. METHOD AND THEORY

3·1. *Sparse discriminant analysis via penalized least squares*

Our approach is motivated by the intimate connection between linear discriminant analysis and least squares in the classical $p < n$ setting; see Chapter 4 of Hastie et al. (2008). We code the class labels as $y_1 = -n/n_1$ and $y_2 = n/n_2$, where $n = n_1 + n_2$. Let

$$(\hat{\beta}^{\mathrm{ols}}, \hat{\beta}_0^{\mathrm{ols}}) = \arg\min_{\beta,\beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2. \tag{5}$$

Then $\hat{\beta}^{\mathrm{ols}} = c\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ for some positive constant c. In other words, the least squares formulation in (5) exactly derives the usual linear discriminant analysis direction.

This connection is lost in high-dimensional problems because the sample covariance estimate is not invertible and the linear discriminant analysis direction is not well defined. However, we may consider a penalized least squares formulation to produce a classification direction. Let $P_\lambda(\cdot)$ be a generic sparsity-inducing penalty, such as the lasso penalty (Tibshirani, 1996) where $P_\lambda(t) = \lambda t$ $(t \geqslant 0)$. We first compute the solution to a penalized least squares problem,

$$(\hat{\beta}^{\lambda}, \hat{\beta}_0^{\lambda}) = \arg\min_{\beta,\beta_0} \left\{ n^{-1} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \right\}. \tag{6}$$

Then our classification rule is to assign x to class 2 if

$$x^{\mathrm{T}}\hat{\beta}^{\lambda} + \hat{\beta}_0 > 0. \tag{7}$$

Note that $\hat{\beta}_0$ in (7) differs from $\hat{\beta}_0^{\lambda}$ in (6). In the $p \ll n$ case, consider the ordinary least squares estimator and the usual linear discriminant analysis. Let us write $\hat{\beta}^{\mathrm{ols}} = c\hat{\beta}^{\mathrm{LDA}}$, where $\hat{\beta}^{\mathrm{LDA}} = \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$. We should use $\hat{\beta}_0 = c\hat{\beta}_0^{\mathrm{LDA}}$ in (7), where $\hat{\beta}_0^{\mathrm{LDA}} = \log(n_2/n_1) - \{(\hat{\mu}_1 + \hat{\mu}_2)/2\}^{\mathrm{T}}\hat{\beta}^{\mathrm{LDA}}$, such that the ordinary least squares classifier and the linear discriminant analysis rule yield an identical classification. If we use $\hat{\beta}_0^{\mathrm{ols}}$ in (7), then these two classifiers are not the same in general. Finding the right intercept is critical for classification but receives little attention in the literature. Hastie et al. (2008) mentioned that one could choose the intercept $\hat{\beta}_0$ empirically by minimizing the training error. Fortunately, there is a closed-form formula for computing the optimal intercept.

PROPOSITION 2. *Suppose that a linear classifier assigns x to class 2 if $x^{\mathrm{T}}\tilde{\beta} + \tilde{\beta}_0 > 0$. Given $\tilde{\beta}$, if $(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta} > 0$, then the optimal intercept is*

$$\beta_0^{\mathrm{opt}} = -(\mu_1 + \mu_2)^{\mathrm{T}}\tilde{\beta}/2 + \tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}\{(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta}\}^{-1}\log(\pi_2/\pi_1), \tag{8}$$

which can be estimated by

$$\hat{\beta}_0^{\mathrm{opt}} = -(\hat{\mu}_1 + \hat{\mu}_2)^{\mathrm{T}}\tilde{\beta}/2 + \tilde{\beta}^{\mathrm{T}}\hat{\Sigma}\tilde{\beta}\{(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\tilde{\beta}\}^{-1}\log(n_2/n_1). \tag{9}$$

By Proposition 2, we calculate $\hat{\beta}_0$ and then the classifier given in (7) assigns x to class 2 if

$$\{x - (\hat{\mu}_1 + \hat{\mu}_2)/2\}^{\mathrm{T}}\hat{\beta}^{\lambda} + (\hat{\beta}^{\lambda})^{\mathrm{T}}\hat{\Sigma}\hat{\beta}^{\lambda}\{(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\hat{\beta}^{\lambda}\}^{-1}\log(n_2/n_1) > 0.$$

Remark 1. The condition $(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta} > 0$ in Proposition 2 is very mild. If the linear classifier actually yields $(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta} < 0$, then we can always use $\tilde{\beta}_{\mathrm{new}} = -\tilde{\beta}$, which obeys $(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta}_{\mathrm{new}} > 0$.

Remark 2. If $\pi_1 = \pi_2 = 0{\cdot}5$, then we can take $\hat{\beta}_0^{\mathrm{opt}} = -(\hat{\mu}_1 + \hat{\mu}_2)^{\mathrm{T}}\tilde{\beta}/2$. In general, we need to include the second term in the right-hand side of (9). If $\pi_1 \neq \pi_2$ and without the sparsity condition on $\tilde{\beta}$, the second term in (9) would not work well when $p \gg n$. Fortunately, when $\tilde{\beta}$ is sparse, we have

$$\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta} = \sum_{\substack{i,j:\tilde{\beta}_i \neq 0,\\ \tilde{\beta}_j \neq 0}} \Sigma_{ij}\tilde{\beta}_i\tilde{\beta}_j, \quad \tilde{\beta}^{\mathrm{T}}\hat{\Sigma}\tilde{\beta} = \sum_{\substack{i,j:\tilde{\beta}_i \neq 0,\\ \tilde{\beta}_j \neq 0}} \hat{\Sigma}_{ij}\tilde{\beta}_i\tilde{\beta}_j.$$

Thus, even when $p \gg n$, as long as $\|\tilde{\beta}\|_0 \ll n$, $\tilde{\beta}^{\mathrm{T}}\hat{\Sigma}\tilde{\beta}$ is a good estimator for $\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}$. Using a regularized estimate of Σ could provide some further improvement. For example, for banded covariance matrices, the banding estimator (Bickel & Levina, 2008) and the tapering estimator (Cai et al., 2010) are better estimators for Σ than the sample covariance. However, in this work, our primary focus is $\hat{\beta}^{\lambda}$ and we do not want to entangle estimating large covariance matrices with feature selection.

In principle, (6) can work with any sparsity-inducing penalty function. We choose the lasso in this work because it is the most popular penalty in the literature. Other popular penalty functions include the smoothly clipped absolute deviation (Fan & Li, 2001), the elastic net (Zou & Hastie, 2005), the adaptive lasso (Zou, 2006), the fused lasso (Tibshirani et al., 2005), the grouped

lasso (Yuan & Lin, 2006) and the minimum concavity penalty (Zhang, 2010). For convenience, we call the resulting classifier lassoed discriminant analysis from now on. We can use either the least angle regression algorithm (Efron et al., 2004) or the coordinate descent algorithm (Friedman et al., 2010) to compute the lasso-penalized least squares estimator.

3·2. *Theory*

We first introduce some necessary notation. For a general $m \times n$ matrix M, define $\|M\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |M_{ij}|$. For any vector b, let $\|b\|_\infty = \max_j |b_j|$ and $|b|_{\min} = \min_j |b_j|$. We let $\beta^{\text{Bayes}} = \Sigma^{-1}(\mu_2 - \mu_1)$ represent the Bayes classifier coefficient vector. So, $A = \{j : \beta_j^{\text{Bayes}} \neq 0\}$ and let s be the cardinality of A. We use $C = \text{cov}(x)$ to represent the marginal covariance matrix of the predictors and partition C as

$$C = \begin{pmatrix} C_{AA} & C_{AA^c} \\ C_{A^cA} & C_{A^cA^c} \end{pmatrix}.$$

We define three quantities that frequently appear in our analysis:

$$\kappa = \|C_{A^cA}(C_{AA})^{-1}\|_\infty, \quad \varphi = \|(C_{AA})^{-1}\|_\infty, \quad \Delta = \|\mu_{2A} - \mu_{1A}\|_\infty.$$

Suppose that X is the predictor matrix and let $\tilde{X}$ be the centred predictor matrix, whose column-wise mean is zero. Obviously, $C^{(n)} = \tilde{X}^{\mathrm{T}}\tilde{X}/n$ is an empirical version of C. Likewise, we can write $\tilde{X}_A^{\mathrm{T}}\tilde{X}_A/n = C_{AA}^{(n)}$ and $\tilde{X}_{A^c}^{\mathrm{T}}\tilde{X}_A/n = C_{A^cA}^{(n)}$.

Denote $\beta^* = (C_{AA})^{-1}(\mu_{2A} - \mu_{1A})$. Now we can define $\tilde{\beta}^{\text{Bayes}}$ by letting $\tilde{\beta}_A^{\text{Bayes}} = \beta^*$ and $\tilde{\beta}_{A^c}^{\text{Bayes}} = 0$. The following proposition shows the equivalence between $\tilde{\beta}^{\text{Bayes}}$ and β^{Bayes}.

PROPOSITION 3. *The quantities $\tilde{\beta}^{\text{Bayes}}$ and β^{Bayes} are equivalent in the sense that $\tilde{\beta}^{\text{Bayes}} = c\beta^{\text{Bayes}}$ for some positive constant c and the Bayes classifier can be written as assigning x to class 2 if*

$$\{x - (\mu_1 + \mu_2)/2\}^{\mathrm{T}}\tilde{\beta}^{\text{Bayes}} + (\tilde{\beta}^{\text{Bayes}})^{\mathrm{T}}\Sigma\tilde{\beta}^{\text{Bayes}}\{(\mu_2 - \mu_1)^{\mathrm{T}}\tilde{\beta}^{\text{Bayes}}\}^{-1}\log(\pi_2/\pi_1) > 0.$$

Proposition 3 tells us that it suffices to show that the proposed sparse discriminant analysis can consistently recover the support of $\tilde{\beta}^{\text{Bayes}}$ and estimate β^*.

Throughout our analysis we assume that the variance of each variable is bounded by a finite constant. In practice, one often standardizes the data beforehand. Then the finite constant can be taken as unity. In this subsection, ϵ_0 and c_1, c_2 are positive constants.

The lassoed discriminant analysis direction is computed by

$$(\hat{\beta}^{\text{lasso}}, \hat{\beta}_0^{\lambda}) = \arg\min_{\beta,\beta_0} \left\{ n^{-1}\sum_{i=1}^n (y_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2 + \lambda\sum_{j=1}^p |\beta_j| \right\}. \tag{10}$$

If lassoed discriminant analysis finds the support of the Bayes rule, then we should have $\hat{\beta}_{A^c}^{\text{lasso}} = 0$ and $\hat{\beta}_A^{\text{lasso}}$ should be identical to $\hat{\beta}_A$, where

$$\hat{\beta}_A = \arg\min_{\beta,\beta_0} \left\{ n^{-1}\sum_{i=1}^n (y_i - \beta_0 - \sum_{j\in A} x_{ij}\beta_j)^2 + \sum_{j\in A}\lambda|\beta_j| \right\}. \tag{11}$$

We introduce $\hat{\beta}_A$ only for mathematical analysis. It is not a real estimator, because its definition depends on knowing A. To ensure that the lassoed discriminant analysis has the variable selection consistency property, we impose a condition on the covariance matrix of the predictors:

$$\kappa = \|C_{A^cA}(C_{AA})^{-1}\|_\infty < 1. \tag{12}$$

The condition in (12) is an analogue of the irrepresentability condition for the lasso regression estimator (Meinshausen & Bühlmann, 2006; Zou, 2006; Zhao & Yu, 2006; Wainwright, 2009). This condition can be relaxed if using a concave penalty to derive $\hat{\beta}^\lambda$; see § 5.

THEOREM 1. *Pick any* λ *such that* $\lambda < \min\{|\beta^*|_{\min}/(2\varphi), \Delta\}$. *Then:*

1. *assuming the condition in* (12), *with probability at least* $1-\delta_1$, $\hat{\beta}^{\text{lasso}}_A = \hat{\beta}_A$ *and* $\hat{\beta}^{\text{lasso}}_{A^c} = 0$, *where*

$$\delta_1 = 2ps\exp(-c_1 n s^{-2}\epsilon^2) + 2p\exp\{-c_2 n\lambda^2(1-\kappa-2\epsilon\varphi)^2(1+\kappa)^{-2}/16\} \tag{13}$$

   and ϵ *is any positive constant less than* $\min[\epsilon_0, \lambda(1-\kappa)(4\varphi)^{-1}\{\lambda/2+(1+\kappa)\Delta\}^{-1}]$;
2. *with probability at least* $1-\delta_2$, *none of the elements of* $\hat{\beta}_A$ *is zero, where*

$$\delta_2 = 2s^2\exp(-c_1 n s^{-2}\epsilon^2) + 2s\exp(-c_2 n\epsilon^2) \tag{14}$$

   and ϵ *is any positive constant less than* $\min\{\epsilon_0, \zeta(3+\zeta)^{-1}/\varphi, \Delta\zeta(6+2\zeta)^{-1}\}$;
3. *for any positive* ϵ *satisfying* $\epsilon < \min\{\epsilon_0, \lambda(2\varphi\Delta)^{-1}, \lambda\}$, *we have*

$$\text{pr}(\|\hat{\beta}_A - \beta^*\|_\infty \leqslant 4\varphi\lambda) \geqslant 1 - 2s^2\exp(-c_1 n s^{-2}\epsilon^2) - 2s\exp(-c_2 n\epsilon^2). \tag{15}$$

The non-asymptotic results in Theorem 1 can be easily translated into asymptotic arguments when allowing the triple (n, s, p) to tend to infinity at suitable rates. To highlight the main points, we assume that Δ, κ, φ are constants. In addition, we need the following regularity conditions:

Condition 1. $n, p \to \infty$ and $\log(ps)s^2/n \to 0$;

Condition 2. $|\beta^*|_{\min} \gg \{\log(ps)s^2/n\}^{1/2}$.

Condition 1 restricts p. Clearly, we cannot expect the proposed method to work for an arbitrarily large p. However, the restriction is rather loose. Consider the case where $s = O(n^{1/2-\gamma})$ for some $\gamma < 1/2$. Condition 1 holds as long as $p \ll e^{n^{2\gamma}}$. Therefore, p is allowed to grow faster than any polynomial order of n, referred to as nonpolynomial-dimension asymptotics. Condition 2 requires the nonzero elements of the Bayes rule to be large enough such that we could consistently separate them from zero using the observed data. The lower bound actually converges to zero under Condition 1, so Condition 2 is not strong.

THEOREM 2. *Let* $\hat{A} = \{j : \hat{\beta}^{\text{lasso}}_j \neq 0\}$. *Under Conditions* 1 *and* 2, *if we choose* $\lambda = \lambda_n$ *such that* $\lambda_n \ll |\beta^*|_{\min}$ *and* $\lambda_n \gg \{\log(ps)s^2/n\}^{1/2}$, *and further assume* $\kappa < 1$, *then* $\text{pr}(\hat{A} = A) \to 1$ *and* $\text{pr}(\|\hat{\beta}^{\text{lasso}}_A - \beta^*\|_\infty \leqslant 4\varphi\lambda_n) \to 1$.

Remark 3. Although we use penalized least squares to estimate the classification direction, there is a fundamental difference between Theorem 1 and theoretical results derived for lasso-penalized least squares regression (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2009). The previous work assumes that the data obey a linear regression model with additive noise, which is not true for y and x in (10).

Table 1. *Simulation models. The choices of n, p, Σ and β^{Bayes} are listed*

Model	n	p	Σ	β^{Bayes}
1	100	400	$\Sigma_{ij} = 0{\cdot}5^{\lvert i-j\rvert}$.	$0{\cdot}556\ (3, 1{\cdot}5, 0, 0, 2, 0_{p-5})^{\mathrm{T}}$
2	100	400	$\Sigma_{ij} = 0{\cdot}5^{\lvert i-j\rvert}$.	$0{\cdot}582\ (3, 2{\cdot}5, -2{\cdot}8, 0_{p-3})^{\mathrm{T}}$
3	400	800	$\Sigma_{jj} = 1$, $\Sigma_{ij} = 0{\cdot}5$, $i \neq j$.	$0{\cdot}395\ (3, 1{\cdot}7, -2{\cdot}2, -2{\cdot}1, 2{\cdot}55, 0_{p-5})^{\mathrm{T}}$
4	300	800	$\Sigma = \mathrm{I}_5 \otimes \tilde{\Sigma}$, I_5 is an identity matrix; $\tilde{\Sigma}_{jj} = 1$, $\tilde{\Sigma}_{ij} = 0{\cdot}6$, $i \neq j$.	$0{\cdot}916\ (1{\cdot}2, -1{\cdot}4, 1{\cdot}15, -1{\cdot}64, 1{\cdot}5, -1, 2, 0_{p-7})^{\mathrm{T}}$
5	400	800	$\Sigma_{jj} = 1$, $\Sigma_{ij} = 0{\cdot}5$, $i \neq j$.	$0{\cdot}551\ (3, 1{\cdot}7, -2{\cdot}2, -2{\cdot}1, 2{\cdot}55, (p-5)^{-1}1_{p-5})^{\mathrm{T}}$
6	400	800	$\Sigma_{jj} = 1$, $\Sigma_{ij} = 0{\cdot}5$, $i \neq j$.	$0{\cdot}362\ (3, 1{\cdot}7, -2{\cdot}2, -2{\cdot}1, 2{\cdot}55, (p-5)^{-1}1_{p-5})^{\mathrm{T}}$

Remark 4. Our method is also fundamentally different from those based on high-dimensional covariance estimation. In the current literature on covariance or inverse-covariance matrix estimation, a commonly used assumption is that the target matrix has some sparsity structure (Bickel & Levina, 2008; Cai et al., 2010). Such assumptions are not needed in our method.

4. NUMERICAL RESULTS

4·1. *Simulation*

We use simulated data to demonstrate the good performance of our proposal. For comparison, we included the nearest shrunken centroids classifier (Tibshirani et al., 2002), the features annealed independence rule (Fan & Fan, 2008), the ℓ_1-penalized linear discriminant (Witten & Tibshirani, 2011) and the ℓ_1-constrained Fisher discriminant (Wu et al., 2009). The nearest shrunken centroids classifier is implemented in the R package pamr; see http://cran.r-project.org/web/packages/pamr/index.html. The ℓ_1-penalized linear discriminant is implemented in the R package penalizedLDA; see http://cran.r-project.org/web/packages/penalizedLDA/index.html. We used the code of Wu et al. (2009) to implement the ℓ_1-constrained Fisher discriminant. We also considered the t-test classifier: we first performed Bonferroni-adjusted t-tests with size 0·05 and then did linear discriminant analysis only using these features that passed the t-test.

We randomly generated n class labels such that $\pi_1 = \pi_2 = 0{\cdot}5$. Conditioning on the class labels g ($g = 1, 2$), we generated the p-dimensional predictor x from a multivariate normal distribution with mean vector μ_g and covariance Σ. Without loss of generality, we set $\mu_1 = 0$ and $\mu_2 = \Sigma\beta^{\text{Bayes}}$. We considered six different simulation models. The choices of n, p, Σ and β^{Bayes} are shown in Table 1. Models 1–4 are sparse discriminant models with different covariance and mean structure; Models 5 and 6 are practically sparse in the sense that their Bayes rules depend on all variables in theory but can be well approximated by sparse discriminant functions. Table 2 summarizes the simulation results based on 2000 replications.

Our method, lassoed discriminant analysis, is the only one that shows good performance in all six simulation settings. It closely mimics the Bayes rule, regardless of the Bayes error and covariance structure. Tibshirani's method and Fan's method have very comparable performance, but they are much worse than ours except for Model 1. In Model 1, the first five elements of $\mu_2 - \mu_1$ are much larger than the rest, which implies that independence rules can include all three discriminative variables. On the other hand, although Model 2 uses the same Σ as in Model 1, it has a very different mean structure: the first two elements of $\mu_2 - \mu_1$ are huge while the rest are much smaller. This means that independence rules have difficulty in selecting variable 3, resulting in inferior classification. Wu's method has good classification accuracy overall, but

Table 2. *Simulation results. The methods are named after the first author of the original papers. The reported numbers are medians with their standard errors, obtained by bootstrap, in parentheses. TRUE selection and FALSE selection denote the numbers of selected variables from the discriminative set and its complement, respectively. Fitted model size is the total number of selected variables*

	Bayes	Our method	Wu	Witten	Tibshirani	Fan	*t*-test classifier
Model 1							
Error (%)	10	10·89	13·71	10·81	10·94	11·47	10·46
		(0·03)	(0·01)	(0·01)	(0·02)	(0·05)	(0·01)
TRUE selection	3	3	3	1	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	2	0	26	6	7	1
		(0·16)	(0·49)	(0·11)	(0·61)	(0·66)	(0)
Model 2							
Error (%)	10	12·84	14·5	14·25	15·12	15·67	14·12
		(0·05)	(0·01)	(0·02)	(0·05)	(0·07)	(0·01)
TRUE selection	3	3	1	2	2	2	2
		(0)	(0·14)	(0)	(0·34)	(0)	(0)
FALSE selection	0	6	0	4	9	8	0
		(0·27)	(0)	(0·61)	(0·73)	(0·29)	(0)
Model 3							
Error (%)	20	21·93	22·37	33·69	27·48	25·69	38·94
		(0·03)	(0·05)	(0·01)	(0·07)	(0·02)	(0·03)
TRUE selection	5	5	5	3	3	2	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	14	2	419·5	2	0	790
		(0·59)	(0)	(10·19)	(0·31)	(0)	(0·42)
Model 4							
Error (%)	10	12·50	13·99	23·90	19·25	18·56	24·59
		(0·02)	(0·03)	(0·01)	(0·04)	(0·00)	(0·06)
TRUE selection	7	7	6	4	4	3	6
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	18	2	35	1	0	153
		(0·70)	(0)	(4·43)	(0·48)	(0)	(0)
Model 5							
Error (%)	10	11·11	12·07	21·99	14·72	14·27	25·79
		(0·02)	(0·07)	(0·01)	(0·03)	(0·01)	(0·04)
Fitted model size	800	21	7	737	3	3	800
		(0·65)	(0·16)	(2·29)	(0·46)	(0)	(0)
Model 6							
Error (%)	20	22·22	23·34	30·43	26·13	24·14	36·80
		(0·03)	(0·05)	(0·01)	(0·07)	(0)	(0·04)
Fitted model size	800	20	5	592·5	8	3	798
		(0·53)	(0·49)	(7·46)	(0·51)	(0)	(0)

it can often miss some important features. Witten and Tibshirani's method has rather poor performance, which is somewhat surprising because the basic idea is similar to Wu's. Witten and Tibshirani's formulation in (4) is nonconvex while Wu's formulation in (3) is convex, which may help explain their different performances. The *t*-test classifier is best for Model 1, second best for Model 2 and worst for Models 3–6.

Table 3. *Comparing lassoed discriminant analysis with other approaches on the colon and the prostate datasets*

		Our method	Witten	Wu	Tibshirani	Fan	*t*-test classifier
Colon	Error (%)	86·4 (1·54)	86·4 (0·49)	84·1 (2·17)	86·4 (1·20)	86·4 (0·61)	81·0 (1·67)
	Fitted model size	5 (0·63)	10 (1·39)	1 (0)	89 (29·95)	11 (1·19)	16 (1·15)
Prostate	Error (%)	94·1 (0·55)	91·2 (0·24)	91·2 (0·70)	91·2 (0·96)	76·5 (0·54)	76·5 (1·62)
	Fitted model size	10 (0·77)	18 (4·45)	1 (0)	10 (0·84)	4 (0·40)	58 (2·65)

Table 4. *Adjusted classification accuracies of four methods by forcing them to select a similar number of genes as does our method*

	Witten	Wu	Tibshirani	Fan
Colon	86·4 (0·51)	86·4 (1·06)	63·6 (0·70)	77·3 (2·16)
Prostate	94·1 (1·25)	91·2 (1·24)	91·2 (1·39)	73·5 (1·11)

4·2. *Real data*

We further compare the methods on two benchmark datasets: the colon and prostate cancer datasets. The basic task here is to predict whether an observation is tumour or is normal tissue. We randomly split the datasets into the training and test sets with ratio 2:1. Model fitting was done on the training set and classification accuracy was evaluated on the test set. This procedure was repeated 100 times. Shown in Table 3 are the classification accuracies and the numbers of genes selected by each competitor.

The colon and prostate datasets have been previously used to test classification and feature selection methods. See Alon et al. (1999), Singh et al. (2002) and Dettling (2004). Dettling (2004) reported that BagBoost was the most accurate classifier for the prostate data, with 92·5% classification accuracy, and the nearest shrunken centroids classifier was the most accurate classifier for the colon data. Table 2 shows that our method is as accurate as the nearest shrunken centroids classifier on the colon data and significantly outperforms BagBoost on the prostate data. Since BagBoost does not do gene selection, we do not include it in Table 2. Witten and Tibshirani's method works quite well on these two real datasets. As suggested by a referee, we adjusted the tuning parameters of Witten and Tibshirani's, Wu's, Tibshirani's and Fan's methods such that they selected a similar number of genes to that by our method on each dataset. Their adjusted classification errors are reported in Table 4. This adjustment helps Witten and Tibshirani's method but degrades Tibshirani's and Fan's methods.

5. DISCUSSION

Sparse discriminant analysis based on independence rules is computationally attractive for high-dimensional classification. However, independence rules may lead to misleading feature selection and hence poor classification performance, due to the difference between discriminative and signal variables. When doing feature selection in classification, one should aim to recover the discriminative set, not the signal set, which is the goal of large-scale hypothesis testing. Discovering the signal set is the fundamental question of research in many scientific studies (Efron, 2010), but identifying features for classification could be very different from identifying interesting signals, and hence the statistical tools for data analysis should be carefully chosen.

If one feels that the condition (12) is somewhat strong for establishing the nonpolynomial-dimension theory, one may use a concave penalty other than the lasso penalty. We have tried

using the smoothly clipped absolute deviation penalty (Fan & Li, 2001) and have shown that the resulting sparse discriminant algorithm enjoys a strong oracle property without requiring the condition (12). These results are given in a technical report which is available upon request. We only focus on lassoed discriminant analysis in the current paper because our primary goal is to demonstrate the effectiveness of the penalized least squares formulation of sparse discriminant analysis. We do not want to overly emphasize the penalty function. For a given dataset, one penalty function may be more appropriate than others. For example, in some situations, the predictors may have a natural ordering, so ordered variable selection is preferred, and the nested lasso (Levina et al., 2008) is a better choice than the lasso or smoothly clipped absolute deviation penalty. It is straightforward to implement the nested lasso in our proposal, but a detailed treatment is beyond the scope of this paper.

Acknowledgement

The authors thank the editor, associate editor and referees for their helpful comments and suggestions. Mai is supported by an Alumni Fellowship from the School of Statistics at the University of Minnesota. Zou and Yuan are supported by the National Science Foundation, U.S.A.

Appendix

Proof of Proposition 1. 1. Let $\Omega = \Sigma^{-1}$ and $\beta^{\text{Bayes}} = \Omega(\mu_2 - \mu_1)$. Note that $A \subseteq \tilde{A}$ is equivalent to $\beta^{\text{Bayes}}_{\tilde{A}^c} = 0$, and $\beta^{\text{Bayes}}_{\tilde{A}^c} = \Omega_{\tilde{A}^c,\tilde{A}}(\mu_{2,\tilde{A}} - \mu_{1,\tilde{A}})$, where $\Omega_{\tilde{A}^c,\tilde{A}} = -(\Sigma_{\tilde{A}^c,\tilde{A}^c} - \Sigma_{\tilde{A}^c,\tilde{A}}\Sigma^{-1}_{\tilde{A},\tilde{A}}\Sigma_{\tilde{A},\tilde{A}^c})^{-1}\Sigma_{\tilde{A}^c,\tilde{A}}\Sigma^{-1}_{\tilde{A},\tilde{A}}$. Therefore, part 1 is proven.

2. By definition, $\tilde{A} \subseteq A$ is equivalent to $\mu_{2,A^c} = \mu_{1,A^c}$. Now using $\mu_2 - \mu_1 = \Sigma\beta^{\text{Bayes}}$ we have $\mu_{2,A} - \mu_{1,A} = \Sigma_{A,A}\beta^{\text{Bayes}}_A$ and $\mu_{2,A^c} - \mu_{1,A^c} = \Sigma_{A^c,A}\beta^{\text{Bayes}}_A$. Hence, $\mu_{2,A^c} - \mu_{1,A^c} = \Sigma_{A^c,A}\Sigma^{-1}_{A,A}(\mu_{2,A} - \mu_{1,A})$. Thus part 2 is proven. □

Proof of Proposition 2. We recode the response variable as $y^* = \pm 1$. Note that $\tilde{\beta}_0^{\text{opt}} = \arg\min_{\tilde{\beta}_0} E\{y^*_{\text{new}} \neq \text{sign}(x^{\mathrm{T}}_{\text{new}}\tilde{\beta} + \tilde{\beta}_0) \mid \text{training data}\}$. Since $y^*_{\text{new}}, x_{\text{new}}$ are independent from the training data, $(Y^*_{\text{new}}, z_{\text{new}} = x^{\mathrm{T}}_{\text{new}}\tilde{\beta})$ obeys a one-dimensional linear discriminant analysis model, that is, $z_{\text{new}} \mid y^*_{\text{new}} = 1 \sim N(\tilde{\beta}^{\mathrm{T}}\mu_2, \tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta})$, $\text{pr}(y^*_{\text{new}} = 1) = \pi_2$ and $z_{\text{new}} \mid y^*_{\text{new}} = -1 \sim N(\tilde{\beta}^{\mathrm{T}}\mu_1, \tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta})$, $\text{pr}(y^*_{\text{new}} = -1) = \pi_1$. Then a straightforward calculation gives (8). □

Proof of Proposition 3. By definition we can write $C_{AA} = \Sigma_{AA} + \pi_1\pi_2(\mu_{2A} - \mu_{1A})(\mu_{2A} - \mu_{1A})^{\mathrm{T}}$ and $C_{AA}\tilde{\beta}^{\text{Bayes}}_A = \mu_{2A} - \mu_{1A}$. Let $c = n\{n - 2 + \pi_1\pi_2(\mu_{2A} - \mu_{1A})^{\mathrm{T}}\Sigma^{-1}_{AA}(\mu_{2A} - \mu_{1A})\}^{-1} > 0$, then $\tilde{\beta}^{\text{Bayes}}_A = c\beta^{\text{Bayes}}_A$. □

We now prove Theorems 1 and 2. With $y = n/n_2$, $-n/n_1$ and the centred predictor matrix $\tilde{X}$, we can write

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta}\left\{ n^{-1}\beta^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}\tilde{X})\beta - 2(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\beta + \lambda\sum_{j=1}^{p}|\beta_j| \right\}.$$

The following two lemmas are repeatedly used in the proof.

Lemma A1. *There exist constants ϵ_0 and c_1, c_2 such that for any $\epsilon \leqslant \epsilon_0$ we have*

$$\text{pr}(|C^{(n)}_{ij} - C_{ij}| \geqslant \epsilon) \leqslant 2\exp(-n\epsilon^2 c_1) \ (i, j = 1, \ldots, p); \tag{A1}$$

$$\text{pr}\{|(\hat{\mu}_{2j} - \hat{\mu}_{1j}) - (\mu_{2j} - \mu_{1j})| \geqslant \epsilon\} \leqslant 2\exp(-n\epsilon^2 c_2) \ (j = 1, \ldots, p); \tag{A2}$$

$$\text{pr}(\|C^{(n)}_{AA} - C_{AA}\|_\infty \geqslant \epsilon) \leqslant 2s^2\exp(-ns^{-2}\epsilon^2 c_1); \tag{A3}$$

$$\text{pr}(\|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \geqslant \epsilon) \leqslant 2(p-s)s\exp(-ns^{-2}\epsilon^2 c_1); \tag{A4}$$

$$\text{pr}\{\|(\hat\mu_2 - \hat\mu_1) - (\mu_2 - \mu_1)\|_\infty \geqslant \epsilon\} \leqslant 2p\exp(-n\epsilon^2 c_2); \tag{A5}$$

$$\text{pr}\{\|(\hat\mu_{2A} - \hat\mu_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty \geqslant \epsilon\} \leqslant 2s\exp(-n\epsilon^2 c_2). \tag{A6}$$

LEMMA A2. *There exist constants ϵ_0, c_1 such that for any $\epsilon \leqslant \min(\epsilon_0, 1/\varphi)$, we have*

$$\text{pr}\{\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \geqslant \epsilon\varphi(\kappa+1)(1-\varphi\epsilon)^{-1}\} \leqslant 2ps\exp(-ns^{-2}\epsilon^2 c_1).$$

Proof of Lemma A1. We only prove (A1) and (A2). Inequalities (A3)–(A6) can be obtained from (A1)–(A2) by union bounds. First, $\text{pr}(|\hat\mu_{1j} - \mu_{1j}| \geqslant \epsilon \mid Y) \leqslant 2\exp\{-n_1\epsilon^2/(2\sigma_j^2)\}$. Also, $n_1 \sim \text{Ber}(n, \pi_1)$ and $\text{pr}(|n_1 - \pi_1 n| \geqslant n\epsilon) \leqslant 2\exp(-nc_2'\epsilon^2)$. Therefore, $\text{pr}(|\hat\mu_1 - \mu_1| \geqslant \epsilon) \leqslant 2\exp\{-n\pi_1\epsilon^2/(4\sigma_j^2)\} + 2\exp(-nc_2'\pi^2/4) \leqslant 2\exp(-nc_2^{(1)}\epsilon^2)$. The same inequality also holds for class 2. Thus (A2) holds.

To prove (A1), note that $C^{(n)}_{ij} = n^{-1}\sum_{k=1}^n x_{ki}x_{kj} - \bar x_i\bar x_j$. Since $\bar x_v = \hat\pi_1\hat\mu_{1v} + \hat\pi_2\hat\mu_{2v}$ $(v = i, j)$, by previous arguments, we know that there exists $c_1' > 0$ such that

$$\text{pr}\{|\bar x_i\bar x_j - E(x_i)E(x_j)| \geqslant \epsilon\} \leqslant 2\exp(-nc_1'\epsilon^2).$$

We further have that $n^{-1}\sum_{k=1}^n x_{ki}x_{kj} - E(x_ix_j) = \sum_{l=1}^2 n_l/n\{n_l^{-1}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j \mid g=l)\} + \sum_{l=1}^2 E(x_ix_j \mid g=l)(n_l/n - \pi_l)$ and $E(x_ix_j \mid g=l) = \Sigma_{ij} + \mu_{li}\mu_{lj}$ for $l = 1, 2$. Note that

$$n_l^{-1}\sum_{g_k=l} x_{ki}x_{kj} = n_l^{-1}\sum_{g_k=l}(x_{ki} - \mu_{li})(x_{kj} - \mu_{lj}) + \mu_{li}(\mu_{lj} - \hat\mu_{lj}) + \mu_{lj}(\mu_{li} - \hat\mu_{li}) + \mu_{li}\mu_{lj}.$$

Bickel & Levina (2008) showed that, for $\epsilon < \epsilon_0$,

$$\text{pr}\{|n_l^{-1}\sum_{g_k=l}(x_{ki} - \mu_{li})(x_{kj} - \mu_{lj}) - \Sigma_{ij}| > \epsilon \mid Y\} \leqslant 2\exp(-c_3 n\epsilon^2). \tag{A7}$$

Combining the concentration results for $\hat\mu_{lv}$, n_l and (A7), we have (A1). □

Proof of Lemma A2. Let $\eta_1 = \|C_{AA} - C^{(n)}_{AA}\|_\infty$, $\eta_2 = \|C_{A^cA} - C^{(n)}_{A^cA}\|_\infty$ and $\eta_3 = \|(C^{(n)}_{AA})^{-1} - (C_{AA})^{-1}\|_\infty$. First we have

$$\begin{aligned}
\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty &\leqslant \|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \times \|(C^{(n)}_{AA})^{-1} - (C_{AA})^{-1}\|_\infty \\
&\quad + \|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \times \|(C_{AA})^{-1}\|_\infty \\
&\quad + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \times \|C_{AA} - C^{(n)}_{AA}\|_\infty \times \|(C_{AA})^{-1}\|_\infty \\
&\quad + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \times \|C_{AA} - C^{(n)}_{AA}\|_\infty \\
&\quad \times \|(C^{(n)}_{AA})^{-1} - (C_{AA})^{-1}\|_\infty \\
&\leqslant (\kappa\eta_1 + \eta_2)(\varphi + \eta_3).
\end{aligned}$$

Moreover, $\eta_3 \leqslant \|(C^{(n)}_{AA})^{-1}\|_\infty \times \|(C^{(n)}_{AA} - C_{AA})\|_\infty \times \|(C_{AA})^{-1}\|_\infty = (\varphi + \eta_3)\varphi\eta_1$. So, as long as $\varphi\eta_1 < 1$, we have $\eta_3 \leqslant \varphi^2\eta_1(1 - \varphi\eta_1)^{-1}$ and hence

$$\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \leqslant (\kappa\eta_1 + \eta_2)\varphi(1 - \varphi\eta_1)^{-1}.$$

Then we consider the event $\max(\eta_1, \eta_2) \leqslant \epsilon$ and use Lemma 1 to obtain Lemma 2. □

Proof of Theorem 1. We first prove conclusion 1. By (11) we can write $\hat\beta_A = (n^{-1}\tilde X_A^{\mathrm T}\tilde X_A)^{-1}\{(\hat\mu_{2A} - \hat\mu_{1A}) - \lambda t_A/2\}$, where t_A represents the subgradient such that $t_j = \text{sign}(\hat\beta_j)$ if $\hat\beta_j \neq 0$ and $-1 < t_j < 1$ if

$\hat{\beta}_j = 0$. Write

$$\hat{\beta}_A = (C_{AA})^{-1}(\mu_{2A} - \mu_{1A}) + (C_{AA}^{(n)})^{-1}\{(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\} - \{(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\}(\mu_{2A} - \mu_{1A}) - \lambda(C_{AA}^{(n)})^{-1}t_A/2. \tag{A8}$$

In order to show that $\hat{\beta}^{\text{lasso}} = (\hat{\beta}_A, 0)$, it suffices to verify that

$$\|n^{-1}\tilde{X}_{A^c}^{\mathrm{T}}\tilde{X}_A\hat{\beta}_A - (\hat{\mu}_{2A^c} - \hat{\mu}_{1A^c})\|_\infty \leqslant \lambda/2. \tag{A9}$$

The left-hand side of (A9) is equal to

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}\lambda t_A/2 - (\hat{\mu}_{2A^c} - \hat{\mu}_{1A^c})\|_\infty. \tag{A10}$$

Using $C_{A^cA}C_{AA}^{-1}(\mu_{2A} - \mu_{1A}) = (\mu_{2A^c} - \mu_{1A^c})$, (A10) is bounded from above by

$$\begin{aligned} U_1 = {} & \|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty\Delta + \|(\hat{\mu}_{2A^c} - \hat{\mu}_{1A^c}) - (\mu_{2A^c} - \mu_{1A^c})\|_\infty \\ & + \{\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty + \kappa\}\|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty \\ & + (\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty + \kappa)\lambda/2. \end{aligned}$$

If $\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty \leqslant (\kappa + 1)\epsilon\varphi(1 - \varphi\epsilon)^{-1}$, and

$$\|(\hat{\mu}_2 - \hat{\mu}_1) - (\mu_2 - \mu_1)\|_\infty \leqslant 4^{-1}\lambda(1 - \kappa - 2\epsilon\varphi)/(1 + \kappa),$$

then $U_1 \leqslant \lambda/2$. Therefore, by Lemmas A1 and A2, we have

$$\begin{aligned} 1 - \delta_1 & \equiv \text{pr}\{\|n^{-1}\tilde{X}_{A^c}^{\mathrm{T}}\tilde{X}_A\hat{\beta}_A - (\mu_{2A^c} - \mu_{1A^c})\|_\infty \leqslant \lambda/2\} \\ & \geqslant 1 - 2ps\exp(-ns^{-2}\epsilon^2c_1) - 2p\exp[-nc_2\{4^{-1}\lambda(1 - \kappa - 2\epsilon\varphi)/(1 + \kappa)\}^2]. \end{aligned}$$

Thus (13) is proven.

We now prove conclusion 2. Let $\zeta = |\beta^*|_{\min}/(\Delta\varphi)$. Write $\eta_1 = \|C_{AA} - C_{AA}^{(n)}\|_\infty$ and $\eta_3 = \|(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\|_\infty$. Then for any $j \in A$,

$$|\hat{\beta}_j| \geqslant \zeta\Delta\varphi - (\eta_3 + \varphi)\{\lambda/2 + \|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty\} - \eta_3\Delta.$$

When $\eta_1\varphi < 1$ we have shown that $\eta_3 < \varphi^2\eta_1(1 - \eta_1\varphi)^{-1}$, thus

$$|\hat{\beta}_j| \geqslant \zeta\Delta\varphi - (1 - \eta_1\varphi)^{-1}\{\lambda\varphi/2 + \|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty\varphi + \varphi^2\eta_1\Delta\} \equiv L_1.$$

Because $\|\beta^*\|_\infty \leqslant \Delta\varphi$, $\zeta \leqslant 1$. Hence $\lambda \leqslant |\beta^*|_{\min}/(2\varphi) \leqslant 2|\beta^*|_{\min}/\{(3 + \zeta)\varphi\}$. Under the events $\eta_1 \leqslant \epsilon$ and $\|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty \leqslant \epsilon$ we have $L_1 > 0$. Therefore,

$$\text{pr}(L_1 > 0) \geqslant 1 - 2s^2\exp(-c_1ns^{-2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2).$$

Thus (14) is proven.

We now prove conclusion 3. By (A8) and $\eta_1\varphi < 1$, we have

$$\|\hat{\beta}_A - \beta^*\|_\infty \leqslant (1 - \eta_1\varphi)^{-1}\{\lambda\varphi/2 + \|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty\varphi + \varphi^2\eta_1\Delta\}.$$

Under the events $\eta_1 < \epsilon$ and $\|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty \leqslant \epsilon$ we have $\|\hat{\beta}_A - \beta^*\|_\infty \leqslant 4\varphi\lambda$. Thus,

$$\text{pr}(\|\hat{\beta}_A - \beta^*\|_\infty \leqslant 4\varphi\lambda) \geqslant 1 - 2s^2\exp(-c_1ns^{-2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2).$$

Thus (15) is proven. □

Proof of Theorem 2. Theorem 2 directly follows from Theorem 1, so its proof is omitted. □

REFERENCES

ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., MACK, S. & LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.* **96**, 6745–50.

BICKEL, P. J. & LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

BICKEL, P. J. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

CAI, T., ZHANG, C. H. & ZHOU, H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, 2118–44.

DETTLING, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583–93.

EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge: Cambridge University Press.

EFRON, B., HASTIE, T. J., JOHNSTONE, I. M. & TIBSHIRANI, R. J. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.

FAN, J. & FAN, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–37.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

HALL, P., MARRON, J. S. & NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

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

HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2008). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Berlin: Springer.

LEVINA, E., ROTHMAN, A. J. & ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Statist.* **2**, 245–63.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

MICHIE, D., SPIEGELHALTER, D. J. & TAYLOR, C. C. (1994). *Machine Learning, Neural and Statistical Classification*, Chichester: Ellis Horwood.

ROTHMAN, A. J., BICKEL, P., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V., RICHIE, J. P., ET AL. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–9.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

TIBSHIRANI, R. J., HASTIE, T. J., NARASIMHAN, B. & CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.* **99**, 6567–72.

TIBSHIRANI, R. J., SAUNDERS, M., ROSSET, S., ZHU, J. & KEITH, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B **67**, 91–108.

TRENDAFILOV, N. T. & JOLLIFFE, I. T. (2007). DALASS: Variable selection in discriminant analysis via the lasso. *Comp. Statist. Data Anal.* **51**, 3718–36.

WAINWRIGHT, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Info. Theory* **55**, 2183–202.

WITTEN, D. M. & TIBSHIRANI, R. J. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc.* B **73**, 753–72.

WU, M. C., ZHANG, L., WANG, Z., CHRISTIANI, D. C. & LIN, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics* **25**, 1145–51.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.

ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

ZOU, H. & HASTIE, T. J. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received February* 2011. *Revised September* 2011]