The Annals of Statistics
2011, Vol. 39, No. 2, 1241–1265
DOI: 10.1214/10-AOS870

SPARSE LINEAR DISCRIMINANT ANALYSIS BY THRESHOLDING FOR HIGH DIMENSIONAL DATA

BY JUN SHAO[1], YAZHEN WANG[2], XINWEI DENG AND SIJIAN WANG

East China Normal University and University of Wisconsin

In many social, economical, biological and medical studies, one objective is to classify a subject into one of several classes based on a set of variables observed from the subject. Because the probability distribution of the variables is usually unknown, the rule of classification is constructed using a training sample. The well-known linear discriminant analysis (LDA) works well for the situation where the number of variables used for classification is much smaller than the training sample size. Because of the advance in technologies, modern statistical studies often face classification problems with the number of variables much larger than the sample size, and the LDA may perform poorly. We explore when and why the LDA has poor performance and propose a sparse LDA that is asymptotically optimal under some sparsity conditions on the unknown parameters. For illustration of application, we discuss an example of classifying human cancer into two classes of leukemia based on a set of 7,129 genes and a training sample of size 72. A simulation is also conducted to check the performance of the proposed method.

1. Introduction. The objective of a classification problem is to classify a subject to one of several classes based on a p-dimensional vector $\mathbf{x}$ of characteristics observed from the subject. In most applications, variability exists, and hence $\mathbf{x}$ is random. If the distribution of $\mathbf{x}$ is known, then we can construct an optimal classification rule that has the smallest possible misclassification rate. However, the distribution of $\mathbf{x}$ is usually unknown, and a classification rule has to be constructed using a training sample. A statistical issue is how to use the training sample to construct a classification rule that has a misclassification rate close to that of the optimal rule.

In traditional applications, the dimension p of $\mathbf{x}$ is fixed while the training sample size n is large. Because of the advance in technologies, nowadays a much larger amount of information can be collected, and the resulting $\mathbf{x}$ is of a high dimension. In many recent applications, p is much larger than the training sample size, which is referred to as the large-p-small-n problem or ultra-high dimension problem when $p = O(e^{n^\beta})$ for some $\beta \in (0, 1)$. An example is a study with genetic

Received February 2010; revised September 2010.
[1]Supported in part by the NSF Grant SES-0705033.
[2]Supported in part by the NSF Grant DMS-10-05635.
MSC2010 subject classifications. Primary 62H30; secondary 62F12, 62G12.
Key words and phrases. Classification, high dimensionality, misclassification rate, normality, optimal classification rule, sparse estimates.

or microarray data. In our example presented in Section 5, for instance, a crucial step for a successful chemotherapy treatment is to classify human cancer into two classes of leukemia, acute myeloid leukemia and acute lymphoblastic leukemia, based on $p = 7{,}129$ genes and a training sample of 72 patients. Other examples include data from radiology, biomedical imaging, signal processing, climate and finance. Although more information is better when the distribution of $\mathbf{x}$ is known, a larger dimension p produces more uncertainty when the distribution of $\mathbf{x}$ is unknown and, hence, results in a greater challenge for data analysis since the training sample size n cannot increase as fast as p.

The well-known linear discriminant analysis (LDA) works well for fixed-p-large-n situations and is asymptotically optimal in the sense that, when n increases to infinity, its misclassification rate over that of the optimal rule converges to one. In fact, we show in this paper that the LDA is still asymptotically optimal when p diverges to infinity at a rate slower than $\sqrt{n}$. On the other hand, Bickel and Levina (2004) showed that the LDA is asymptotically as bad as random guessing when $p > n$; some similar results are also given in this paper. The main purpose of this paper is to construct a sparse LDA and show it is asymptotically optimal under some sparsity conditions on unknown parameters and some condition on the divergence rate of p (e.g., $n^{-1}\log p \to 0$ as $n \to \infty$). Our proposed sparse LDA is based on the thresholding methodology, which was developed in wavelet shrinkage for function estimation [Donoho and Johnstone (1994), Donoho et al. (1995)] and covariance matrix estimation [Bickel and Levina (2008)]. There exist a few other sparse LDA methods, for example, Guo, Hastie and Tibshirani (2007), Clemmensen, Hastie and Ersbøll (2008) and Qiao, Zhou and Huang (2009). The key differences between the existing methods and ours are the conditions on sparsity and the construction of sparse estimators of parameters. However, no asymptotic results were established in the existing papers.

For high-dimensional $\mathbf{x}$ in regression, there exist some variable selection methods [see a recent review by Fan and Lv (2010)]. For constructing a classification rule using variable selection, we must identify not only components of $\mathbf{x}$ having mean effects for classification, but also components of $\mathbf{x}$ having effects for classification through their correlations with other components [see, e.g., Kohavi and John (1997), Zhang and Wang (2010)]. This may be a very difficult task when p is much larger than n, such as $p = 7{,}129$ and $n = 72$ in the leukemia example in Section 5. Ignoring the correlation, Fan and Fan (2008) proposed the features annealed independence rule (FAIR), which first selects m components of $\mathbf{x}$ having mean effects for classification and then applies the naive Bayes rule (obtained by assuming that components of $\mathbf{x}$ are independent) using the selected m components of $\mathbf{x}$ only. Although no sparsity condition on the covariance matrix of $\mathbf{x}$ is required, the FAIR is not asymptotically optimal because the correlation between components of $\mathbf{x}$ is ignored. Our approach is not a variable selection approach, that is, we do not try to identify a subset of components of $\mathbf{x}$ with a size smaller than n. We use thresholding estimators of the mean effects as well as Bickel and Levina's (2008)

thresholding estimator of the covariance matrix of $\mathbf{x}$, but we allow the number of nonzero estimators (for the mean differences or covariances) to be much larger than n to ensure the asymptotic optimality of the resulting classification rule.

The rest of this paper is organized as follows. In Section 2, after introducing some notation and terminology, we establish a sufficient condition on the divergence of p under which the LDA is still asymptotically close to the optimal rule. We also show that, when p is large compared with n ($p/n \to \infty$), the performance of the LDA is not good even if we know the covariance matrix of $\mathbf{x}$, which indicates the need of sparse estimators for both the mean difference and covariance matrix. Our main result is given in Section 3, along with some discussions about various sparsity conditions and divergence rates of p for which the proposed sparse LDA performs well asymptotically. Extensions of the main result are discussed in Section 4. In Section 5, the proposed sparse LDA is illustrated in the example of classifying human cancer into two classes of leukemia, along with some simulation results for examining misclassification rates. All technical proofs are given in Section 6.

2. The optimal rule and linear discriminant analysis. We focus on the classification problem with two classes. The general case with three or more classes is discussed in Section 4. Let $\mathbf{x}$ be a p-dimensional normal random vector belonging to class k if $\mathbf{x} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2$, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ is positive definite. The misclassification rate of any classification rule is the average of the probabilities of making two types of misclassification: classifying $\mathbf{x}$ to class 1 when $\mathbf{x} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and classifying $\mathbf{x}$ to class 2 when $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$.

If $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, then the optimal classification rule, that is, the rule with the smallest misclassification rate, classifies $\mathbf{x}$ to class 1 if and only if $\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \bar{\boldsymbol{\mu}}) \geq 0$, where $\bar{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and $\mathbf{a}'$ denotes the transpose of the vector $\mathbf{a}$. This rule is also the Bayes rule with equal prior probabilities for two classes. Let R_{OPT} denote the misclassification rate of the optimal rule. Using the normal distribution, we can show that

$$R_{\mathrm{OPT}} = \Phi(-\Delta_p/2), \qquad \Delta_p = \sqrt{\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}, \tag{1}$$

where Φ is the standard normal distribution function. Although $0 < R_{\mathrm{OPT}} < 1/2$, $R_{\mathrm{OPT}} \to 0$ if $\Delta_p \to \infty$ as $p \to \infty$ and $R_{\mathrm{OPT}} \to 1/2$ if $\Delta_p \to 0$. Since $1/2$ is the misclassification rate of random guessing, we assume the following regularity conditions: there is a constant c_0 (not depending on p) such that

$$c_0^{-1} \leq \text{all eigenvalues of } \boldsymbol{\Sigma} \leq c_0 \tag{2}$$

and

$$c_0^{-1} \leq \max_{j \leq p} \delta_j^2 \leq c_0, \tag{3}$$

where δ_j is the jth component of $\boldsymbol{\delta}$. Under (2)–(3), $\Delta_p \geq c_0^{-1}$, and hence $R_{\mathrm{OPT}} \leq \Phi(-(2c_0)^{-1}) < 1/2$. Also, $\Delta_p^2 = O(\|\boldsymbol{\delta}\|^2)$ and $\|\boldsymbol{\delta}\|^2 = O(\Delta_p^2)$ so that the rate of $\|\boldsymbol{\delta}\|^2 \to \infty$ is the same as the rate of $\Delta_p^2 \to \infty$, where $\|\mathbf{a}\|$ is the L_2-norm of the vector $\mathbf{a}$.

In practice, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are typically unknown, and we have a training sample $\mathbf{X} = \{\mathbf{x}_{ki}, i = 1, \ldots, n_k, k = 1, 2\}$, where n_k is the sample size for class k, $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2$, all $\mathbf{x}_{ki}$'s are independent and $\mathbf{X}$ is independent of $\mathbf{x}$ to be classified. The limiting process considered in this paper is the one with $n = n_1 + n_2 \to \infty$. We assume that n_1/n converges to a constant strictly between 0 and 1; p is a function of n, but the subscript n is omitted for simplicity. When $n \to \infty$, p may diverge to ∞, and the limit of p/n may be 0, a positive constant, or ∞.

For a classification rule T constructed using the training sample, its performance can be assessed by the conditional misclassification rate $R_T(\mathbf{X})$ defined as the average of the conditional probabilities of making two types of misclassification, where the conditional probabilities are with respect to $\mathbf{x}$, given the training sample $\mathbf{X}$. The unconditional misclassification rate is $R_T = E[R_T(\mathbf{X})]$. The asymptotic performance of T refers to the limiting behavior of $R_T(\mathbf{X})$ or R_T as $n \to \infty$. Since $0 \leq R_T(\mathbf{X}) \leq 1$, by the dominated convergence theorem, if $R_T(\mathbf{X}) \to_P c$, where c is a constant and $\to_P$ denotes convergence in probability, then $R_T \to c$. Hence, in this paper we focus on the limiting behavior of the conditional misclassification rate $R_T(\mathbf{X})$.

We hope to find a rule T such that $R_T(\mathbf{X})$ converges in probability to the same limit as R_{OPT}, the misclassification rate of the optimal rule. If $R_{\mathrm{OPT}} \to 0$, however, we hope not only $R_T(\mathbf{X}) \to_P 0$, but also $R_T(\mathbf{X})$ and R_{OPT} have the same convergence rate. This leads to the following definition.

DEFINITION 1. Let T be a classification rule with conditional misclassification rate $R_T(\mathbf{X})$, given the training sample $\mathbf{X}$.

(i) T is asymptotically optimal if $R_T(\mathbf{X})/R_{\mathrm{OPT}} \to_P 1$.
(ii) T is asymptotically sub-optimal if $R_T(\mathbf{X}) - R_{\mathrm{OPT}} \to_P 0$.
(iii) T is asymptotically worst if $R_T(\mathbf{X}) \to_P 1/2$.

If $\lim_{n\to\infty} R_{\mathrm{OPT}} > 0$ [i.e., Δ_p in (1) is bounded], then the asymptotic sub-optimality is the same as the asymptotic optimality. Part (iii) of Definition 1 comes from the fact that $1/2$ is the misclassification rate of random guessing.

In this paper we focus on the classification rules of the form

$$\text{classifying } \mathbf{x} \text{ to class 1 if and only if } \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\bar{\boldsymbol{\mu}}}) \geq 0, \tag{4}$$

where $\hat{\boldsymbol{\delta}}$, $\hat{\bar{\boldsymbol{\mu}}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ are estimators of $\boldsymbol{\delta}$, $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}^{-1}$, respectively, constructed using the training sample $\mathbf{X}$.

The well-known linear discriminant analysis (LDA) uses the maximum likelihood estimators $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\mathbf{S}$, where

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_{ki}, \qquad k = 1, 2, \qquad \mathbf{S} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'.$$

The LDA is given by (4) with $\hat{\boldsymbol{\delta}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, $\hat{\bar{\boldsymbol{\mu}}} = \bar{\mathbf{x}} = (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$, $\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{S}^{-1}$ when $\mathbf{S}^{-1}$ exists, and $\hat{\boldsymbol{\Sigma}}^{-1}$ = a generalized inverse $\mathbf{S}^{-}$ when $\mathbf{S}^{-1}$ does not exist (e.g., when $p > n$). A straightforward calculation shows that, given $\mathbf{X}$, the conditional misclassification rate of the LDA is

$$\frac{1}{2}\sum_{k=1}^{2}\Phi\biggl(\frac{(-1)^k\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_k - \bar{\mathbf{x}}_k) - \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\delta}}/2}{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\delta}}}}\biggr). \tag{5}$$

Is the LDA asymptotically optimal or sub-optimal according to Definition 1? Bickel and Levina [(2004), Theorem 1] showed that, if $p > n$ and $p/n \to \infty$, then the unconditional misclassification rate of the LDA converges to $1/2$ so that the LDA is asymptotically worst. A natural question is, for what kind of p (which may diverge to ∞), is the LDA asymptotically optimal or sub-optimal. The following result provides an answer.

THEOREM 1. *Suppose that* (2)–(3) *hold and* $s_n = p\sqrt{\log p}/\sqrt{n} \to 0$.

(i) *The conditional misclassification rate of the LDA is equal to*

$$R_{\mathrm{LDA}}(\mathbf{X}) = \Phi\bigl(-[1 + O_P(s_n)]\Delta_p/2\bigr).$$

(ii) *If* Δ_p *is bounded, then the LDA is asymptotically optimal and*

$$\frac{R_{\mathrm{LDA}}(\mathbf{X})}{R_{\mathrm{OPT}}} - 1 = O_P(s_n).$$

(iii) *If* $\Delta_p \to \infty$, *then the LDA is asymptotically sub-optimal.*

(iv) *If* $\Delta_p \to \infty$ *and* $s_n\Delta_p^2 = (p\sqrt{\log p}/\sqrt{n})\Delta_p^2 \to 0$, *then the LDA is asymptotically optimal.*

REMARK 1. Since $\Delta_p \not\to 0$ under conditions (2) and (3), when Δ_p is bounded, $s_n\Delta_p^2 \to 0$ is the same as $s_n \to 0$, which is satisfied if $p = O(n^\lambda)$ with $0 \le \lambda < 1/2$. When $\Delta_p \to \infty$, $s_n\Delta_p^2 \to 0$ is stronger than $s_n \to 0$. Under (2)–(3), $\Delta_p^2 = O(p)$. Hence, the extreme case is Δ_p^2 is a constant times p, and the condition in part (iv) becomes $p^2\sqrt{\log p}/\sqrt{n} \to 0$, which holds when $p = O(n^\lambda)$ with $0 \le \lambda < 1/4$. In the traditional applications with a fixed p, Δ_p is bounded, $s_n \to 0$ as $n \to \infty$ and thus Theorem 1 proves that the LDA is asymptotically optimal.

The proof of part (iv) of Theorem 1 (see Section 6) utilizes the following lemma, which is also used in the proofs of other results in this paper.

LEMMA 1. *Let ξ_n and τ_n be two sequences of positive numbers such that $\xi_n \to \infty$ and $\tau_n \to 0$ as $n \to \infty$. If $\lim_{n\to\infty} \tau_n \xi_n = \gamma$, where γ may be 0, positive, or ∞, then*

$$\lim_{n\to\infty} \frac{\Phi(-\sqrt{\xi_n}(1-\tau_n))}{\Phi(-\sqrt{\xi_n})} = e^{\gamma}.$$

Since the LDA uses $\mathbf{S}^-$ to estimate $\mathbf{\Sigma}^{-1}$ when $p > n$ and is asymptotically worst as Bickel and Levina (2004) showed, one may think that the bad performance of the LDA is caused by the fact that $\mathbf{S}^-$ is not a good estimator of $\mathbf{\Sigma}^{-1}$. Our following result shows that the LDA may still be asymptotically worst even if we can estimate $\mathbf{\Sigma}^{-1}$ perfectly.

THEOREM 2. *Suppose that (2)–(3) hold, $p/n \to \infty$ and that $\mathbf{\Sigma}$ is known so that the LDA is given by (4) with $\hat{\mathbf{\Sigma}}^{-1} = \mathbf{\Sigma}^{-1}$, $\hat{\boldsymbol{\delta}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ and $\hat{\bar{\boldsymbol{\mu}}} = \bar{\mathbf{x}}$.*

(i) *If $\Delta_p^2/\sqrt{p/n} \to 0$ (which is true if $\Delta_p \not\to \infty$), then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 1/2$.*

(ii) *If $\Delta_p^2/\sqrt{p/n} \to c$ with $0 < c < \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P$ a constant strictly between 0 and $1/2$ and $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.*

(iii) *If $\Delta_p^2/\sqrt{p/n} \to \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 0$ but $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.*

Theorem 2 shows that even if $\mathbf{\Sigma}$ is known, the LDA may be asymptotically worst and the best we can hope is that the LDA is asymptotically sub-optimal. It can also be shown that, when $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known and we apply the LDA with $\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}$ and $\hat{\bar{\boldsymbol{\mu}}} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, the LDA is still not asymptotically optimal when $\|\boldsymbol{\delta}\|^2 - \|\boldsymbol{\delta}_n\|^2 \not\to 0$, where $\boldsymbol{\delta}_n$ is any sub-vector of $\boldsymbol{\delta}$ with dimension n. This indicates that, in order to obtain an asymptotically optimal classification rule when p is much larger than n, we need sparsity conditions on $\mathbf{\Sigma}$ and $\boldsymbol{\delta}$ when both of them are unknown. For bounded Δ_p (in which case the asymptotic optimality is the same as the asymptotic sub-optimality), by imposing sparsity conditions on $\mathbf{\Sigma}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, Theorem 2 of Bickel and Levina (2004) shows the existence of an asymptotically optimal classification rule. In the next section, we obtain a result by relaxing the boundedness of Δ_p and by imposing sparsity conditions on $\mathbf{\Sigma}$ and $\boldsymbol{\delta}$. Since the difference of the two normal distributions is in $\boldsymbol{\delta}$, imposing a sparsity condition on $\boldsymbol{\delta}$ is weaker and more reasonable than imposing sparsity conditions on both $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

3. Sparse linear discriminant analysis. We focus on the situation where the limit of p/n is positive or ∞. The following sparsity measure on $\mathbf{\Sigma}$ is considered in Bickel and Levina (2008):

$$C_{h,p} = \max_{j \le p} \sum_{l=1}^{p} |\sigma_{jl}|^h, \tag{6}$$

where σ_{jl} is the (j,l)th element of $\mathbf{\Sigma}$, h is a constant not depending on p, $0 \le h < 1$ and 0^0 is defined to be 0. In the special case of $h = 0$, $C_{0,p}$ in (6) is the maximum of the numbers of nonzero elements of rows of $\mathbf{\Sigma}$ so that a $C_{0,p}$ much smaller than p implies many elements of $\mathbf{\Sigma}$ are equal to 0. If $C_{h,p}$ is much smaller than p for a constant $h \in (0,1)$, then $\mathbf{\Sigma}$ is sparse in the sense that many elements of $\mathbf{\Sigma}$ are very small. An example of $C_{h,p}$ much smaller than p is $C_{h,p} = O(1)$ or $C_{h,p} = O(\log p)$.

Under conditions (2) and

$$\frac{\log p}{n} \to 0, \tag{7}$$

Bickel and Levina (2008) showed that

$$\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_P(d_n) \quad \text{and} \quad \|\tilde{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\| = O_P(d_n), \tag{8}$$

where $d_n = C_{h,p}(n^{-1}\log p)^{(1-h)/2}$, $\tilde{\mathbf{\Sigma}}$ is $\mathbf{S}$ thresholded at $t_n = M_1\sqrt{\log p}/\sqrt{n}$ with a positive constant M_1; that is, the (j,l)th element of $\tilde{\mathbf{\Sigma}}$ is $\hat{\sigma}_{jl} I(|\hat{\sigma}_{jl}| > t_n)$, $\hat{\sigma}_{jl}$ is the (j,l)th element of $\mathbf{S}$ and $I(A)$ is the indicator function of the set A. We consider a slight modification, that is, only off-diagonal elements of $\mathbf{S}$ are thresholded. The resulting estimator is still denoted by $\tilde{\mathbf{\Sigma}}$ and it has property (8) under conditions (2) and (7).

We now turn to the sparsity of $\boldsymbol{\delta}$. On one hand, a large Δ_p results in a large difference between $N_p(\mu_1, \mathbf{\Sigma})$ and $N_p(\mu_2, \mathbf{\Sigma})$ so that the optimal rule has a small misclassification rate. On the other hand, a larger divergence rate of Δ_p results in a more difficult task of constructing a good classification rule, since $\boldsymbol{\delta}$ has to be estimated based on the training sample $\mathbf{X}$ of a size much smaller than p. We consider the following sparsity measure on $\boldsymbol{\delta}$ that is similar to the sparsity measure $C_{h,p}$ on $\mathbf{\Sigma}$:

$$D_{g,p} = \sum_{j=1}^{p} \delta_j^{2g}, \tag{9}$$

where δ_j is the jth component of $\boldsymbol{\delta}$, g is a constant not depending on p and $0 \le g < 1$. If $D_{g,p}$ is much smaller than p for a $g \in [0,1)$, then $\boldsymbol{\delta}$ is sparse. For Δ_p^2 defined in (1), under (2)–(3), $\Delta_p^2 \le c_0\|\boldsymbol{\delta}\|^2 \le c_0^{1+2(1-g)} D_{g,p}$. Hence, the rate of divergence of Δ_p^2 is always smaller than that of $D_{g,p}$ and, in particular, Δ_p is bounded when $D_{g,p}$ is bounded for a $g \in [0,1)$.

We consider the sparse estimator $\tilde{\boldsymbol{\delta}}$ that is $\hat{\boldsymbol{\delta}}$ thresholded at

$$a_n = M_2\left(\frac{\log p}{n}\right)^{\alpha} \tag{10}$$

with constants $M_2 > 0$ and $\alpha \in (0, 1/2)$, that is, the jth component of $\tilde{\boldsymbol{\delta}}$ is $\hat{\delta}_j I(|\hat{\delta}_j| > a_n)$, where $\hat{\delta}_j$ is the jth component of $\hat{\boldsymbol{\delta}}$. The following result is useful.

LEMMA 2. *Let δ_j be the jth component of $\boldsymbol{\delta}$, $\hat{\delta}_j$ be the jth component of $\hat{\boldsymbol{\delta}}$, a_n be given by (10) and $r > 1$ be a fixed constant.*

(i) *If (7) holds, then*

$$P\Bigg(\bigcap_{1\le j\le p,|\delta_j|\le a_n/r}\{|\hat{\delta}_j| \le a_n\}\Bigg) \to 1 \tag{11}$$

and

$$P\Bigg(\bigcap_{1\le j\le p,|\delta_j|> ra_n}\{|\hat{\delta}_j| > a_n\}\Bigg) \to 1. \tag{12}$$

(ii) *Let q_{n0} = the number of j's with $|\delta_j| > ra_n$, q_n = the number of j's with $|\delta_j| > a_n/r$ and $\hat{q}$ = the number of j's with $|\hat{\delta}_j| > a_n$. If (7) holds, then*

$$P(q_{n0} \le \hat{q} \le q_n) \to 1.$$

We propose a sparse linear discriminant analysis (SLDA) for high-dimension p, which is given by (4) with $\hat{\delta} = \tilde{\delta}$, $\hat{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}}$ and $\hat{\bar{\boldsymbol{\mu}}} = \bar{\mathbf{x}}$. The following result establishes the asymptotic optimality of the SLDA under some conditions on the rate of divergence of p, $C_{h,p}$, $D_{g,p}$, q_n and Δ_p^2.

THEOREM 3. *Let $C_{h,p}$ be given by (6), $D_{g,p}$ be given by (9), a_n be given by (10), q_n be as defined in Lemma 2 and $d_n = C_{h,p}(n^{-1}\log p)^{(1-h)/2}$. Assume that conditions (2), (3) and (7) hold and*

$$b_n = \max\Bigg\{d_n, \frac{a_n^{1-g}\sqrt{D_{g,p}}}{\Delta_p}, \frac{\sqrt{C_{h,p}q_n}}{\Delta_p\sqrt{n}}\Bigg\} \to 0. \tag{13}$$

(i) *The conditional misclassification rate of the SLDA is equal to*

$$R_{\mathrm{SLDA}}(\mathbf{X}) = \Phi\big(-[1 + O_P(b_n)]\Delta_p/2\big).$$

(ii) *If Δ_p is bounded, then the SLDA is asymptotically optimal and*

$$\frac{R_{\mathrm{SLDA}}(\mathbf{X})}{R_{\mathrm{OPT}}} - 1 = O_P(b_n).$$

(iii) *If $\Delta_p \to \infty$, then the SLDA is asymptotically sub-optimal.*

(iv) *If $\Delta_p \to \infty$ and $b_n\Delta_p^2 \to 0$, then the SLDA is asymptotically optimal.*

REMARK 2. Condition (13) may be achieved by an appropriate choice of α in a_n, given the divergence rates of $C_{h,p}$, $D_{g,p}$, q_n and Δ_p.

REMARK 3. When Δ_p is bounded and (2)–(3) hold, condition (13) is the same as

$$d_n \to 0, \qquad D_{g,p} a_n^{2(1-g)} \to 0 \quad \text{and} \quad C_{h,p} q_n / n \to 0. \tag{14}$$

REMARK 4. When $\Delta_p \to \infty$, condition (13), which is sufficient for the asymptotic sub-optimality of the SLDA, is implied by $d_n \to 0$, $D_{g,p} a_n^{2(1-g)} = O(1)$ and $C_{h,p} q_n / n = O(1)$. When $\Delta_p \to \infty$, the condition $b_n \Delta_p^2 \to 0$, which is sufficient for the asymptotic optimality of the SLDA, is the same as

$$\Delta_p^2 d_n \to 0, \qquad \Delta_p^2 D_{g,p} a_n^{2(1-g)} \to 0 \quad \text{and} \quad \Delta_p^2 C_{h,p} q_n / n \to 0. \tag{15}$$

We now study when condition (13) holds and when $b_n \Delta_p^2 \to 0$ with $\Delta_p \to \infty$. By Remarks 3 and 4, (13) is the same as condition (14) when Δ_p is bounded, and $b_n \Delta_p^2 \to 0$ is the same as condition (15) when $\Delta_p \to \infty$.

1. If there are two constants c_1 and c_2 such that $0 < c_1 \le |\delta_j| \le c_2$ for any nonzero δ_j, then q_n is exactly the number of nonzero δ_j's. Under condition (3), Δ_p^2 and $D_{0,p}$ have exactly the order q_n.
   (a) If q_n is bounded (e.g., there are only finitely many nonzero δ_j's), then Δ_p is bounded and condition (13) is the same as condition (14). The last two convergence requirements in (14) are implied by $d_n = C_{h,p}(n^{-1} \times \log p)^{(1-h)/2} \to 0$, which is the condition for the consistency of $\tilde{\mathbf{\Sigma}}$ proposed by Bickel and Levina (2008).
   (b) When $q_n \to \infty$ ($\Delta_p \to \infty$), we assume that $q_n = O(n^\eta)$ and $C_{h,p} = O(n^\gamma)$ with $\eta \in (0, 1)$ and $\gamma \in [0, 1)$. Then, condition (15) is implied by

$$n^{\eta+\gamma}(n^{-1}\log p)^{(1-h)/2} \to 0, \qquad n^{2\eta}(n^{-1}\log p)^{2\alpha} \to 0, \qquad n^{2\eta+\gamma-1} \to 0. \tag{16}$$

   If we choose $\alpha = (1-h)/4$, then condition (16) holds when $2\eta + \gamma < 1$ and $n^{\eta+\gamma}(n^{-1}\log p)^{(1-h)/2} \to 0$. To achieve (16) we need to know the divergence rate of p. If $p = O(n^\kappa)$ for a $\kappa \ge 1$, then $(n^{-1}\log p)^{(1-h)/2} = O((n^{-1}\log n)^{(1-h)/2})$, and thus condition (16) holds when $\eta + \gamma < (1-h)/2$ and $\eta < (1+h)/2$. If $p = O(e^{n^\beta})$ for a $\beta \in (0, 1)$, which is referred to as an ultra-high dimension, then $(n^{-1}\log p)^{(1-h)/2} = (n^{\beta-1})^{(1-h)/2}$, and condition (16) holds if $\eta + \gamma < (1-h)(1-\beta)/2$ and $\eta < 1 - (1-h)(1-\beta)/2$.
2. Since

$$\Delta_p^2 \ge \sum_{j:|\delta_j| > a_n/r} \delta_j^2 \ge q_n (a_n/r)^2$$

   and

$$D_{g,p} \ge \sum_{j:|\delta_j| > a_n/r} \delta_j^{2g} \ge q_n (a_n/r)^{2(1-g)},$$

we conclude that

$$q_n = O\left(\min\left\{\frac{\Delta_p^2}{a_n^2}, \frac{D_{g,p}}{a_n^{2(1-g)}}\right\}\right). \tag{17}$$

The right-hand side of (17) can be used as a bound of the divergence rate of q_n when $q_n \to \infty$, although it may not be a tight bound. For example, if $\Delta_p^2 = O(\log p)$ and the right-hand side of (17) is used as a bound for q_n, then the last convergence requirement in (14) or (15) is implied by the first convergence requirement in (14) or (15) when $\alpha \le (1+h)/4$.

3. If $D_{g,p} = O(C_{h,p})$, then the second convergence requirement in (14) or (15) is implied by the first convergence requirement in (14) or (15) when $\alpha \ge (1-h)/[4(1-g)]$.
4. Consider the case where $C_{h,p} = O(\log p)$, $D_{g,p} = O(\log p)$ and an ultra-high dimension, that is, $p = O(e^{n^\beta})$ for a $\beta \in (0,1)$. From the previous discussion, condition (14) holds if $d_n \to 0$, and (15) holds if $d_n \log p \to 0$. Since $\log p = O(n^\beta)$, $d_n = O(n^{\beta+(\beta-1)(1-h)/2})$, which converges to 0 if $\beta < (1-h)/(3-h)$. If Δ_p is bounded, then $d_n \to 0$ is sufficient for condition (13). If $\Delta_p \to \infty$, then the largest divergence rate of Δ_p^2 is $O(\log p) = O(n^\beta)$ and $\Delta_p^2 d_n \to 0$ (i.e., the SLDA is asymptotically optimal) when $\beta < (1-h)/(5-h)$. When $h = 0$, this means $\beta < 1/5$.
5. If the divergence rate of p is smaller than $O(e^{n^\beta})$ then we can afford to have a larger than $O(\log p)$ divergence rate for $C_{h,p}$ and $D_{g,p}$. For example, if $p = O(n^\kappa)$ for a $\kappa \ge 1$ and $\max\{C_{h,p}, D_{g,p}\} = cn^\gamma$ for a $\gamma \in (0,1)$ and a positive constant c, then $\log p = O(\log n)$ diverges to ∞ at a rate slower than n^γ. We now study when condition (14) holds. First, $d_n = C_{h,p}(n^{-1}\log p)^{(1-h)/2} = O(n^{\gamma-(1-h)/2}(\log n)^{(1-h)/2})$, which converges to 0 if $\gamma < (1-h)/2 \le 1/2$. Second, $a^{2(1-g)} D_{g,p} = O(n^{\gamma-2(1-g)\alpha}(\log n)^{2(1-g)\alpha})$, which converges to 0 if α is chosen so that $\alpha > \gamma/[2(1-g)]$. Finally, if we use the right-hand side of (17) as a bound for q_n, then $C_{h,p} q_n/n = O(n^{2(1-g)\alpha+\gamma-1}/(\log n)^{2(1-g)\alpha})$, which converges to 0 if $\alpha \le (1-\gamma)/[2(1-g)]$. Thus, condition (14) holds if $\gamma < (1-h)/2$ and $\gamma/[2(1-g)] < \alpha \le (1-\gamma)/[2(1-g)]$. For condition (15), we assume that $\Delta_p^2 = O(n^{\rho\gamma})$ with a $\rho \in [0,1]$ ($\rho = 0$ corresponds to a bounded Δ_p). Then, a similar analysis leads to the conclusion that condition (15) holds if $(1+\rho)\gamma \le (1-h)/2$ and $(1+\rho)\gamma/[2(1-g)] < \alpha \le [1-(1+\rho)\gamma]/[2(1-g)]$.

To apply the SLDA, we need to choose two constants, M_1 in the thresholding estimator $\tilde{\mathbf{\Sigma}}$ and M_2 in the thresholding estimator $\tilde{\boldsymbol{\delta}}$. We suggest a data-driven method via a cross-validation procedure. Let $\mathbf{X}_{ki}$ be the data set containing the entire training sample but with $\mathbf{x}_{ki}$ deleted, and let T_{ki} be the SLDA rule based on $\mathbf{X}_{ki}$, $i = 1, \ldots, n_k$, $k = 1, 2$. The leave-one-out cross-validation estimator of the

misclassification rate of the SLDA is

$$\hat{R}_{\mathrm{SLDA}} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k} r_{ki},$$

where r_{ki} is the indicator function of whether T_{ki} classifies $\mathbf{x}_{ki}$ incorrectly. Let $R(n_1, n_2)$ denote R_{SLDA} when the sample sizes are n_1 and n_2. Then

$$E(\hat{R}_{\mathrm{SLDA}}) = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k} E(r_{ki}) = \frac{n_1 R(n_1 - 1, n_2) + n_2 R(n_1, n_2 - 1)}{n},$$

which is close to $R(n_1, n_2) = R_{\mathrm{SLDA}}$ for large n_k. Let $\hat{R}_{\mathrm{SLDA}}(M_1, M_2)$ be the cross-validation estimator when (M_1, M_2) is used in thresholding $\hat{\mathbf{S}}$ and $\hat{\boldsymbol{\delta}}$. Then, a data-driven method of selecting (M_1, M_2) is to minimize $\hat{R}_{\mathrm{SLDA}}(M_1, M_2)$ over a suitable range of (M_1, M_2). The resulting $\hat{R}_{\mathrm{SLDA}}$ can also be used as an estimate of the misclassification rate of the SLDA.

4. Extensions. We first consider an extension of the main result in Section 3 to nonnormal $\mathbf{x}$ and $\mathbf{x}_{ki}$'s. For nonnormal $\mathbf{x}$, the LDA with known $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$, that is, the rule classifying $\mathbf{x}$ to class 1 if and only if $\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \bar{\boldsymbol{\mu}}) \geq 0$, is still optimal when $\mathbf{x}$ has an elliptical distribution [see, e.g., Fang and Anderson (1990)] with density

$$c_p|\boldsymbol{\Sigma}|^{-1/2} f\big((\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\big), \tag{18}$$

where $\boldsymbol{\mu}$ is either $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, f is a monotone function on $[0, \infty)$, and c_p is a normalizing constant. Special cases of (18) are the multivariate t-distribution and the multivariate double-exponential distribution. Although this rule is not necessarily optimal when the distribution of $\mathbf{x}$ is not of the form (18), it is still a reasonably good rule when $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are known. Thus, when $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are unknown, we study whether the misclassification rate of the SLDA defined in Section 3 is close to that of the LDA with known $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$.

From the proofs for the asymptotic properties of the SLDA in Section 3, the results depending on the normality assumption are:

(i) result (8), the consistency of $\tilde{\boldsymbol{\Sigma}}$;
(ii) results (11) and (12) in Lemma 2;
(iii) the form of the optimal misclassification rate given by (1);
(iv) the result in Lemma 1.

Thus, if we relax the normality assumption, we need to address (i)–(iv). For (i), it was discussed in Section 2.3 of Bickel and Levina (2008) that result (8) still holds when the normality assumption is replaced by one of the following two conditions. The first condition is

$$\sup_{k,j} E(e^{tx_{kij}^2}) < \infty \qquad \text{for all } |t| \leq t_0 \tag{19}$$

for a constant $t_0 > 0$, where x_{kij} is the jth component of $\mathbf{x}_{ki}$. Under condition (19), result (8) holds without any modification. The second condition is

$$\sup_{k,j} E|x_{kij}|^{2\nu} < \infty \tag{20}$$

for a constant $\nu > 0$. Under condition (20), result (8) holds with $n^{-1}\log p$ changed to $n^{-1}p^{4/\nu}$. The same argument can be used to address (ii), that is, results (11) and (12) hold under condition (19) or condition (20) with $n^{-1}\log p$ replaced by $n^{-1}p^{4/\nu}$. For (iii), the normality of $\mathbf{x}$ can be relaxed to that, for any p-dimensional nonrandom vector $\mathbf{l}$ with $\|\mathbf{l}\| = 1$ and any real number t,

$$P\big(\mathbf{l}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \le t\big) = \Psi(t), \tag{21}$$

where Ψ is an unknown distribution function symmetric about 0 but it does not depend on $\mathbf{l}$. Distributions satisfying (21) include elliptical distributions [e.g., a distribution of the form (18)] and the multivariate scale mixture of normals [Fang and Anderson (1990)]. Under (21), when $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are known, the LDA has misclassification rate $\Psi(-\Delta_p/2)$ with Δ_p given by (1). It remains to address (iv). Note that the following result,

$$\frac{x}{1+x^2}e^{-x^2/2} \le \Phi(-x) \le \frac{1}{x}e^{-x^2/2}, \qquad x > 0, \tag{22}$$

is the key for Lemma 1. Without assuming normality, we consider the condition

$$0 < \lim_{x\to\infty} \frac{x^{\omega}e^{-cx^{\varphi}}}{\Psi(-x)} < \infty, \tag{23}$$

where φ is a constant, $0 \le \varphi \le 2$, ω is a constant and c is a positive constant. For the case where Ψ is standard normal, condition (23) holds with $\varphi = 2$, $\omega = -1$ and $c = 1/2$. Under condition (23), we can show that the result in Lemma holds for the case of $\gamma = 0$, which is needed to extend the result in Theorem 3(iv). This leads to the following extension.

THEOREM 4. *Assume condition* (21) *and either condition* (19) *or* (20). *When condition* (19) *holds, let* b_n *be defined by* (13). *When condition* (20) *holds, let* a_n *and* b_n *be defined by* (10) *and* (13), *respectively, with* $n^{-1}\log p$ *replaced by* $n^{-1}p^{4/\nu}$. *Assume that* $a_n \to 0$ *and* $b_n \to 0$.

(i) *The conditional misclassification rate of the SLDA is*

$$R_{\mathrm{SLDA}}(\mathbf{X}) = \Psi\big(-[1 + O_P(b_n)]\Delta_p/2\big).$$

(ii) *If* Δ_p *is bounded, then*

$$\frac{R_{\mathrm{SLDA}}(\mathbf{X})}{\Psi(-\Delta_p/2)} - 1 = O_P(b_n),$$

where $\Psi(-\Delta_p/2)$ *is the misclassification rate of the LDA when* $\boldsymbol{\mu}_k$ *and* $\boldsymbol{\Sigma}$ *are known.*

(iii) *If* $\Delta_p \to \infty$, *then* $R_{\mathrm{SLDA}}(\mathbf{X}) \to_P 0$.
(iv) *If* $\Delta_p \to \infty$ *and* $b_n \Delta_p^2 \to 0$, *then*

$$\frac{R_{\mathrm{SLDA}}(\mathbf{X})}{\Psi(-\Delta_p/2)} \to_P 1.$$

We next consider extending the results in Sections 2 and 3 to the classification problem with $K \geq 3$ classes. Let $\mathbf{x}$ be a p-dimensional normal random vector belonging to class k if $\mathbf{x} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, \ldots, K$, and the training sample be $\mathbf{X} = \{\mathbf{x}_{ki}, i = 1, \ldots, n_k, k = 1, \ldots, K\}$, where n_k is the sample size for class k, $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, \ldots, K$, and all $\mathbf{x}_{ki}$'s are independent. The LDA classifies $\mathbf{x}$ to class k if and only if $\hat{\boldsymbol{\delta}}_{kl}' \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\bar{\boldsymbol{\mu}}}_{kl}) \geq 0$ for all $l \neq k$, $l = 1, \ldots, K$, where $\hat{\boldsymbol{\delta}}_{kl} = \bar{\mathbf{x}}_k - \bar{\mathbf{x}}_l$, $\hat{\bar{\boldsymbol{\mu}}}_{kl} = (\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_l)/2$, $\bar{\mathbf{x}}_k = n_k^{-1} \sum_{i=1}^{n_k} \mathbf{x}_{ki}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ is an inverse or a generalized inverse of $\mathbf{S} = n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'$, and $n = n_1 + \cdots + n_K$. The conditional misclassification rate of the LDA is

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{j \neq k} P_k\big(\hat{\boldsymbol{\delta}}_{jl}' \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\bar{\boldsymbol{\mu}}}_{jl}) \geq 0, l \neq j\big),$$

where P_k is the probability with respect to $\mathbf{x} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, \ldots, K$. The SLDA and its conditional misclassification rate can be obtained by simply replacing $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}_{kl}$ by their thresholding estimators $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\delta}}_{kl}$, respectively. For simplicity of computation, we suggest the use of the same thresholding constant (10) for all $\tilde{\boldsymbol{\delta}}_{kl}$'s.

The optimal rate can be calculated as

$$R_{\mathrm{OPT}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j \neq k} P_k\big(\boldsymbol{\delta}_{jl}' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \bar{\boldsymbol{\mu}}_{jl}) \geq 0, l \neq j\big), \tag{24}$$

where $\boldsymbol{\delta}_{jl} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_l$ and $\bar{\boldsymbol{\mu}}_{jl} = (\boldsymbol{\mu}_j + \boldsymbol{\mu}_l)/2$, $j, l = 1, \ldots, K$, $j \neq l$. Asymptotic properties of the LDA and SLDA can be obtained, under the asymptotic setting with $n \to \infty$ and $n_k/n \to$ a constant in $(0, 1)$ for each k. Sparsity conditions should be imposed to each $\boldsymbol{\delta}_{kl}$. If the probabilities in expression (24) do not converge to 0, then the asymptotic optimality of the LDA (under the conditions in Theorem 1) or the SLDA (under the conditions in Theorem 3) can be established using the same proofs as those in Section 6. When R_{OPT} in (24) converges to 0, to consider convergence rates, the proof of the asymptotic optimality of the LDA or SLDA requires an extension of Lemma 1. Specifically, we need an extension of result (22) to the case of multivariate normal distributions. This technical issue, together with empirical properties of the SLDA with $K \geq 3$, will be investigated in our future research.

5. Numerical studies. Golub et al. (1999) applied gene expression microarray techniques to study human acute leukemia and discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Distinguishing ALL from AML is crucial for successful treatment, since chemotherapy regimens for ALL can be harmful for AML patients. An accurate classification based solely on gene expression monitoring independent of previous biological knowledge is desired as a general strategy for discovering and predicting cancer classes.

We considered a dataset that was used by many researchers [see, e.g., Fan and Fan (2008)]. It contains the expression levels of $p = 7{,}129$ genes for $n = 72$ patients. Patients in the sample are known to come from two distinct classes of leukemia: $n_1 = 47$ are from the ALL class, and $n_2 = 25$ are from the AML class.

Figure 1 displays the cumulative proportions defined as $\sum_{j=1}^{l} \hat{\delta}_{(j)}^2 / \|\hat{\boldsymbol{\delta}}\|^2$, $l = 1, \ldots, p$, where $\hat{\delta}_{(j)}^2$ is the jth largest value among the squared components of $\hat{\boldsymbol{\delta}}$. These proportions indicate the importance of the contribution of each $\hat{\delta}_{(j)}$. It can be seen from Figure 1 that the first 1,000 $\hat{\delta}_{(j)}$'s contribute a cumulative proportion nearly 98%. Figure 2 plots the absolute values of the off-diagonal elements of the sample covariance matrix $\mathbf{S}$. It can be seen that many of them are relatively small. If we ignore a factor of 10^8, then among a total of 25,407,756 values in Figure 2, only 0.45% of them vary from 0.35 to 9.7 and the rest of them are under 0.35.

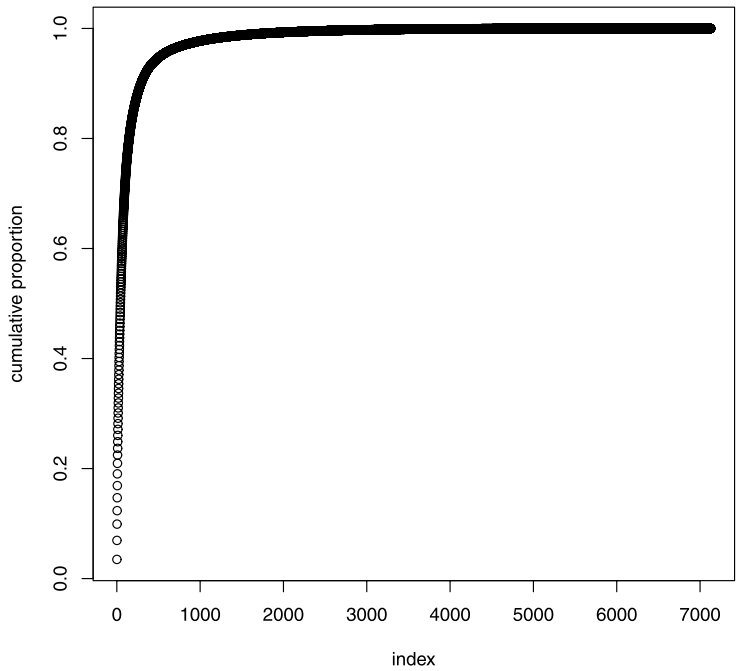

FIG. 1. *Cumulative proportions.*

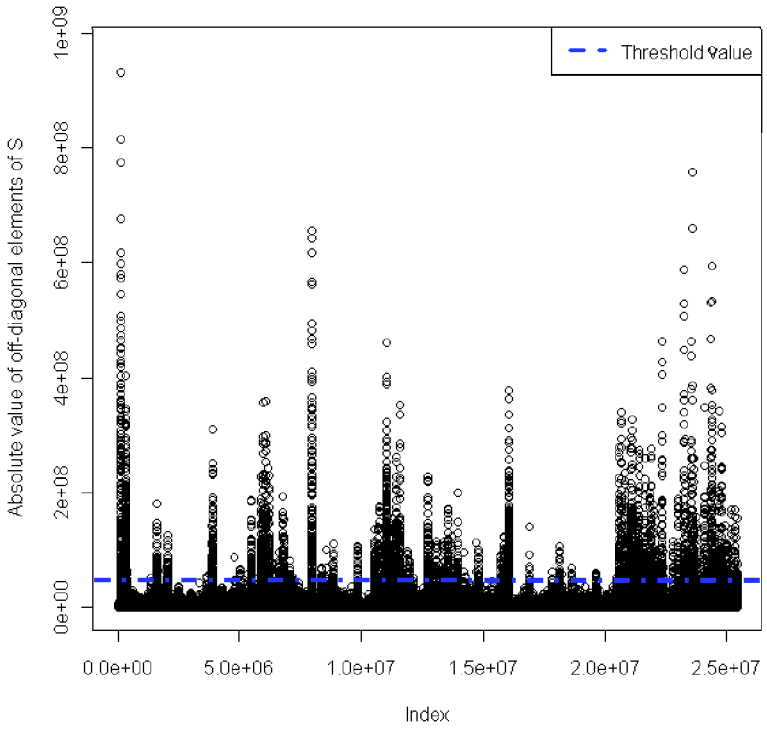

FIG. 2. *Plot of off-diagonal elements of* $\mathbf{S}$.

For the SLDA, to construct sparse estimates of $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ by thresholding, we applied the cross-validation method described in the end of Section 3 to choose the constants M_1 and M_2 in the thresholding values $t_n = M_1(n^{-1}\log p)^{0.5}$ and $a_n = M_2(n^{-1}\log p)^{0.3}$. Figure 3 shows the cross validation scores $\hat{R}_{\mathrm{SLDA}}(M_1, M_2)$ over a range of (M_1, M_2). The minimum cross validation score is achieved at $M_1 = 10^7$ and $M_2 = 300$. These thresholding values resulted in a $\tilde{\boldsymbol{\delta}}$ with exactly 2,492 nonzero components, which is about 35% of all components of $\hat{\boldsymbol{\delta}}$, and a $\tilde{\boldsymbol{\Sigma}}$ with exactly 227,083 nonzero elements, which is about 0.45% of all elements of $\mathbf{S}$. Note that the number of nonzero estimates of $\boldsymbol{\delta}$ is still much larger than $n = 72$, but the SLDA does not require it to be smaller than n. The resulting SLDA has an estimated (by cross validation) misclassification rate 0.0278. In fact, 1 of the 47 ALL cases and 1 of the 25 AML cases are misclassified under the cross validation evaluation of the SLDA.

For comparison, we carried out the LDA with a generalized inverse $\mathbf{S}^-$. In the leave-one-out cross-validation evaluation of the LDA, 2 of the 47 ALL cases and 5 of the 25 AML cases are misclassified by the LDA, which results in an estimated misclassification rate 0.0972. Compared with the LDA, the SLDA reduces the misclassification rate by nearly 70%. From Figure 5 of Fan and Fan (2008), the misclassification rate of the FAIR method, estimated by the average of 100 ran-

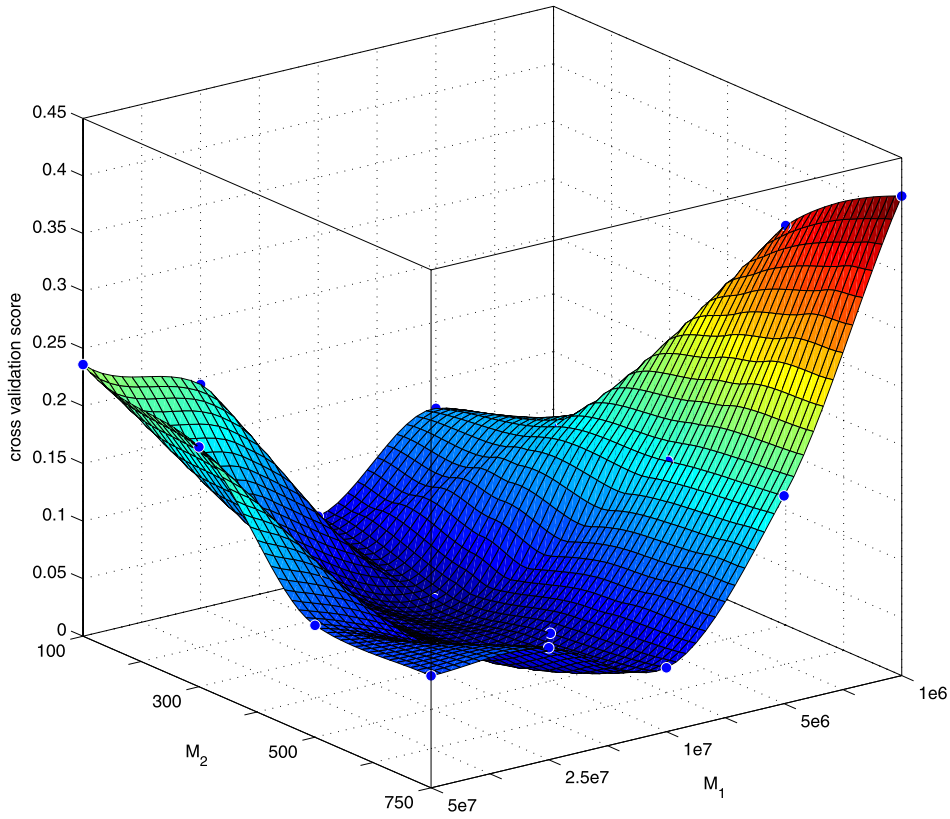

FIG. 3. *Cross-validation score vs* (M_1, M_2).

domly constructed cross validations with πn data points for constructing classifier and $(1-\pi)n$ data points for validation ($\pi = 0.4, 0.5$ and 0.6), ranges from 5% to 7%, which is smaller than the misclassification rate of the LDA but larger than the misclassification rate of the SLDA.

We also performed a simulation study on the conditional misclassification rate of SLDA under a population constructed using estimates from the real data set and a smaller dimension $p = 1{,}714$. The smaller dimension was used to reduce the computational cost and the 1,714 variables were chosen from the 7,129 variables with p-values (of the two sample t-tests for the mean effects) smaller than 0.05. In each of the 100 independently generated data sets, independent $\{\mathbf{x}_{1i}, i = 1, \ldots, 47\}$ and $\{\mathbf{x}_{2i}, i = 1, \ldots, 25\}$ were generated from $N_p(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}})$ and $N_p(\hat{\boldsymbol{\mu}}_2, \tilde{\boldsymbol{\Sigma}})$, respectively, where $p = 1{,}714$ and $\hat{\boldsymbol{\mu}}_k$ and $\tilde{\boldsymbol{\Sigma}}$ are estimates from the real data set. The sparse estimate $\tilde{\boldsymbol{\Sigma}}$ was used instead of the sample covariance matrix $\mathbf{S}$, because $\mathbf{S}$ is not positive definite. Since the population means and covariance matrix are known in the simulation, we were able to compute the conditional misclassification rate $R_{\mathrm{SLDA}}(\mathbf{X})$ for each generated data set. A boxplot of 100 values of $R_{\mathrm{SLDA}}(\mathbf{X})$ in the

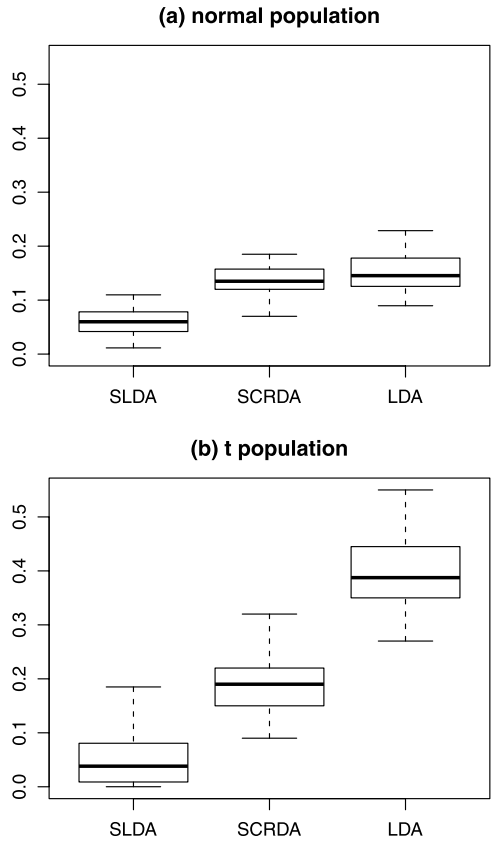

FIG. 4. *Boxplots of conditional misclassification rates of SLDA, SCRDA and LDA.*

simulation is given in Figure 4(a). The unconditional misclassification rate of the SLDA can be approximated by averaging over the 100 conditional misclassification rates. In this simulation, the unconditional misclassification rate for the SLDA is 0.069. Since the population is known in simulation, the optimal misclassification rate R_{OPT} is known to be 0.03.

For comparison, in the simulation we computed the conditional misclassification rates, $R_{\mathrm{LDA}}(\mathbf{X})$ for the LDA and $R_{\mathrm{SCRDA}}(\mathbf{X})$ for the shrunken centroids regularized discriminant analysis (SCRDA) proposed by Guo, Hastie and Tibshirani (2007). Since $R_{\mathrm{SCRDA}}(\mathbf{X})$ does not have an explicit form, it is approximated by an independent test data set of size 100 in each simulation run. Boxplots of $R_{\mathrm{LDA}}(\mathbf{X})$ and $R_{\mathrm{SCRDA}}(\mathbf{X})$ for 100 simulated data sets are included in Figure 4(a). It can be seen that the conditional misclassification rate of the LDA varies more than that of the SLDA. The unconditional misclassification rate for the LDA, approximated by the 100 simulated $R_{\mathrm{LDA}}(\mathbf{X})$ values, is 0.152, which indicates a 53% improvement of the SLDA over the LDA in terms of the unconditional misclassification

rate. The SCRDA has a simulated unconditional misclassification rate 0.137 and its performance is better than that of the LDA but worse than that of the SLDA. In this simulation, we also found that the conditional misclassification rate of the FAIR method was similar to that of the LDA.

To examine the performance of these classification methods in the case of non-normal data, we repeated the same simulation with the multivariate normal distribution replaced by the multivariate t-distribution with 3 degrees of freedom. The boxplots are given in Figure 4(b) and the simulated unconditional misclassification rates are 0.059, 0.194 and 0.399 for the SLDA, SCRDA and LDA, respectively. Since the t-distribution has a larger variability than the normal distribution, all conditional misclassification rates in the t-distribution case vary more than those in the normal distribution case.

6. Proofs.

PROOF OF THEOREM 1. (i) Let $\hat{\sigma}_{j,l}$ and $\sigma_{j,l}$ be the (j,l)th elements of $\mathbf{S}$ and $\mathbf{\Sigma}$, respectively. From result (10) in Bickel and Levina (2008), $\max_{j,l\le p}|\hat{\sigma}_{j,l}-\sigma_{j,l}| = O_P(\sqrt{\log p}/\sqrt{n})$. Then,

$$\|\mathbf{S}-\mathbf{\Sigma}\| \le \max_{j\le p}\sum_{l=1}^{p}|\hat{\sigma}_{j,l}-\sigma_{j,l}| = O_P\big(p\sqrt{\log p}/\sqrt{n}\big) = O_P(s_n),$$

where $\|\mathbf{A}\|$ is the norm of the matrix $\mathbf{A}$ defined as the maximum of all eigenvalues of $\mathbf{A}$. By (2)–(3) and $s_n \to 0$, $\mathbf{S}^{-1}$ exists and

$$\|\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\| = \|\mathbf{S}^{-1}(\mathbf{S}-\mathbf{\Sigma})\mathbf{\Sigma}^{-1}\| \le \|\mathbf{S}^{-1}\|\|\mathbf{S}-\mathbf{\Sigma}\|\|\mathbf{\Sigma}^{-1}\| = O_P(s_n).$$

Consequently,

$$\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}[1+O_P(s_n)] = \hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}}[1+O_P(s_n)].$$

Since $E[(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})'\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})] = O(p/n)$ and $E[\boldsymbol{\delta}'\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})]^2 \le \Delta_p^2 E[(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})'\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})]$, we have

$$\begin{aligned}
\hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}} &= \boldsymbol{\delta}'\mathbf{\Sigma}^{-1}\boldsymbol{\delta} + 2\boldsymbol{\delta}'\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}) + (\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})'\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}) \\
&= \Delta_p^2 + O_P\left(\frac{\sqrt{p}\Delta_p}{\sqrt{n}}\right) + O_P\left(\frac{p}{n}\right) \\
&= \Delta_p^2\left[1 + O_P\left(\frac{\sqrt{p}}{\sqrt{n}\Delta_p}\right) + O_P\left(\frac{p}{n\Delta_p^2}\right)\right] \\
&= \Delta_p^2[1+O_P(s_n)],
\end{aligned}$$

where the last equality follows from $\sqrt{p}/(s_n\sqrt{n}\Delta_p) = 1/(\sqrt{p\log p}\Delta_p) = O(1)$. Combining these results, we obtain that

$$\begin{aligned}
\hat{\boldsymbol{\delta}}\mathbf{S}^{-1}\hat{\boldsymbol{\delta}} &= \hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}}[1+O_P(s_n)] = \Delta_p^2[1+O_P(s_n)]^2 \\
&= \Delta_p^2[1+O_P(s_n)].
\end{aligned}$$

Then

$$\begin{aligned}\frac{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)-\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}/2}{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}}}&=-\frac{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}}}{2\sqrt{1+O_P(s_n)}}+\frac{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)}{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}}}\\&=-\frac{\sqrt{\Delta_p^2[1+O_P(s_n)]}}{2\sqrt{1+O_P(s_n)}}+O_P\left(\sqrt{\frac{p}{n}}\right)\\&=-\frac{\Delta_p}{2}[1+O_P(s_n)]+O_P\left(\sqrt{\frac{p}{n}}\right)\\&=-\frac{\Delta_p}{2}\left[1+O_P(s_n)+O_P\left(\frac{\sqrt{p}}{\sqrt{n}\Delta_p}\right)\right]\\&=-\frac{\Delta_p}{2}[1+O_P(s_n)].\end{aligned}$$

Similarly, we can show that

$$\frac{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}(\boldsymbol{\mu}_2-\bar{\mathbf{x}}_2)-\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}/2}{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\hat{\boldsymbol{\delta}}}}=-\frac{\Delta_p}{2}[1+O_P(s_n)].$$

These results and formula (5) imply the result in (i).

(ii) Let ϕ be the density of Φ. By the result in (i),

$$R_{\mathrm{LDA}}(\mathbf{X})-R_{\mathrm{OPT}}=\phi(\omega_n)O_P(s_n),$$

where ω_n is between $-\Delta_p/2$ and $-[1+O_P(s_n)]\Delta_p/2$. Since $\phi(\omega_n)$ is bounded by a constant, the result follows from the fact that R_{OPT} is bounded away from 0 when Δ_p is bounded.

(iii) When $\Delta_p\to\infty$, $R_{\mathrm{OPT}}\to 0$, and, by the result in (i), $R_{\mathrm{LDA}}(\mathbf{X})\to_P 0$.

(iv) If $\Delta_p\to\infty$, then, by Lemma 1 and the condition $s_n\Delta_p^2\to 0$, we conclude that $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}}\to_P 1$. □

PROOF OF LEMMA 1. It follows from result (22) that

$$\begin{aligned}\frac{\xi_n(1-\tau_n)}{1+\xi_n(1-\tau_n)^2}e^{[\xi_n-\xi_n(1-\tau_n)^2]/2}&\le\frac{\Phi(-\sqrt{\xi_n}(1-\tau_n))}{\Phi(-\sqrt{\xi_n})}\\&\le\frac{1+\xi_n}{\xi_n(1-\tau_n)}e^{[\xi_n-\xi_n(1-\tau_n)^2]/2}.\end{aligned}$$

Since $\xi_n\to\infty$ and $\tau_n\to 0$,

$$\frac{\xi_n(1-\tau_n)}{1+\xi_n(1-\tau_n)^2}\to 1\quad\text{and}\quad\frac{1+\xi_n}{\xi_n(1-\tau_n)}\to 1.$$

The result follows from $[\xi_n-\xi_n(1-\tau_n)^2]/2=\xi_n\tau_n(1-\tau_n/2)\to\gamma$ regardless of whether γ is 0, positive, or ∞. □

PROOF OF THEOREM 2. For simplicity, we prove the case of $n_1 = n_2 = n/2$.

(i) The conditional misclassification rate of the LDA in this case is given by (5) with $\hat{\mathbf{\Sigma}}$ replaced by $\mathbf{\Sigma}$. Note that $\mathbf{\Sigma}^{-1/2}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_k) \sim N_p(\mathbf{0}, n_1^{-1}\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of order p. Let ζ_j be the jth component of $\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}$. Then, $\sum_{j=1}^p \zeta_j^2 = \Delta_p^2$ and the jth component of $\mathbf{\Sigma}^{-1/2}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_k)$ is $n_1^{-1/2}\varepsilon_{kj}$, and the jth component of $\mathbf{\Sigma}^{-1/2}\hat{\boldsymbol{\delta}}$ is $\zeta_j + n_1^{-1/2}(\varepsilon_{1j} - \varepsilon_{2j})$, $j = 1, \ldots, p$, where ε_{kj}, $j = 1, \ldots, p$, $k = 1, 2$, are independent standard normal random variables. Consequently,

$$\begin{aligned}
\hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1) - \hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}}/2 &= \sum_{j=1}^p \left(-\frac{\zeta_j^2}{2} + \frac{\varepsilon_{1j}^2 - \varepsilon_{2j}^2}{n} + \frac{\zeta_j \varepsilon_{2j}}{\sqrt{n_1}} \right) \\
&= -\frac{\Delta_p^2}{2} + \frac{1}{n}\sum_{j=1}^p (\varepsilon_{1j}^2 - \varepsilon_{2j}^2) + \frac{1}{\sqrt{n_1}}\sum_{j=1}^p \zeta_j \varepsilon_{2j} \\
&= -\frac{\Delta_p^2}{2} + O_P\left(\frac{\sqrt{p}}{n}\right) + O_P\left(\frac{\Delta_p}{\sqrt{n}}\right)
\end{aligned}$$

and

$$\begin{aligned}
\hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}} &= \sum_{j=1}^p \left(\zeta_j + \frac{\varepsilon_{1j} - \varepsilon_{2j}}{\sqrt{n_1}} \right)^2 \\
&= \Delta_p^2 + \frac{1}{n_1}\sum_{j=1}^p (\varepsilon_{1j} - \varepsilon_{2j})^2 + \frac{2}{\sqrt{n_1}}\sum_{j=1}^p \zeta_j(\varepsilon_{1j} - \varepsilon_{2j}) \\
&= \Delta_p^2 + \frac{4p}{n}[1 + o_P(1)] + O_P\left(\frac{\Delta_p}{\sqrt{n}}\right) \\
&= \Delta_p^2 + \frac{4p}{n}[1 + o_P(1)],
\end{aligned}$$

where the last equality follows from $\Delta_p^2 = O(p)$ under (2)–(3). Combining these results, we obtain that

$$\frac{\hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1) - \hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}}/2}{\sqrt{\hat{\boldsymbol{\delta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\delta}}}} = -\frac{\Delta_p^2}{2\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}} + o_P(1). \tag{25}$$

Similarly, we can prove that (25) still holds if $\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1$ is replaced by $\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_2$. If $\Delta_p^2/\sqrt{p/n} \to 0$, then the quantity in (25) converges to 0 in probability. Hence, $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 1/2$.

(ii) Since $p/n \to \infty$, $\Delta_p^2/(p/n) \to 0$. Then, the quantity in (25) converges to $-c/4$ in probability and, hence, $R_{\mathrm{LDA}}(\mathbf{X}) \to_P \Phi(-c/4)$, which is a constant between 0 and 1/2. Since $\Delta_p \to \infty$, $R_{\mathrm{OPT}} \to 0$ and, hence, $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.

(iii) When $\Delta_p^2/\sqrt{p/n} \to \infty$, it follows from (25) that the quantity on the left-hand side of (25) diverges to $-\infty$ in probability. This proves that $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 0$. To show $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$, we need a more refined analysis. The quantity on the left-hand side of (25) is equal to

$$-\frac{\Delta_p^2 + O_P(\sqrt{p}/n) + O_P(\Delta_p/\sqrt{n})}{2\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}} = -\frac{\Delta_p}{2}(1 - \tau_n),$$

where

$$\tau_n = 1 - \frac{\Delta_p + O_P(\sqrt{p}/n)/\Delta_p + O_P(1/\sqrt{n})}{\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}}$$

and $P(0 \le \tau_n \le 1) \to 1$. Note that

$$\begin{aligned}\tau_{1n} &= 1 - \frac{\Delta_p}{\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}} \\ &= \frac{(4p/n)[1 + o_P(1)]}{\Delta_p^2 + (4p/n)[1 + o_P(1)] + \Delta_p\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}}\end{aligned}$$

and

$$\begin{aligned}\tau_{2n} &= \frac{O_P(\sqrt{p}/n)/\Delta_p + O_P(1/\sqrt{n})}{\sqrt{\Delta_p^2 + (4p/n)[1 + o_P(1)]}} = \frac{O_P(\sqrt{p}/n)}{\Delta_p^2} + \frac{O_P(1/\sqrt{n})}{\Delta_p} \\ &= \frac{O_P(\sqrt{p/n})}{\Delta_p^2}\end{aligned}$$

under (2) and (3). Then

$$\tau_n\Delta_p^2 = \tau_{1n}\Delta_p^2 + \tau_{2n}\Delta_p^2 = \tau_{1n}\Delta_p^2 + O_P\big(\sqrt{p/n}\big).$$

If $\Delta_p^2/(p/n)$ is bounded, then $\tau_{1n} \ge c$ for a constant $c > 0$ and

$$\tau_n\Delta_p^2 \ge c\Delta_p^2 + O_P\big(\sqrt{p/n}\big),$$

which diverges to ∞ in probability since $\Delta_p^2/\sqrt{p/n} \to \infty$. If $\Delta_p^2/(p/n) \to \infty$, then $\tau_{1n}\Delta_p^2 \ge cp/n$ for a constant $c > 0$ and

$$\tau_n\Delta_p^2 \ge cp/n + O_P\big(\sqrt{p/n}\big),$$

which diverges to ∞ in probability since $p/n \to \infty$. Thus, $\tau_n\Delta_p^2 \to \infty$ in probability, and the result follows from Lemma 1. □

PROOF OF LEMMA 2. (i) It follows from (22) that, for all t,

$$P(|\hat{\delta}_j - \delta_j| > t) \le c_1 e^{-c_2 n t^2},$$

where c_1 and c_2 are positive constants. Then, the probability in (11) is

$$\begin{aligned} 1 - P\Bigg(\bigcup_{1\le j\le p, |\delta_j|\le a_n/r} \{|\hat{\delta}_j| > a_n\}\Bigg) &\ge 1 - \sum_{j=1}^{p} P\big(|\hat{\delta}_j - \delta_j| > a_n(r-1)/r\big) \\ &\ge 1 - pc_1 e^{-c_2 n a_n^2 (r-1)^2/r^2}. \end{aligned}$$

Because

$$\frac{na_n^2}{\log p} = \left(\frac{n}{\log p}\right)^{1-2\alpha} \to \infty$$

when $\alpha < 1/2$, we conclude that $pc_1 e^{-c_2 n a_n^2 (r-1)^2/r^2} \to 0$, and thus (11) holds. The proof of (12) is similar since

$$\begin{aligned} 1 - P\Bigg(\bigcup_{1\le j\le p, |\delta_j|> r a_n} \{|\hat{\delta}_j| \le a_n\}\Bigg) &\ge 1 - \sum_{j=1}^{p} P\big(|\hat{\delta}_j - \delta_j| > a_n(r-1)\big) \\ &\ge 1 - pc_1 e^{-c_2 n a_n^2 (r-1)^2}. \end{aligned}$$

(ii) The result follows from results (11) and (12). □

PROOF OF THEOREM 3. The conditional misclassification rate $R_{\mathrm{SLDA}}(\mathbf{X})$ is given by

$$\frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^k \tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_k - \bar{\mathbf{x}}_k) - \hat{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}/2}{\sqrt{\tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}}}\right).$$

From result (8),

$$\tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}[1 + O_P(d_n)] = \tilde{\boldsymbol{\delta}}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\delta}}[1 + O_P(d_n)].$$

Without loss of generality, we assume that $\tilde{\boldsymbol{\delta}} = (\tilde{\boldsymbol{\delta}}_1', \mathbf{0}')'$, where $\tilde{\boldsymbol{\delta}}_1$ is the $\hat{q}$-vector containing nonzero components of $\tilde{\boldsymbol{\delta}}$. Let $\boldsymbol{\delta} = (\boldsymbol{\delta}_1', \boldsymbol{\delta}_0')'$, where $\boldsymbol{\delta}_1$ has dimension $\hat{q}$. From Lemma 2(ii), $\|\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1\|^2 = O_P(q_n/n)$ and, with probability tending to 1,

$$\|\boldsymbol{\delta}_0\|^2 = \sum_{j:|\hat{\delta}_j|\le a_n} \delta_j^2 \le \sum_{j:|\delta_j|\le r a_n} \delta_j^2 \le (r a_n)^{2(1-g)} \sum_{j:|\delta_j|\le r a_n} \delta_j^{2g} = O\big(a_n^{2(1-g)} D_{g,p}\big).$$

Let $k_n = \max\{a_n^{2(1-g)} D_{g,p}, q_n/n\}$. Then $\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 = \|\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1\|^2 + \|\boldsymbol{\delta}_0\|^2 = O_P(k_n)$. This together with (2)–(3) implies that $(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta})' \boldsymbol{\Sigma}^{-1} (\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}) = O_P(k_n)$,

and hence

$$\begin{aligned}\tilde{\boldsymbol{\delta}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\delta}} &= \Delta_p^2 + 2\boldsymbol{\delta}'\boldsymbol{\Sigma}(\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta}) + (\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta}) \\ &= \Delta_p^2[1 + O_P(\sqrt{k_n}/\Delta_p) + O_P(k_n/\Delta_p^2)] \\ &= \Delta_p^2[1 + O_P(\sqrt{k_n}/\Delta_p)].\end{aligned}$$

Write

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_2 \end{pmatrix},$$
$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_1 & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{12}' & \tilde{\boldsymbol{\Sigma}}_2 \end{pmatrix}, \qquad \tilde{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} \tilde{\mathbf{C}}_1 & \tilde{\mathbf{C}}_{12} \\ \tilde{\mathbf{C}}_{12}' & \tilde{\mathbf{C}}_2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_1$, $\tilde{\boldsymbol{\Sigma}}_1$, $\mathbf{C}_1$ and $\tilde{\mathbf{C}}_1$ are $q_n \times q_n$ matrices with q_n defined in Lemma 2(ii). Then

$$\mathbf{C}_{12} = -\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{C}_2 \quad \text{and} \quad \tilde{\mathbf{C}}_{12} = -\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2.$$

If $\check{\boldsymbol{\delta}}_1 = (\tilde{\boldsymbol{\delta}}_1', \mathbf{0}')'$ and $\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1 = (\boldsymbol{\xi}_1', \boldsymbol{\xi}_0')'$, where $\check{\boldsymbol{\delta}}_1$ and $\boldsymbol{\xi}_1$ have dimension q_n, then

$$\tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1) = \check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1 + \check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_{12}\boldsymbol{\xi}_0 = \check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1 - \check{\boldsymbol{\delta}}_1'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0.$$

Since $\boldsymbol{\xi}_1$ has dimension q_n,

$$(\check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1)^2 \le (\boldsymbol{\xi}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1)(\check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\check{\boldsymbol{\delta}}_1) = (\boldsymbol{\xi}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1)(\tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\delta}}) = O_P(q_n/n)(\tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\delta}})$$

and hence

$$\check{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\boldsymbol{\xi}_1 = O_P(\sqrt{k_n})\sqrt{\tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\delta}}}.$$

Since $\tilde{\boldsymbol{\Sigma}}_1^{-1} \le \tilde{\mathbf{C}}_1$,

$$\begin{aligned}(\check{\boldsymbol{\delta}}_1'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0)^2 &\le (\check{\boldsymbol{\delta}}_1'\tilde{\boldsymbol{\Sigma}}_1^{-1}\check{\boldsymbol{\delta}}_1)(\boldsymbol{\xi}_0'\tilde{\mathbf{C}}_2\tilde{\boldsymbol{\Sigma}}_{12}'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0) \\ &\le (\tilde{\boldsymbol{\delta}}_1'\tilde{\mathbf{C}}_1\tilde{\boldsymbol{\delta}}_1)(\boldsymbol{\xi}_0'\tilde{\mathbf{C}}_2\tilde{\boldsymbol{\Sigma}}_{12}'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0) \\ &= (\tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\delta}})(\boldsymbol{\xi}_0'\tilde{\mathbf{C}}_2\tilde{\boldsymbol{\Sigma}}_{12}'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0).\end{aligned}$$

From result (8),

$$\boldsymbol{\xi}_0'\tilde{\mathbf{C}}_2\tilde{\boldsymbol{\Sigma}}_{12}'\tilde{\boldsymbol{\Sigma}}_1^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\mathbf{C}}_2\boldsymbol{\xi}_0 = \boldsymbol{\xi}_0'\mathbf{C}_2\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{C}_2\boldsymbol{\xi}_0[1 + O_P(d_n)].$$

Under condition (2), all eigenvalues of sub-matrices of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ are bounded by c_0. Repeatedly using condition (2), we obtain that

$$\begin{aligned}E(\boldsymbol{\xi}_0'\mathbf{C}_2\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{C}_2\boldsymbol{\xi}_0) &\le c_0 E(\boldsymbol{\xi}_0'\mathbf{C}_2\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{12}\mathbf{C}_2\boldsymbol{\xi}_0) \\ &= c_0 n^{-1}\operatorname{trace}(\boldsymbol{\Sigma}_{12}\mathbf{C}_2\boldsymbol{\Sigma}_2\mathbf{C}_2\boldsymbol{\Sigma}_{12}') \\ &\le c_0^4 n^{-1}\operatorname{trace}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}')\end{aligned}$$

$$
\begin{aligned}
&= \frac{c_0^4}{n} \sum_{j=1}^{q_n} \sum_{l=q_n+1}^{p} \sigma_{jl}^2 \\
&\le \frac{c_0^{6-h} q_n}{n} \max_{l \le p} \sum_{j=1}^{p} |\sigma_{jl}|^h \\
&= O(C_{h,p} q_n / n),
\end{aligned}
$$

where h and $C_{h,p}$ are given in (6). This proves that

$$
\frac{\tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1)}{\sqrt{\tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}}} = \frac{O_P(\sqrt{k_n}) + O_P(\sqrt{C_{h,p} q_n / n})}{\sqrt{1 + O_P(d_n)}},
$$

which also holds when $\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1$ is replaced by $\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2$ or $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}$. Note that

$$
\begin{aligned}
\hat{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} &= \tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} + (\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} \\
&= \tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}} + \Delta_p O_P(\sqrt{k_n}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{(-1)^k \tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu}_k - \bar{\mathbf{x}}_k) - \hat{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}/2}{\sqrt{\tilde{\boldsymbol{\delta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\delta}}}} &= \frac{O_P(\sqrt{k_n}) + O_P(\sqrt{C_{h,p} q_n / n})}{\sqrt{1 + O_P(d_n)}} \\
&\quad - \frac{\Delta_p \sqrt{1 + O_P(\sqrt{k_n}/\Delta_p)}}{2\sqrt{1 + O_P(d_n)}} \\
&= O_P(\sqrt{k_n}) + O_P\bigl(\sqrt{C_{h,p} q_n / n}\bigr) \\
&\quad - \frac{\Delta_p}{2} [1 + O_P(\sqrt{k_n}/\Delta_p) + O_P(d_n)] \\
&= -\frac{\Delta_p}{2} \biggl[1 + O_P\biggl(\frac{\sqrt{C_{h,p} q_n}}{\Delta_p \sqrt{n}} \biggr) \\
&\qquad + O_P\biggl(\frac{\sqrt{k_n}}{\Delta_p} \biggr) + O_P(d_n) \biggr] \\
&= -\frac{\Delta_p}{2} [1 + O_P(b_n)].
\end{aligned}
$$

This proves the result in (i). The proofs of (ii)–(iv) are the same as the proofs for Theorem 1(ii)–(iv) with s_n replaced by b_n. This completes the proof. □

Acknowledgments. The authors would like to thank two referees and an associate editor for their helpful comments and suggestions, and Dr. Weidong Liu for his help in correcting an error in the proof of Theorem 3.

REFERENCES

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

CLEMMENSEN, L., HASTIE, T. and ERSBØLL, B. (2008). Sparse discriminant analysis. Technical report, Technical Univ. Denmark and Stanford Univ.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369. MR1323344

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FANG, K. T. and ANDERSON, T. W. (eds.) (1990). *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York. MR1066887

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. and LANDER, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2007). Regularized linear discriminant analysis and its applications in microarrays. *Biostatistics* **8** 86–100.

KOHAVI, R. and JOHN, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97** 273–324.

QIAO, Z., ZHOU, L. and HUANG, J. Z. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG Int. J. Appl. Math.* **39** 48–60. MR2493149

ZHANG, Q. and WANG, H. (2010). On BIC's selection consistency for discriminant analysis. *Statist. Sinica* **20**. To appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1300 UNIVERSITY AVE.
MADISON, WISCONSIN 53706
USA
E-MAIL: shao@stat.wisc.edu
yzwang@stat.wisc.edu
xdeng@stat.wisc.edu
wangs@stat.wisc.edu