A Direct Approach to Sparse Discriminant Analysis in Ultra-high Dimensions

Qing Mai, Hui Zou and Ming Yuan

University of Minnesota

Georgia Institute of Technology

This version:July 22, 2011

Abstract

Sparse discriminant methods based on independence rules, such as the nearest shrunken centroids classifier (Tibshirani et al. 2002) and features annealed independent 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 selections and inferior classifications. We propose a new recipe for sparse discriminant analysis, motivated by least squares formulation of linear discriminant analysis. To demonstrate our proposal, we study the numerical and theoretical properties of discriminant analysis constructed via Lasso/SCAD penalized least squares. Our theory shows that both the proposed methods 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 our methods compare favorably with other popular sparse discriminant proposals in the literature.

Keywords: Discriminant analysis, Feature selection, High-dimensional data, Lasso, SCAD, Nearest shrunken centroid classifier, NP-dimension asymptotics.

1 Introduction

Consider a binary classification problem where $\mathrm{x} = (x_1, \ldots, x_p)$ represents the predictor vector and $G = 1, 2$ denotes the class label. Linear discriminant analysi(LDA) is perhaps the oldest classification technique that is still being used routinely in real world applications. The linear discriminant analysis model assumes $\mathrm{x} \mid (G = g) \sim N(\mu_g, \Sigma), \Pr(G = 1) = \pi_1, \Pr(G = 2) = \pi_2$. 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

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

Let $\hat{\mu}_1$, n_1 and $\hat{\mu}_2$, n_2 be the sample mean vector and sample size within class 1 and class 2, respectively. Let $\hat{\Sigma}$ be the sample estimate of Σ. To implement the Bayes rule, linear discriminant analysis substitutes $\mu_1 = \hat{\mu}_1$, $\mu_2 = \hat{\mu}_2$, $\Sigma = \hat{\Sigma}$, $\pi_1 = n_1/n, \pi_2 = n_2/n$ in (1). Despite its simplicity, linear discriminant analysis has been proven 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 rapid advance of technology, high dimensional data appear more and more frequently in contemporary statistical problems, such as tumor classification using microarray data. In such data the dimension (p) can be much larger than the sample size (n). It has been empirically observed by many that for classification problems with high-dimension-and-low-sample-size data some simple linear classifiers perform as well as much 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) provides some geometric insight into this interesting phenomenon. In recent years, many papers have considered ways to modify the usual LDA such that the modified discriminant analysis method is suitable for high dimensional classification. A seemingly obvious choice is by using more sophisticated estimates of the inverse covariance matrix Σ^{-1} to replace the naive sample estimate. Under some sparsity assumption, 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. Consider an ideal scenario where we know Σ is an identity matrix. Even so, 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, μ_2. Therefore, effectively exploiting sparsity is critically

important for high-dimensional classification.

Tibshirani et al. (2002) proposed the nearest shrunken centroid classifier(NSC) for tumor classification and gene selection using microarray data. The shrunken centroid classifier is defined as follows. For each variable x_j, we compute $d_{j1} = \frac{n}{n_1 n_2}\frac{\hat{\mu}_{j1} - \hat{\mu}_{j2}}{s_j + s_0}$ and $d_{j2} = -d_{j1}$, where $\hat{\mu}_{j1}$ and $\hat{\mu}_{j2}$ are 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 consideration. For simplicity, we can think $s_0 = 0$. Define the shrunken centroid mean by

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

where $\bar{x}_j = \frac{n_1\hat{\mu}_{1j} + n_2\hat{\mu}_{2j}}{n}$ 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}: $d_{jg}^{\lambda} = \text{sign}(d_{jg})(|d_{jg}| - \lambda)_+, g = 1, 2$. The NSC classifies x to class 2 if

$$\sum_{j=1}^{p} \left(x_j - \frac{(\hat{\mu}'_{j2} + \hat{\mu}'_{j1})}{2} \right) \frac{(\hat{\mu}'_{j2} - \hat{\mu}'_{j1})}{s_j^2} + \log \frac{n_2}{n_1} > 0. \tag{2}$$

Comparing (2) and (1), we see that the nearest shrunken centroid classifier modifies the usual LDA in two directions. First, it only uses the diagonal sample covariance matrix to estimate Σ. If $\lambda = 0$, NSC reduces to the so-called diagonal linear discriminant analysis. As shown in Bickel & Levina (2004), the diagonal linear discriminant analysis may work much better than the usual LDA in high dimensions. Second, NSC classifier uses the shrunken centroid mean to estimate μ_1, μ_2 in order to perform feature selection. Note that if we use a sufficiently large λ, then the soft-thresholding operation will force $\mu'_{j1} = \hat{\mu}'_{j2} = \bar{x}_j$ for some variables and those variables have no contribution to the classifier defined in (2). NSC is implemented in the `R` package `pamr` written by Hastie, Tibshirani, Narasimhan and Chu. See http://cran.r-project.org/web/packages/pamr/index.html. Many empirical experiments have shown that NSC is very competitive for high-dimensional classification. Variants of the shrunken centroid idea have been considered in other sparse discriminant analysis proposals (Guo et al. 2006, Wang & Zhu 2007). More recently, Fan & Fan (2008) proposed features annealed independence rules in which gene 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 genes/features that contribute most to classification, the target of an ideal feature selection should be the discriminative set which contains all "discriminative genes" that contributes to the Bayes rule. This is

a very natural argument 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 genes/features. The performance of feature selection by a sparse discriminative method is measured by its probability of discovering the discriminative set. There is little theoretical work for justifying NSC and its variants. To our knowledge, only Fan & Fan (2008) provided some detailed theoretical analysis of features annealed independent rules(FAIR), where the fundamental assumption is that Σ is a diagonal matrix. However, such an assumption is too restrictive to hold in real applications, because strong correlations exist in microarrays and other types of high-dimensional data. It is not hard to see that ignoring the important correlation structure may lead to misleading feature selection results. In fact, we argue that both NSC and FAIR aim to discover the so-called signal set whose definition is given explicitly in Section 2. We further provide a necessary and sufficient condition under which the signal set is identical to the discriminative set. The necessary and sufficient condition can be easily violated and hence independent rules could select wrong features.

In this work we propose a new recipe for sparse discriminant analysis in high dimensions. Our proposal is motivated by the well-known fact that in the traditional low dimension setting the LDA classifier can be reconstructed exactly via least squares (Hastie et al. 2008). We suggest using penalized sparse least squares methods to derive sparse discriminant methods. Our proposal is computationally efficient in high dimensions with the help of efficient algorithms for computing penalized least squares. We further provide theoretical justifications for our proposal. Suppose the Bayes rule has a sparse representation. Our theoretical results show that the proposed sparse discriminant 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.

The rest of the paper is organized as follows. In Section 2 we discuss the differences between the signal set and the discriminative set. In Section 3 we introduce the penalized least squares formulation of sparse discriminant analysis. In Section 4 we establish the theoretical properties of Lasso-LDA and SCAD-LDA, where the Lasso penalty and the SCAD penalty are used to do feature selection. Numerical results are presented in Section 5. Technical proofs are relegated to an Appendix.

2 The signal set and the discriminative set

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

We begin with some necessary notation. 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 informative or discriminative variables. Define $\tilde{A} = \{j : \mu_{1j} \neq \mu_{2j}\}$ which is referred to as the signal set and variables in $\tilde{A}$ are called signals. Independent rules select genes by comparing their within-class means. In an ideal situation, $\tilde{A}$ is the gene selection outcome of an independent rule.

Finding $\tilde{A}$ is of course an interesting and valid statistical inference problem, which is often formulated as a multiple hypothesis testing problem (Dudoit & Van der Laan 2008, Efron 2010). Various new methods and theories have been developed for doing thousands of hypotheses testing at the same time. See Benjamini & Hochberg (1995), Storey (2002), Storey et al. (2004), Genovese & Wasserman (2004), Efron (2005), Donoho & Jin (2004), Sun & Cai (2007), among others. Efron (2009) discussed the connection between large-scale classification and large-scale testing under a special LDA model assuming a diagonal covariance matrix. When Σ is diagonal $A = \tilde{A}$. For a general covariance matrix, however, the informative set and the signal set can be very different, as shown in the following proposition.

Proposition 1. *Let us decompose* Σ *as* $\Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,A^c} \\ \Sigma_{A^c,A} & \Sigma_{A^c,A^c} \end{pmatrix}$ *and* $\Sigma = \begin{pmatrix} \Sigma_{\tilde{A},\tilde{A}} & \Sigma_{\tilde{A},\tilde{A}^c} \\ \Sigma_{\tilde{A}^c,\tilde{A}} & \Sigma_{\tilde{A}^c,\tilde{A}^c} \end{pmatrix}$.

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

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

Based on Proposition 1 it is very easy to construct concrete examples to show that a non-signal can be informative, and vice versa. Here are two examples. Consider a LDA model with $\mu_1 = 0_p$, $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = 0.5$, $1 \leq i, j \leq 25$ and $i \neq j$, where $p = 25$. If $\mu_2 = (1, 1, 1, 1, 1, 0, \cdots, 0)^T$, then $\tilde{A} = \{1, 2, 3, 4, 5\}$ and $A = \{j : j = 1, \cdots, 25\}$, i.e., all variables are informative. Similarly, if let $\mu_2 = (3, 3, 3, 3, 3, 2.5, \cdots, 2.5)^T$, then all variables are signals but $A = \{1, 2, 3, 4, 5\}$.

The above arguments warn us that sparse discriminant analysis using independent rules could end up with a wrong set of features. A different sparse discriminant analysis method was recently proposed by Wu et al. (2008). Their proposal starts with Fisher's view of LDA, that is, the LDA direction is obtained by maximizing $\beta^T\hat{B}\beta/\beta^T\hat{\Sigma}\beta$ where $\hat{\Sigma}$ is the within covariance matrix and $\hat{B} = (\hat{\mu}_2 - \hat{\mu}_1)^T(\hat{\mu}_2 - \hat{\mu}_1)$ is the between covariance matrix. Note that $\beta^T\hat{B}\beta = \|(\hat{\mu}_2 - \hat{\mu}_1)^T\beta\|_2^2$. Wu et al. (2008) proposed the following sLDA:

$$\min \beta^T\hat{\Sigma}\beta \quad \text{s.t.} \quad (\hat{\mu}_2 - \hat{\mu}_1)^T\beta = 1 \text{ and } \|\beta\|_1 \leq \tau. \tag{3}$$

Witten & Tibshirani (2011) proposed another ℓ_1-penalized linear discriminant analysis:

$$\max\{\beta^T\hat{B}\beta - \lambda\sum_{j=1}^{p}|s_j\beta_j|\}, \text{subject to } \beta^T\hat{\Sigma}\beta \leq 1. \tag{4}$$

Little is known about the theoretical properties of the estimators defined in (3) and (4). However, it is not our interest in this paper to prove or disprove these two methods, although we do include them in our numerical experiments.

3 Methodology

3.1 Sparse LDA via penalized least squares

Our approach to sparse LDA is motivated by the intimate connection between linear discriminant analysis and least squares in the classical $p < n$ setting (Hastie et al. 2008). Suppose we numerically code the class labels as $y_1 = -n/n_1$ and $y_2 = n/n_2$ where $n = n_1 + n_2$. Let

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

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

The connection is lost in high dimensional problems because the sample covariance estimate is no longer invertible and the LDA direction is not well defined in its original form. However, we may consider a penalized least squares formulation to produce a classification direction. Let $P_\lambda(\cdot)$ be a generic sparsity-inducing penalty. Specific choices of $P_\lambda(\cdot)$ are given in Section 2.2. We first compute the solution to a penalized least squares problem

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

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

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

It is important to note that $\hat{\beta}_0$ in (7) differs from $\hat{\beta}_0^\lambda$ in (6). In the $p \ll n$ case, consider the OLS estimator and the usual LDA. Let us write $\hat{\beta}^{\mathrm{ols}} = c\hat{\beta}^{\mathrm{LDA}}$, $\hat{\beta}^{\mathrm{LDA}} = \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$. We should use $\hat{\beta}_0 = c\hat{\beta}_0^{\mathrm{LDA}}$ in (7) where

$$\hat{\beta}_0^{\mathrm{LDA}} = \log\left(\frac{n_2}{n_1}\right) - \left(\frac{\hat{\mu}_1 + \hat{\mu}_2}{2}\right)^T \hat{\beta}^{\mathrm{LDA}}$$

such that the OLS classifier and the LDA rule yield identical classification. If we use $\hat{\beta}_0^{\mathrm{ols}}$ in (7), the the OLS classifier is in general not identical to the LDA rule.

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. We show here that for a given classification direction, there is a nice closed-form formula for the optimal intercept.

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

$$\tilde{\beta}_0^{opt.} = -\frac{1}{2}(\mu_1 + \mu_2)^T\tilde{\beta} + \frac{\tilde{\beta}^T\Sigma\tilde{\beta}}{(\mu_2 - \mu_1)^T\tilde{\beta}}\log\frac{\pi_2}{\pi_1}, \tag{8}$$

which can be estimated by

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

Without sparsity condition on $\tilde{\beta}$, the estimator given in (9) would not work well when $p > n$. However, when $\tilde{\beta}$ is sparse, we have $\tilde{\beta}^T\Sigma\tilde{\beta} = \sum_{i,j:\tilde{\beta}_i\neq 0,\tilde{\beta}_j\neq 0}\Sigma_{ij}\tilde{\beta}_i\tilde{\beta}_j$ and $\tilde{\beta}^T\hat{\Sigma}\tilde{\beta} = \sum_{i,j:\tilde{\beta}_i\neq 0,\tilde{\beta}_j\neq 0}\hat{\Sigma}_{ij}\tilde{\beta}_i\tilde{\beta}_j$. Even when $p \gg n$, as long as $\|\tilde{\beta}\|_0 \ll n$, $\tilde{\beta}^T\hat{\Sigma}\tilde{\beta}$ is a good estimator for $\tilde{\beta}^T\Sigma\tilde{\beta}$. 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 the issue of estimating large covariance matrices with the problem of feature selection.

The condition $(\mu_2 - \mu_1)^T\tilde{\beta} > 0$ in proposition 2 is very mild. Suppose the linear classifier actually yields $(\mu_2 - \mu_1)^T\tilde{\beta} < 0$, then it is easy to show that such a classifier is dominated by the other linear classifier using direction $\tilde{\beta}_{new} = -\tilde{\beta}$ that obeys $(\mu_2 - \mu_1)^T\tilde{\beta}_{new} > 0$.

By proposition 2 , the sparse LDA classifier is defined as follows: assigning x to class 2 if

$$\left(\mathrm{x} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right)^T \hat{\beta}^\lambda + \frac{(\hat{\beta}^\lambda)^T \hat{\Sigma} \hat{\beta}^\lambda}{(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta}^\lambda} \log \frac{n_2}{n_1} > 0. \tag{10}$$

The sparse LDA classifier depends on the regularization parameter λ. In practice we need to select a good regularization parameter such that the generalization error is as small as possible. Cross-validation is a popular method for tuning. In this paper we use cross-validation to select λ under the 0-1 loss.

3.2 Choice of penalty and computing algorithm

We now discuss the choice of penalty functions in (6). We note first that our sparse LDA approach can work with any sparsity-inducing penalty function. In recent years, many papers have been devoted to designing nice penalty functions for sparse regression. Some well-known examples are lasso (Tibshirani 1996), SCAD (Fan & Li 2001), elastic net (Zou & Hastie 2005), fused lasso (Tibshirani et al. 2005), grouped lasso (Yuan & Lin 2006), adaptive lasso (Zou 2006), MCP (Zhang 2010) and SICA (Lv & Fan 2009), among others. Fan & Lv (2010) provided a good review on feature selection and penalized regression models. Roughly speaking, these penalty functions can be classified into two categories: the convex family and the concave family, with lasso and SCAD being the representing examples. To fix idea, we focus on the lasso and SCAD penalties when constructing sparse LDA classifiers. The lasso penalty function is $P_\lambda(t) = \lambda t$ for $t \geq 0$. The SCAD penalty function is defined by $P_{\lambda,a}(0) = 0$ and $P'_{\lambda,a}(t) = \lambda I(t \leq \lambda) + \frac{(a\lambda - t)_+}{a-1} I(t > \lambda)$ for $t > 0$ where $a > 2$. Following Fan and Lv (2001), we used $a = 3.7$ in our numerical experiments. If the lasso penalty is used in (6), we call the resulting classifier Lasso-LDA. Likewise, if the SCAD penalty is used, we call the resulting classifier SCAD-LDA.

There has been considerable progress in developing efficient algorithms for computing sparse regularized regression models in high dimensions. The LARS algorithm (Efron et al. 2004), implemented in the `R` package `lars`, computes the entire solution path for the lasso regression with the same order of computational cost as a single ordinary least squares fit. Friedman et al. (2008) implemented the coordinate descent algorithm for computing the lasso regression in the `R` package `glmnet` and showed that `glmnet` can be even faster than `lars`. Zou & Li (2008) showed that using the LLA algorithm one can solve the concave penalized regression problem via an iterative weighted-lasso regression procedure. One could combine the LLA algorithm and `glmnet` to solve any concave penalized least squares for each fixed

λ. For the SCAD penalty, it turns out that an even faster algorithm is possible by directly applying the coordinate descent principle. The coordinate decent algorithm works well for the lasso because the univariate lasso regression solution is given by the soft-thresholding rule (Tibshirani 1996, Friedman et al. 2008). Likewise, the univariate SCAD solution also has a closed-form formula given by the SCAD thresholding rule (Fan & Li 2001). Therefore, with some proper modification, `glmnet` can be used to compute the SCAD penalized regression. In a word, both Lasso-LDA and SCAD-LDA can be computed very efficiently even when $p \gg n$. Hence they are practically useful for high dimensional classification problems.

4 Statistical Theory

In this section we study the theoretical properties of the sparse LDA classifiers based on lasso/SCAD penalized least squares. In the literature there are many results on sparse penalized least squares (Fan & Li 2001, Zou 2006, Zhao & Yu 2006, Zhang & Huang 2008, Zhang 2010, Fan & Lv 2008, Lv & Fan 2009). But they cannot be directly applied to our setting although we borrow the least squares criterion to derive the sparse LDA classifier, because the linear model assumption ($y = \sum_j x_j\beta_j + error$), the foundation for these existing theoretical work, does not hold for the LDA model. Furthermore, if we regard the predictor matrix as the "design" matrix, then our theory always deals with the random design case, whereas the fixed design theory is common in the existing work on high-dimensional penalized least squares regression.

4.1 Notation and definitions

We first introduce some necessary notation to be used in the theoretical analysis. For a general $m \times n$ matrix M, define $\|M\|_\infty = \max_{i=1,..,m} \sum_{j=1}^n |M_{ij}|$. 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 $A = \{j : \beta(\text{Bayes})_j \neq 0\}$ and let $s = |A|$. 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:

$$\kappa = \|C_{A^cA}(C_{AA})^{-1}\|_\infty, \tag{11}$$

$$\varphi = \|(C_{AA})^{-1}\|_\infty, \tag{12}$$

$$\Delta = \|\mu_{2A} - \mu_{1A}\|_\infty. \tag{13}$$

Suppose X is the predictor matrix and let $\tilde{X}$ be the centered predictor matrix such that the column-wise mean is zero. Obviously, $C^{(n)} = \frac{1}{n}\tilde{X}^T\tilde{X}$ is an empirical version of C. Likewise, we can write $\frac{1}{n}\tilde{X}_A^T\tilde{X}_A = C_{AA}^{(n)}$ and $\frac{1}{n}\tilde{X}_{A^c}^T\tilde{X}_A = C_{A^cA}^{(n)}$.

Denote $\beta^* = (C_{AA})^{-1}(\mu_{2A} - \mu_{1A})$. Now we can define $\tilde{\beta}(\text{Bayes})$ by letting $\tilde{\beta}(\text{Bayes})_A = \beta^*$ and $\tilde{\beta}(\text{Bayes})_{A^c} = 0$. The following is a simple but very useful result, showing the equivalence between $\tilde{\beta}(\text{Bayes})$ and $\beta(\text{Bayes})$ in the context of LDA model.

Proposition 3. *$\tilde{\beta}(Bayes)$ and $\beta(Bayes)$ are equivalent in the sense that $\tilde{\beta}(Bayes) = c\beta(Bayes)$ for some positive constant c and the Bayes classifier is also equivalent to assigning* x *to class 2 if*

$$(\mathrm{x} - \frac{\mu_1 + \mu_2}{2})^T\tilde{\beta}(Bayes) + \frac{(\tilde{\beta}(Bayes))^T\Sigma\tilde{\beta}(Bayes)}{(\mu_2 - \mu_1)^T\tilde{\beta}(Bayes)}\log\frac{\pi_2}{\pi_1} > 0. \tag{14}$$

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

4.2 Main results

We now present the main theoretical results. In our analysis we assume the variance of each variables is bounded by a finite constant. This regularity condition usually holds. In practice, one often standardizes the data beforehand. Then the finite constant can be taken as one. In this subsection, ϵ_0 and c_1, c_2 are some positive constants.

Suppose the Lasso-LDA estimator does find the support of the Bayes rule, A, then we have $\hat{\beta}(\text{lasso})_{A^c} = 0$ and $\hat{\beta}(\text{lasso})_A$ should be identical to $\hat{\beta}_A$, where

$$\hat{\beta}_A = \arg\min_{\beta,\beta_0} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j\in A} x_{ij}\beta_j)^2 + \sum_{j\in A}\lambda|\beta_j|. \tag{15}$$

We introduce $\hat{\beta}_A$ only for mathematical analysis. It is not a real estimator, because its definition depends on knowing A.

To ensure the Lasso-LDA classifier has the variable selection consistency property, we impose a condition on the covariance matrix of the predictors:

$$\kappa = \|C_{A^cA}(C_{AA})^{-1}\|_\infty < 1. \tag{16}$$

The above condition is an analogue of the irrepresentable condition for the Lasso regression estimator (Zhao & Yu 2006, Zou 2006).

Theorem 1 (Analysis of Lasso-LDA)**.** *Pick any* λ *such that* $\lambda < \min(\frac{1}{2}|\beta^*|_{min}/\varphi, \Delta)$.

1. *Assuming the condition in (16), with probability at least* $1 - \delta_1$, $\hat{\beta}_A(lasso) = \hat{\beta}_A$ *and* $\hat{\beta}_{A^c}(lasso) = 0$, *where*

$$\delta_1 = 2ps\exp(-\frac{n}{s^2}\epsilon^2 c_1) + 2p\exp(-nc_2(\frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa})^2) \tag{17}$$

   and ϵ *is any positive constant satisfying* $\epsilon < \min(\epsilon_0, \frac{\frac{\lambda}{4\varphi}(1-\kappa)}{\frac{\lambda}{2}+(1+\kappa)\Delta})$.

2. *With probability at least* $1 - \delta_2$, *none of the elements of* $\hat{\beta}_A$ *is zero, where*

$$\delta_2 = 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) + 2s\exp(-nc_2\epsilon^2), \tag{18}$$

   where ϵ *is any positive constant satisfying* $\epsilon < \min(\epsilon_0, \frac{1}{\varphi}\frac{\zeta}{3+\zeta})$.

3.

$$\Pr(\|\hat{\beta}_A - \beta^*\|_\infty \le 4\varphi\lambda) \;\; \ge \;\; 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2), \tag{19}$$

   where ϵ *is any positive constant satisfying* $\epsilon < \min(\epsilon_0, \frac{\lambda}{2\varphi\Delta}, \lambda)$.

In our analysis we compare SCAD-LDA to an oracle estimator knowing the true feature set A. We first define

$$\tilde{\beta}(\text{oracle})_A = (C_{AA}^{(n)})^{-1}(\mu_{2A} - \mu_{1A}). \tag{20}$$

It is easy to see that $\tilde{\beta}(\text{oracle})_A$ is the solution to a least squares criterion $\min\sum_{i=1}^n(y_i - \beta_0 - \sum_{j\in A}x_{ij}\beta_j)^2$. Hence, in terms of classification $\tilde{\beta}(\text{oracle})$ is equivalent to the oracle LDA only using the subset A. The oracle estimator is then defined as $\hat{\beta}(\text{oracle})$ such that $\hat{\beta}(\text{oracle})_A = \tilde{\beta}(\text{oracle})_A$ and $\hat{\beta}(\text{oracle})_{A^c} = 0$.

Theorem 2 (Analysis of SCAD-LDA). *1. For any $\epsilon > 0$ satisfying*

$$\frac{\epsilon}{\epsilon + 2\Delta\varphi} \leq \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta}),$$

we have

$$\begin{aligned} & \Pr(\|\hat{\beta}(oracle) - \beta^*\|_\infty \geq \epsilon) \\ \leq\ & 2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2}) + 2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}). \end{aligned} \tag{21}$$

2. For any $\lambda < \frac{|\beta^*|_{min}}{a}$, with probability at least $1-\delta_3$, none of the elements of $\hat{\beta}(oracle)_A$ is zero and $\hat{\beta}(oracle)$ is a local solution to the SCAD-LDA criterion, where

$$\begin{aligned} \delta_3 \ =\ & 2p\exp(-nc_2\epsilon^2) + 2ps\exp(-\frac{nc_1}{s^2}\frac{1}{\varphi^2}(\frac{\epsilon}{\epsilon+\kappa+1})^2) \\ & +2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2}) + 2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}), \end{aligned} \tag{22}$$

where ϵ is any positive constant, $\epsilon < \epsilon_0$ and obeys the following constraints: $\frac{\epsilon}{\epsilon+\kappa+1} < \varphi\epsilon_0$, $\frac{\epsilon}{\epsilon+2\Delta\varphi} \leq \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta})$, and $\epsilon < \min(|\beta^*|_{min} - a\lambda, \frac{\lambda}{6\Delta}, \frac{\lambda}{6}, \frac{\lambda}{6\kappa+\frac{\lambda}{\Delta}})$.

The analysis of SCAD-LDA does not require condition (16). Theorem 2 works for any positive κ.

The non-asymptotic results in Theorems 1 and 2 can be easily translated into some asymptotic arguments when considering the triple of (n, s, p) goes to infinity at some proper rates. To highlight the main points, we assume Δ, κ, φ are constants in the asymptotic arguments. In addition, we need the following two regularity conditions:

(C1). $n, p \to \infty$ and $\log(ps)s^2/n \to 0$,

(C2). $|\beta^*|_{min} \gg \sqrt{\frac{\log(ps)s^2}{n}}$.

Condition (C1) puts some restriction on p. Clearly, we cannot expect the proposed method (or any sensible method) works for an arbitrarily large p. However, the restriction is rather loose. Consider the case where $s = o(n^{\frac{1}{2}-\gamma})$ for some $\gamma < \frac{1}{2}$. (C1) holds as long as $p \ll e^{n^{2\gamma}}$. Therefore, p is allowed to grow faster than any polynomial order of n. This implies the applicability of the Lasso-LDA and SCAD-LDA to real world problems such as gene expression classification.

Condition (C2) requires the non-zero elements of the Bayes rule to be large enough such that we could consistently separate them from zeros by using observed data. The lower

bound actually converges to zero asymptotically under (C1), and hence condition (C2) is not a strong assumption.

Theorem 3 (Asymptotic properties of Lasso-LDA and SCAD-LDA)**.** *Under conditions (C1) and (C2), if we choose some* $\lambda = \lambda_n$ *such that* $\lambda_n \ll |\beta^*|_{min}$ *and* $\lambda_n \gg \sqrt{\log(ps)s^2/n}$*, then , with probability going to one, a SCAD-LDA solution is identical to the oracle LDA that is consistent in feature selection and* $\|\hat{\beta}(oracle)_A - \beta^*\|_\infty = o_P(\sqrt{\log(s)s^2/n})$*. Moreover, if we further assume* $\kappa < 1$*, then the Lasso-LDA is consistent in feature selection and* $\|\hat{\beta}(oracle)_A - \beta^*\|_\infty = o_P(\lambda_n)$*.*

Finally, it is important to point out that our theory does not require any structure assumption on the common covariance matrix Σ, which clearly shows the fundamental difference between our method and 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, Rothman et al. 2008). Such assumptions are not needed in our method.

5 Numerical Results

5.1 Simulation

We use simulated data to demonstrate the good performance of Lasso-LDA and SCAD-LDA. For comparison, we included NSC, FAIR, the sparse LDA proposed by Witten & Tibshirani (2011) and the sparse LDA proposed by Wu et al. (2008). NSC is implemented in the R package `pamr`; see http://cran.r-project.org/web/packages/pamr/index.html. The sparse LDA proposed by Witten & Tibshirani (2011) is implemented in the R package `penalizedLDA`; see http://cran.rproject.org/web/packages/penalizedLDA/index.html. We used the code by Dr. Wu to implement their proposal of sparse LDA.

We randomly generated n class labels such that $\pi_1 = \pi_2 = 0.5$. 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, μ_2, Σ and β^{Bayes} are shown in Table 1. Models 1-4 are sparse discriminant models with different covariance and mean structure, while models 5 and 6 are "practically sparse" in the sense that their Bays rules depend on all variables in theory but can be well approximated

Model	n	p	Σ	β^{Bayes}
1	100	400	Σ_{ij} =$0.5^{\|i-j\|}$	$0.556(3, 1.5, 0, 0, 2, 0_{p-5})^T$
2	100	400	Σ_{ij} =$0.5^{\|i-j\|}$	$0.582(3, 2.5, -2.8, 0_{p-3})^T$
3	400	800	$\Sigma_{jj} = 1, \Sigma_{ij}$=0.5, $i \neq j$.	$0.395(3, 1.7, -2.2, -2.1, 2.55, 0_{p-5})^T$
4	300	800	$\Sigma_{jj} = 1, \Sigma_{ij} = 0, \|i-j\| \geq 160$ Σ_{ij}=0.6, $0 < \|i-j\| < 160$	$0.916(1.2, -1.4, 1.15, -1.64, 1.5, -1, 2, 0_{p-7})^T$
5	400	800	$\Sigma_{jj} = 1, \Sigma_{ij}$=0.5, $i \neq j$.	$0.551(3, 1.7, -2.2, -2.1, 2.55, (p-5)^{-1}1_{p-5})^T$
6	400	800	$\Sigma_{jj} = 1, \Sigma_{ij}$=0.5, $i \neq j$.	$0.362(3, 1.7, -2.2, -2.1, 2.55, (p-5)^{-1}1_{p-5})^T$

Table 1: Simulation settings.

by sparse discriminant functions. Table 2 summarizes the simulation results based on 2000 replications. For each measure we reported its median and the corresponding standard error in parentheses. Only Lasso-LDA and SCAD-LDA show consistently good performance in all six simulation settings. They closely mimic the Bayes rule, regardless of the Bayes error and covariance structure. NSC and FAIR have very comparable performance, but they are much worse than Lasso-LDA and SCAD-LDA except in model 1. By direct calculation one can see that 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 very different mean structure: the first two elements of $\mu_2 - \mu_1$ are dominating while the rest are much smaller. This means that independence rules have difficulty in selecting variable three, resulting inferior classification. Wu's method has good classification accuracy overall, but it can often miss some important features. Witten's method has rather poor performance, which is somewhat surprising because the basic idea behind Witten's method is similar to Wu's. We do notice that Witten's formulation in (4) is nonconvex while Wu's formulation in (3) and is convex, which may help explain their different performance.

	Bayes rule	Lasso-LDA	SCAD-LDA	Wu	Witten	NSC	FAIR
Model 1							
Error(%)	10	10.89	11.39	13.71	10.81	10.94	11.47
		(0.03)	(0.04)	(0.01)	(0.01)	(0.02)	(0.05)
TRUE Selection	3	3	3	3	1	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE Selection	0	2	0	0	26	6	7
		(0.16)	(0)	(0.49)	(0.11)	(0.61)	(0.66)
Model 2							
Error(%)	10	12.84	14.03	14.5	14.25	15.12	15.67
		(0.05)	(0.05)	(0.01)	(0.02)	(0.05)	(0.07)
TRUE Selection	3	3	3	1	2	2	2
		(0)	(0.48)	(0.14)	(0)	(0.34)	(0)
FALSE Selection	0	6	13	0	4	9	8
		(0.27)	(0.64)	(0)	(0.61)	(0.73)	(0.29)
Model 3							
Error(%)	20	21.93	21.37	22.37	33.69	27.48	25.69
		(0.03)	(0.03)	(0.05)	(0.01)	(0.07)	(0.02)
TRUE Selection	5	5	5	5	3	3	2
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE Selection	0	14	12	2	419.5	2	0
		(0.59)	(0.49)	(0)	(10.19)	(0.31)	(0)
Model 4							
Error(%)	10	12.50	12.12	13.99	23.90	19.25	18.56
		(0.02)	(0.03)	(0.03)	(0.01)	(0.04)	(0.00)
TRUE Selection	7	7	7	6	4	4	3
		(0)	(0.03)	(0)	(0)	(0)	(0)
FALSE Selection	0	18	15	2	35	1	0
		(0.70)	(0.42)	(0)	(4.43)	(0.48)	(0)
Model 5							
Error(%)	10	11.11	10.55	12.07	21.99	14.72	14.27
		(0.02)	(0.03)	(0.07)	(0.01)	(0.03)	(0.01)
Fitted model size	800	21	14	7	737	3	3
		(0.65)	(0.19)	(0.16)	(2.29)	(0.46)	(0)
Model 6							
Error(%)	20	22.22	21.62	23.34	30.43	26.13	24.14
		(0.03)	(0.03)	(0.05)	(0.01)	(0.07)	(0)
Fitted model size	800	20	16	5	592.5	8	3
		(0.53)	(0.31)	(0.49)	(7.46)	(0.51)	(0)

Table 2: Simulation results. The standard errors are reported in parentheses.

		Lasso-LDA	SCAD-LDA	Wu	Witten	NSC	FAIR
Colon	Error(%)	86.4	86.4	84.1	86.4	86.4	86.4
		(1.54)	(2.08)	(2.17)	(0.49)	(1.20)	(0.61)
	Fitted model size	5	6	1	10	89	11
		(0.63)	(0.60)	(0)	(1.39)	(29.95)	(1.19)
Prostate	Error(%)	94.1	91.2	91.2	91.2	91.2	76.5
		(0.55)	(1.37)	(0.70)	(0.24)	(0.96)	(0.54)
	Fitted model size	10	8	1	18	10	4
		(0.77)	(0.96)	(0)	(4.45)	(0.84)	(0.40)

Table 3: Real data.

		Wu	Witten	NSC	FAIR
Colon	Error(%)	86.4(1.06)	86.4(0.51)	63.6(0.70)	77.3(2.16)
Prostate	Error(%)	91.2(1.24)	94.1(1.25)	91.2(1.39)	73.5(1.11)

Table 4: Classification accuracies if we force all the methods to select similar numbers of genes as Lasso-LDA and SCAD-LDA.

5.2 Real data

We further compare the methods on two benchmark datasets: Colon and Prostate cancer data. The basic task here is to predict whether an observation is tumor or normal tissue. We randomly split the datasets into the training and test sets with $2:1$ ratio. Model fitting was done on the training set and the classification accuracy was evaluated on the test set. This procedure was repeated 100 times. Shown in Table 4 are the (median) classification accuracy and the number of selected genes by each competitor.

Colon and Prostate data 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 classification accuracy 92.5% and the nearest shrunken centroids classifier was the most accurate classifier for the Colon data. Table 3 shows that both Lasso-LDA and SCAD-LDA are as accurate as the nearest shrunken centroids classifier on the Colon data and Lasso-LDA significantly outperforms BagBoost on the Prostate data. Since BagBoost does not do gene selection, we do not include it in Table 3. Witten's method works quite well on these two real datasets.

Note that Lasso-LDA and SCAD-LDA select similar numbers of genes. If we force other methods to select similar numbers of genes, the results would be as listed in Table 4. Wu's and Witten's methods have improved performance, while NSC and FAIR have worse performance.

6 Discussion

Sparse discriminant analysis based on independence rules is computationally attractive for high dimensional classification. However, they may lead to misleading feature selection results and hence poor classification performance. Their limitation is due to the fundamental difference between discriminative and signal variables. When doing feature selection in classification, one should aim to recover the discriminative set not the signal set. Finding the signal set is the goal of large-scale hypothesis testing. We should point out that our arguments are not against the developments of theory and methodology for large-scale multiple hypothesis testing. Discovering "signals" is the fundamental question of research in many scientific studies. We only wish to warn the practitioners that the problem of identifying features for discrimination could be very different from identifying interesting signals, and hence the statistical tools for data analysis should be carefully chosen.

Built upon such insight, we have proposed a regularized least squares approach towards sparse LDA models. This approach is computationally efficient for handling high dimensional data. We have established some non-asymptotic theory for the Lasso/SCAD penalized LDA classifiers, from which NP-dimension asymptotic consistency results have been shown to hold for the Lasso and SCAD LDA classifiers. In addition, the numerical results are very promising, suggesting the great potential of the proposed sparse LDA classifiers for real world applications.

The regularized least squares can be flexibly modified to accommodate some specific goals. For instance, if we wish to conduct group-wise variable selection when the groups are clearly defined, then we could use the grouped lasso penalty (Yuan & Lin 2006). If we wish to impose certain smoothness structure to the classification coefficients, we could apply the fused lasso penalty (Tibshirani et al. 2005). In some situations the predictors may have a natural ordering where the ordered variable selection is preferred. For that, we could apply the hierarchical LARS algorithm (Yuan & Lin 2007) or the nested lasso penalty (Levina et al. 2008) to obtain the regularized least squares fit. Such generalizations are straightforward to implement, but the detailed treatment is out of the scope of this paper.

Appendix:proofs

Proof of Proposition 1. 1. Let $\Omega = \Sigma^{-1}$ and $\beta^{Bayes} = \Omega(\mu_2 - \mu_1)$. Write $\Omega = \begin{pmatrix} \Omega_{\tilde{A},\tilde{A}} & \Omega_{\tilde{A},\tilde{A}^c} \\ \Omega_{\tilde{A}^c,\tilde{A}} & \Omega_{\tilde{A}^c,\tilde{A}^c} \end{pmatrix}$. Note that $A \subseteq \tilde{A}$ is equivalent to $\beta^{Bayes}_{\tilde{A}^c} = 0$. On the other hand, we have $\beta^{Bayes}_{\tilde{A}^c} = \Omega_{\tilde{A}^c,\tilde{A}}(\mu_{2,\tilde{A}} - \mu_{1,\tilde{A}})$ and $\Omega_{\tilde{A}^c,\tilde{A}} = -(\Sigma_{\tilde{A}^c,\tilde{A}^c} - \Sigma_{\tilde{A}^c,\tilde{A}}\Sigma^{-1}_{\tilde{A},\tilde{A}}\Sigma_{\tilde{A},\tilde{A}^c})^{-1}\Sigma_{\tilde{A}^c,\tilde{A}}\Sigma^{-1}_{\tilde{A},\tilde{A}}$. Therefore, part 1 is proven.

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

Proof of Proposition 2. We recode the response variable as $y^* = 1, -1$. Note that $\tilde{\beta}^{opt.}_0 = \arg\min_{\tilde{\beta}_0} E(y^*_{new} \neq \text{sign}(\mathrm{x}^T_{new}\tilde{\beta} + \tilde{\beta}_0)|\text{training data})$. Since $y^*_{new}, \mathrm{x}_{new}$ are independent from the training data, $(Y^*_{new}, \mathrm{z}_{new} = \mathrm{x}^T_{new}\tilde{\beta})$ obeys a one-dimensional LDA model, that is,

$$\mathrm{z}_{new}|y^*_{new} = 1 \sim N(\tilde{\beta}^T\mu_2, \tilde{\beta}^T\Sigma\tilde{\beta}), \quad \Pr(y^*_{new} = 1) = \pi_2$$

and

$$\mathrm{z}_{new}|y^*_{new} = -1 \sim N(\tilde{\beta}^T\mu_1, \tilde{\beta}^T\Sigma\tilde{\beta}), \quad \Pr(y^*_{new} = -1) = \pi_1.$$

Then by some straightforward calculation we obtain (8). □

Proof of Proposition 3. It suffices to prove that, there exists a constant $c > 0$ such that $\tilde{\beta}(Bayes)_A = c\beta(Bayes)_A$. Note that $C_{AA} = \Sigma_{AA} + \pi_1\pi_2(\mu_{2A} - \mu_{1A})(\mu_{2A} - \mu_{1A})^T$ and $C_{AA}\tilde{\beta}(Bayes)_A = \mu_{2A} - \mu_{1A}$. Let $c = n\left(n - 2 + \pi_1\pi_2(\mu_{2A} - \mu_{1A})^T\Sigma^{-1}_{AA}(\mu_{2A} - \mu_{1A})\right)^{-1} > 0$ then we have $\tilde{\beta}(Bayes)_A = c\beta(Bayes)_A$. □

We now prove theorems 1, 2 and 3. The following two lemmas provide some useful concentration inequalities that are repeatedly used in the proof.

Lemma 1. *There exists some constants ϵ_0 and c_1, c_2 such that for any $\epsilon \leq \epsilon_0$ we have*

$$\Pr(|C^{(n)}_{ij} - C_{ij}| \geq \epsilon) \leq 2\exp(-n\epsilon^2 c_1), \tag{23}$$

for each (i, j) pair; and

$$\Pr(|(\hat{\mu}_{2j} - \hat{\mu}_{1j}) - (\mu_{2j} - \mu_{1j})| \geq \epsilon) \leq 2\exp(-n\epsilon^2 c_2). \tag{24}$$

for each j. Moreover, we have

$$\Pr(\|C_{AA}^{(n)} - C_{AA}\|_\infty \geq \epsilon) \leq 2s^2 \exp(-\frac{n}{s^2}\epsilon^2 c_1), \tag{25}$$

$$\Pr(\|C_{A^cA}^{(n)} - C_{A^cA}\|_\infty \geq \epsilon) \leq 2(p-s)s \exp(-\frac{n}{s^2}\epsilon^2 c_1), \tag{26}$$

$$\Pr(\|(\hat{\mu}_2 - \hat{\mu}_1) - (\mu_2 - \mu_1)\|_\infty \geq \epsilon) \leq 2p \exp(-n\epsilon^2 c_2), \tag{27}$$

$$\Pr(\|(\hat{\mu}_{2A} - \hat{\mu}_{1A}) - (\mu_{2A} - \mu_{1A})\|_\infty \geq \epsilon) \leq 2s \exp(-n\epsilon^2 c_2). \tag{28}$$

Lemma 2. *There exists some constants ϵ_0, c_1 such that for any $\epsilon \leq \min(\epsilon_0, \frac{1}{\varphi})$, we have*

$$\Pr(\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \geq \frac{(\kappa+1)\epsilon\varphi}{1-\varphi\epsilon}) \leq 2ps \exp(-\frac{n}{s^2}\epsilon^2 c_1). \tag{29}$$

Proof of Lemma 1. Note that inequalities in (25)–(28) can be obtained from (23)-(24) by simple union bounds. So we only prove (23) and (24). First, it is easy to see that $\Pr(|\hat{\mu}_{1j} - \mu_{1j}| \geq \epsilon|Y) \leq 2\exp(-n_1\frac{\epsilon^2}{2\sigma_j^2})$. Also, $n_1 \sim Bernoulli(n, \pi_1)$. Hence, $\Pr(|n_1 - \pi_1 n| \geq n\epsilon) \leq 2\exp(-nc_2'\epsilon^2)$ for some $c_2' > 0$. Therefore, $\Pr(|\hat{\mu}_1 - \mu_1| \geq \epsilon) \leq 2\exp(-n\frac{\pi_1}{2}\frac{\epsilon^2}{2\sigma_j^2}) + 2\exp(-nc_2'(\frac{\pi}{2})^2) \leq 2\exp(-nc_2^{(1)}\epsilon^2)$ for some small enough $c_2^{(1)}$ and $\epsilon > 0$. Similarly, we have $\Pr(|\hat{\mu}_2 - \mu_2| \geq \epsilon) \leq 2\exp(-nc_2^{(2)}\epsilon^2)$. Thus (24) holds.

To prove (23), note that $C_{ij}^{(n)} = \frac{1}{n}\sum_{k=1}^n x_{ki}x_{kj} - \bar{x}_i\bar{x}_j$. Since $\bar{x}_v = \hat{\pi}_1\hat{\mu}_{1v} + \hat{\pi}_2\hat{\mu}_{2v}$, for $v = i, j$, by the previous arguments, we know that there exists $c_1' > 0$ such that

$$\Pr(|\bar{x}_i\bar{x}_j - Ex_iEx_j| \geq \epsilon) \leq 2\exp(-nc_0^{''}\epsilon^2). \tag{30}$$

$\frac{1}{n}\sum_{k=1}^n x_{ki}x_{kj} - E(x_ix_j) = \sum_{l=1}^2 \frac{n_l}{n}\left(\frac{1}{n_l}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j|g=l)\right) + \sum_{l=1}^2 E(x_ix_j|g = l)(\frac{n_l}{n} - \pi_l)$ and $E(x_ix_j|g=l) = \Sigma_{ij} + \mu_{li}\mu_{lj}$ for $l = 1, 2$. Then it suffices to show that there exists some constant $c_1^{(l)}$ such that

$$\Pr\left(|\frac{1}{n_l}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j|g=l)| \geq \epsilon|Y\right) \leq 2\exp(-n_l c_1^{(l)}\epsilon^2). \tag{31}$$

We further have that $n^{-1}\sum_{k=1}^n x_{ki}x_{kj} - E(x_ix_j) = \sum_{l=1}^2 n_l/n\left\{n_l^{-1}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j \mid g=l)\right\} + \sum_{l=1}^2 E(x_ix_j \mid g=l)(n_l/n - \pi_l)$ 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 $\epsilon < \epsilon_0$,

$$\Pr(|n_l^{-1}\sum_{g_k=l}(x_{ki} - \mu_{li})(x_{kj} - \mu_{lj}) - \Sigma_{ij}| > \epsilon \mid Y) \leq 2\exp(-c_3 n\epsilon^2). \tag{32}$$

Combining the concentration results for $\hat{\mu}_{l\nu}$, n_l and (32), we have (23). □

Proof of Lemma 2. 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

$$\begin{aligned} & \|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \qquad (33)\\ \leq\ & \|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \cdot \|(C^{(n)}_{AA})^{-1} - (C_{AA})^{-1}\|_\infty + \|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \\ & + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \cdot \|C_{AA} - C^{(n)}_{AA}\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \\ & + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \cdot \|C_{AA} - C^{(n)}_{AA}\|_\infty \cdot \|(C^{(n)}_{AA})^{-1} - (C_{AA})^{-1}\|_\infty \\ \leq\ & (\kappa\eta_1 + \eta_2)(\varphi + \eta_3) \end{aligned}$$

Moreover,

$$\begin{aligned} \eta_3 &\leq \|(C^{(n)}_{AA})^{-1}\|_\infty \cdot \|(C^{(n)}_{AA} - C_{AA})\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \qquad (34)\\ &= (\varphi + \eta_3)\varphi\eta_1. \end{aligned}$$

So as long as $\varphi\eta_1 < 1$ we have $\eta_3 \leq \frac{\varphi^2\eta_1}{1-\varphi\eta_1}$ and hence

$$\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \leq \frac{(\kappa\eta_1 + \eta_2)\varphi}{1 - \varphi\eta_1}. \qquad (35)$$

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

With $y = \frac{n}{n_2}$ or $-\frac{n}{n_1}$ and the centered predictor matrix $\tilde{X}$, we can rewrite the Lasso-LDA estimator as

$$\hat{\beta}(lasso) = \arg\min_\beta \frac{1}{n}\beta^T(\tilde{X}^T\tilde{X})\beta - 2(\hat{\mu}_2 - \hat{\mu}_1)^T\beta + \lambda\sum_{j=1}^p |\beta_j|. \qquad (36)$$

Similar to the Lasso-LDA, the SCAD-LDA estimator can be written as

$$\arg\min \frac{1}{n}\beta^T(\tilde{X}^T\tilde{X})\beta - 2(\hat{\mu}_2 - \hat{\mu}_1)^T\beta + \sum_{j=1}^p P_{\lambda,a}(|\beta_j|), \qquad (37)$$

where $P_{\lambda,a}(\cdot)$ is the SCAD penalty function.

Proof of Theorem 1. Part (1). By definition we can write

$$\hat{\beta}_A = (\frac{1}{n}\tilde{X}_A^T\tilde{X}_A)^{-1}((\hat{\mu}_{2A} - \hat{\mu}_{1A}) - \frac{\lambda}{2}t_A) \qquad (38)$$

where t_A represents the so-called subgradient which is defined as

$$t_j = \begin{cases} \text{sign}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0, \\ t_j \in (-1, 1), & \text{if } \hat{\beta}_j = 0. \end{cases}$$

From (38) we can write

$$\begin{aligned}\hat{\beta}_A &= (C_{AA})^{-1}(\mu_{2A}-\mu_{1A})+(C_{AA}^{(n)})^{-1}((\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})) \\ &\quad -((C_{AA}^{(n)})^{-1}-(C_{AA})^{-1})(\mu_{2A}-\mu_{1A})-\frac{\lambda}{2}(C_{AA}^{(n)})^{-1}t_A.\end{aligned} \tag{39}$$

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

$$\|\frac{1}{n}\tilde{X}_{A^c}^T\tilde{X}_A\hat{\beta}_A-(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})\|_\infty\leq\frac{\lambda}{2}. \tag{40}$$

The left hand side of (40) is equal to

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}(\hat{\mu}_{2A}-\hat{\mu}_{1A})-C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}\frac{\lambda}{2}t_A-(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})\|_\infty. \tag{41}$$

Using $C_{A^cA}C_{AA}^{-1}(\mu_{2A}-\mu_{1A})=(\mu_{2A^C}-\mu_{1A^C})$, (41) has an upper bound:

$$\begin{aligned}U_1 &= \|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}-C_{A^cA}C_{AA}^{-1}\|_\infty\Delta+\|(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})-(\mu_{2A^C}-\mu_{1A^C})\|_\infty \\ &\quad +(\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}-C_{A^cA}C_{AA}^{-1}\|_\infty+\kappa)\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty \\ &\quad +(\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}-C_{A^cA}C_{AA}^{-1}\|_\infty+\kappa)\frac{\lambda}{2}\end{aligned} \tag{42}$$

Pick ϵ such that $\epsilon<\epsilon_0$ and $\epsilon<\frac{\frac{\lambda}{4\varphi}(1-\kappa)}{\frac{\lambda}{2}+(1+\kappa)\Delta}$. Check $\epsilon<\frac{1-\kappa}{2}\frac{1}{\varphi}$. If

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}-C_{A^cA}C_{AA}^{-1}\|_\infty\leq\frac{(\kappa+1)\epsilon\varphi}{1-\varphi\epsilon} \tag{43}$$

and

$$\|(\hat{\mu}_2-\hat{\mu}_1)-(\mu_2-\mu_1)\|_\infty\leq\frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa} \tag{44}$$

then $U_1\leq\frac{\lambda}{2}$. Therefore, by Lemma 1 and Lemma 2, we have

$$\begin{aligned}&\Pr(\|\frac{1}{n}\tilde{X}_{A^c}^T\tilde{X}_A\hat{\beta}_A-(\mu_{2A^C}-\mu_{1A^C})\|_\infty\leq\frac{\lambda}{2}) \\ \equiv\ &1-\delta_1 \\ \geq\ &1-2ps\exp(-\frac{n}{s^2}\epsilon^2c_1)-2p\exp(-nc_2(\frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa})^2).\end{aligned} \tag{45}$$

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

$$|\hat{\beta}_j| \geq \zeta\Delta\varphi-(\eta_3+\varphi)(\frac{\lambda}{2}+\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-\mu_{2A}-\mu_{1A})\|_\infty)-\eta_3\Delta. \tag{46}$$

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

$$|\hat{\beta}_j| \;\geq\; \zeta\Delta\varphi - \frac{1}{1-\eta_1\varphi}(\frac{\lambda\varphi}{2} + \|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty\varphi + \varphi^2\eta_1\Delta) \equiv L_1. \quad (47)$$

Note that $\zeta \leq 1$, because $\|\beta^*\|_\infty \leq \Delta\varphi$. Hence $\lambda \leq \frac{1}{2}|\beta^*|_{min}/\varphi \leq \frac{2}{3+\zeta}|\beta^*|_{min}/\varphi$. Pick ϵ such that $\epsilon < \min(\epsilon_0, \frac{1}{\varphi}\frac{\zeta}{3+\zeta}, \frac{\Delta}{2}\frac{\zeta}{3+\zeta})$ Under the events $\eta_1 \leq \epsilon$ and $\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty \leq \epsilon$ we have $L_1 > 0$. Therefore,

$$\Pr(L_1 > 0) \;\geq\; 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2). \quad (48)$$

Part (3). By (39) and $\eta_1\varphi < 1$, we have

$$\|\hat{\beta}_A - \beta^*\|_\infty \leq \frac{1}{1-\eta_1\varphi}(\frac{\lambda}{2}\varphi + \|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty\varphi + \varphi^2\eta_1\Delta). \quad (49)$$

Pick ϵ such that $\epsilon < \min(\epsilon_0, \frac{\lambda}{2\varphi\Delta}, \lambda)$. Under the events $\eta_1 < \epsilon$ and $\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty \leq \epsilon$ we have $\|\hat{\beta}_A - \beta^*\|_\infty \leq 4\varphi\lambda$. Thus,

$$\Pr(\|\hat{\beta}_A - \beta^*\|_\infty \leq 4\varphi\lambda) \;\geq\; 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2). \quad (50)$$

This completes the proof.

□

Proof of Theorem 2. Part (1). Fix any positive ϵ satisfying $\frac{\epsilon}{\epsilon+2\Delta\varphi} \leq \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta})$. Under the events $\eta_1\varphi < \frac{1}{1+\frac{2\Delta\varphi}{\epsilon}}$ and $\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty \leq \frac{\epsilon}{\epsilon+2\Delta\varphi}\Delta$, we have

$$\begin{aligned}\|\hat{\beta}(oracle) - \beta^*\|_\infty &\leq (\eta_3+\varphi)\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty + \eta_3\|\mu_{2A}-\mu_{1A}\|_\infty \\ &\leq \frac{1}{1-\eta_1\varphi}(\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty\varphi + \varphi^2\eta_1\Delta) \\ &< \epsilon. \end{aligned} \quad (51)$$

Then (21) is obtained by Lemma 1.

Part (2). Let $g(\beta) = \frac{1}{n}\beta^T(\tilde{X}^T\tilde{X})\beta - 2(\hat{\mu}_2-\hat{\mu}_1)^T\beta + \sum_{j=1}^p P_\lambda(|\beta_j|)$. If the following two conditions hold

$$|\hat{\beta}(oracle)_A|_{min} > a\lambda \quad (52)$$

$$\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\mu}_{2A}-\hat{\mu}_{1A}) - (\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})\|_\infty < \frac{\lambda}{2} \quad (53)$$

then $\hat{\beta}(oracle)$ is a local minimizer of $g(\beta)$. To see this, consider any $\beta = \hat{\beta}(oracle) + b$ with a sufficiently small b satisfying $\|b\|_2 < \min(\lambda, \frac{1}{2}(|\hat{\beta}(oracle)_A|_{min} - a\lambda))$. Let $z =$

$C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})$. Then it is easy to check that

$$\begin{aligned} g(\beta)-g(\hat{\beta}(oracle)) &= b^TC^{(n)}b+(\sum_{j\in A^c}\lambda|b_j|)+2z^Tb_{A^c} \\ &\geq \sum_{j\in A^c}(\lambda-2\|z\|_\infty)|b_j| \ \geq \ 0. \end{aligned} \tag{54}$$

Clearly, "=" is taken if and only if $b=0$. Thus, within a sufficiently small ball centered at $\hat{\beta}(oracle)$, $\hat{\beta}(oracle)$ is the unique (strict) minimizer of the objective function.

First, we derive a bound for the probability of (52). Pick some ϵ in part (1) and let $\epsilon<|\beta^*|_{min}-a\lambda$. Then $\Pr(|\hat{\beta}(oracle)_A|_{min}>a\lambda)>\Pr(|\hat{\beta}(oracle)_A|_{min}>|\beta^*|_{min}-\epsilon)$ and (21) implies

$$\Pr(|\hat{\beta}(oracle)_A|_{min}>|\beta^*|_{min}-\epsilon)\geq 1-2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2})-2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}). \tag{55}$$

To derive a bound for the probability of (53), we use similar arguments as in the proof of Theorem 1. Consider three events $\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}-C_{A^cA}C^{-1}_{AA}\|_\infty\leq\epsilon<\frac{\lambda}{6\Delta}$, $\|(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})-(\mu_{2A^C}-\mu_{1A^C})\|_\infty\leq\epsilon<\frac{\lambda}{6}$ and $\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty\leq\epsilon<\frac{\lambda}{6\kappa+\frac{\lambda}{\Delta}}$. Then we have

$$\begin{aligned} & \|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})\|_\infty \\ \leq \ & \|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}-C_{A^cA}C^{-1}_{AA}\|_\infty\Delta+\|(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})-(\mu_{2A^C}-\mu_{1A^C})\|_\infty \\ & +(\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}-C_{A^cA}C^{-1}_{AA}\|_\infty+\kappa)\|(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\mu_{2A}-\mu_{1A})\|_\infty \\ < \ & \frac{\lambda}{2}. \end{aligned} \tag{56}$$

By Lemma 1 and Lemma 2, we also have

$$\begin{aligned} & \Pr(\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\mu}_{2A}-\hat{\mu}_{1A})-(\hat{\mu}_{2A^C}-\hat{\mu}_{1A^C})\|_\infty<\frac{\lambda}{2}) \\ \geq \ & 1-2p\exp(-nc_2\epsilon^2)-2ps\exp(-\frac{nc_1}{s^2}\frac{1}{\varphi^2}(\frac{\epsilon}{\epsilon+\kappa+1})^2). \end{aligned} \tag{57}$$

We obtain the expression for δ_3 in (22) by combining (55) and (57). This completes the proof. □

Proof of Theorem 3. Theorem 3 directly follows Theorems 1 and 2. □

Acknowledgement

Mai is supported by the Alumni Fellowship from the School of Statistics at University of Minnesota. Zou's research is supported in part by NSF grant DMS-0846068. Yuan's research is supported in part by NSF grant DMS-0846234.

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S. & Levine, J. (1999), 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.', *Proc. Nat. Acad. of Sci.* **96**, 6745–6750.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J. R. Statist. Soc. B* **57**, 289–300.

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', *The Ann. Statist.* **36**, 199–227.

Cai, T., Zhang, C. & Zhou, H. (2010), 'Optimal rates of convergence for covariance matrix estimation', *The Ann. Statist.* **38**, 2118–2144.

Dettling, M. (2004), 'Bagboosting for tumor classification with gene expression data', *Bioinformatics* **20**, 3583–3593.

Donoho, D. & Jin, J. (2004), 'Higher criticism for detecting sparse heterogeneous mixtures', *The Ann. Statist* **32**, 962–994.

Dudoit, S. & Van der Laan, M. (2008), *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics. New York: Springer.

Efron, B. (2005), Local false discovery rate, Technical report, Stanford University.

Efron, B. (2009), 'Empirical Bayes estimates for large-scale prediction problems', *J. Am. Statist. Assoc.* **104**, 1015–1028.

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes methods for estimation, testing and prediction*, Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Ann. Statist.* **32**, 407–499.

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

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

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space (with discussion)', *J. R. Statist. Soc. B* **70**, 849–911.

Fan, J. & Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**, 101–148.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Regularization paths for generalized linear models via coordinate descent', *J. Stat. Software* **33**, 1–22.

Genovese, C. & Wasserman, L. (2004), 'A stochastic process approach to false discovery control', *The Ann. Statist* **32**, 1035–1061.

Guo, Y., Hastie, T. & R., T. (2006), 'Regularized linear discriminant analysis and its application in microarrays', *Biostatistics* **8**, 86–100.

Hall, P., Marron, J. S. & Neeman, A. (2005), 'Geometric representation of high dimension, low sample size data', *J. R. Statist. Soc. B* **67**, 427–444.

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

Hastie, T., Tibshirani, R. & Friedman, J. H. (2008), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edn, Springer Verlag.

Levina, E., Rothman, A. & Zhu, J. (2008), 'Sparse estimation of large covariance matrices via a nested lasso penalty', *The Ann. Appl. Statist.* **2**, 245–263.

Lv, J. & Fan, Y. (2009), 'A unified approach to model selection and sparse recovery using regularized least squares', *The Ann. Statist* **37**, 3498–3528.

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

Rothman, A., Bickel, P., Levina, E. & Zhu, J. (2008), 'Sparse permutation invariant covariance estimation', *Electron. J. Statist.* **2**, 494–515.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., M, L., Kantoff, P., Golub, T. & Sellers, W. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell* **1**(2), 203–209.

Storey, J. (2002), 'A direct approach to false discovery rates', *J. R. Statist. Soc. B* **64**, 479–498.

Storey, J., Taylor, J. & Siegmund, D. (2004), 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates; a unified approach', *J. R. Statist. Soc. B* **66**, 187–206.

Sun, W. & Cai, T. (2007), 'Oracle and adaptive compound decision rules for false discovery rate control', *J. Am. Statist. Assoc.* **102**, 901–912.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. R. Statist. Soc. B* **58**, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Nat. Acad. Sci.* **99**, 6567–6572.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Keith, K. (2005), 'Sparsity and smoothness via the fused lasso', *J. R. Statist. Soc. B* **67**, 91–108.

Wang, S. & Zhu, J. (2007), 'Improved centroids estimation for the nearest shrunken centroid classifier', *Bioinformatics* **23**, 972–979.

Witten, D. & Tibshirani, R. (2011), 'Penalized classification using fisher's linear discriminant', *J. R. Statist. Soc. B* .

Wu, M., Zhang, L., Wang, Z., Christiani, D. & 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., J. R. & Lin, Y. (2007), 'An efficient variable selection approach for analyzing designed experiments', *Technometrics* **49**, 430–439.

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

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

Zhang, C. & Huang, J. (2008), 'The sparsity and bias of the lasso selection in high-dimensional linear regression', *The Ann. Statist* **36**, 1567–1594.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *J. Mach. Learn. Res.* **7**, 2541–2567.

Zou, H. (2006), 'The adaptive Lasso and its oracle properties', *J. Am. Statist. Assoc.* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. R. Statist. Soc. B* **67**, 301–320.

Zou, H. & Li, R. (2008), 'One-step sparse estimates in nonconcave penalized likelihood models(with discussion)', *The Ann. Statist* **36**, 1509–1533.