High Dimensional Classification Using Features Annealed Independence Rules *

Jianqing Fan
Princeton University

Yingying Fan
Harvard University

May 16, 2007

ABSTRACT. Classification using high-dimensional features arises frequently in many contemporary statistical studies such as tumor classification using microarray or other high-throughput data. The impact of dimensionality on classifications is largely poorly understood. In a seminal paper, Bickel and Levina (2004) show that the Fisher discriminant performs poorly due to diverging spectra and they propose to use the independence rule to overcome the problem. We first demonstrate that even for the independence classification rule, classification using all the features can be as bad as the random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. In fact, we demonstrate further that almost all linear discriminants can perform as bad as the random guessing. Thus, it is paramountly important to select a subset of important features for high-dimensional classification, resulting in Features Annealed Independence Rules (FAIR). The conditions under which all the important features can be selected by the two-sample t-statistic are established. The choice of the optimal number of features, or equivalently, the threshold value of the test statistics are proposed based on an upper bound of the classification error. Simulation studies and real data analysis support our theoretical results and demonstrate convincingly the advantage of our new classification procedure.

Short Title: Features Annealed Independent Rules.

AMS 2000 subject classifications. Primary 62G08; secondary 62J12, 62F12.

Key words and phrases. Classification, feature extraction, high dimensionality, independence rule, misclassification rates.

*Financial support from the NSF grants DMS-0354223 and DMS-0704337 and NIH grant R01-GM072611 is gratefully acknowledged. The authors acknowledge gratefully the helpful comments of referees that led to the improvement of the presentation and the results of the paper. Address for correspondence: Yingying Fan, Department of Statistics, Harvard University, Cambridge, MA 02138. Phone: (609) 258-9433. E-mail: yingying@princeton.edu.

1 Introduction

With rapid advance of imaging technology, high-throughput data such as microarray and proteomics data are frequently seen in many contemporary statistical studies. For instance, in the analysis of Microarray data, the dimensionality is frequently thousands or more, while the sample size is typically in the order of tens (West *et al.*, 2001; Dudoit *et al.*, 2002). See Fan and Ren (2006) for an overview. The large number of features presents an intrinsic challenge to classification problems. For an overview of statistical challenges associated with high dimensionality, see Fan and Li (2006).

Classical methods of classification break down when the dimensionality is extremely large. For example, even when the covariance matrix is known, Bickel and Levina (2004) demonstrate convincingly that the Fisher discriminant analysis performs poorly in a minimax sense due to the diverging spectra (*e.g.* , the condition number goes to infinity as dimensionality diverges) frequently encountered in the high-dimensional covariance matrices. Even if the true covariance matrix is not ill conditioned, the singularity of the sample covariance matrix will make the Fisher discrimination rule inapplicable when the dimensionality is larger than sample size. Bickel and Levina (2004) show that the independence rule overcomes the above two problems. However, in tumor classification using microarray data, we hope to find tens of genes that have high discriminative power. The independence rule, studied by Bickel and Levina (2004), does not possess this kind of properties.

The difficulty of high-dimensional classification is intrinsically caused by the existence of many noise features that do not contribute to the reduction of misclassification rate. Though the importance of dimension reduction and feature selection has been stressed and many methods have been proposed in the literature, very little research has been done on theoretical analysis of the impacts of high dimensionality on classifi-

cation. For example, using most discrimination rules such as the linear discriminants, we need to estimate the population mean vectors from the sample. When the dimensionality is high, even though each component of the population mean vectors can be estimated with accuracy, the aggregated estimation error can be very large and this has adverse effects on the misclassification rate. Therefore, when there is only a fraction of features that account for most of the variation in the data such as tumor classification using gene expression data, using all features will increase the misclassification rate.

To illustrate the idea, we study independence classification rule. Specifically, we give an explicit formula on how the signal and noise affect the misclassification rates. We show formally how large the signal to noise ratio can be such that the effect of noise accumulation can be ignored, and how small this ratio can be before the independence classifier performs as bad as the random guessing. Indeed, as demonstrated in Section 2, the impact of the dimensionality can be very drastic. For the independence rule, the misclassification rate can be as high as the random guessing even when the problem is perfectly classifiable. In fact, we demonstrate that almost all linear discriminants can not perform any better than random guessing, due to the noise accumulation in the estimation of the population mean vectors, unless the signals are very strong, namely the population mean vectors are very far apart.

The above discussion reveals that feature selection is necessary for high-dimensional classification problems. When the independence rule is applied to selected features, the resulting Feature Annealed Independent Rules (FAIR) overcome both the issues of interpretability and the noise accumulation. One can extract the important features via variable selection techniques such as the penalized quasi-likelihood function. See Fan and Li (2006) for an overview. One can also employ a simple two-sample t-test as in Tibshirani *et al.*(2002) to identify important genes for the tumor classification, resulting

in the nearest shrunken centroids method. Such a simple method corresponds to a componentwise regression method or a ridge regression method with ridge parameters tending to ∞ (Fan and Lv, 2007). Hence, it is a specific and useful example of the penalized quasi-likelihood method for feature selection. It is surprising that such a simple proposal can indeed extract all important features. Indeed, we demonstrate that under suitable conditions, the two sample t-statistic can identify all the features that efficiently characterize both classes.

Another popular class of the dimension reduction methods is projection. They have been widely applied to the classification based on the gene expression data. See, for example, principal component analysis in Ghosh (2002), Zou *et al.*(2004), and Bair *et al.*(2004); partial least squares in Nguyen and Rocke (2002), Huang and Pan (2003), and Boulesteix(2004); and sliced inverse regression in Chiaromonte and Martinelli (2002), Antoniadis *et al.*(2003), and Bura and Pfeiffer (2003). These projection methods attempt to find directions that can result in small classification errors. In fact, the directions found by these methods usually put much more weights on features that have large classification power. In general, however, linear projection methods are likely to perform poorly unless the projection vector is sparse, namely, the effective number of selected features is small. This is due to the aforementioned noise accumulation prominently featured in high-dimensional problems, recalling discrimination based on linear projections onto almost all directions can perform as bad as the random guessing.

As direct application of the independence rule is not efficient, we propose a specific form of FAIR. Our FAIR selects the statistically most significant m features according to the componentwise two-sample t-statistics between two classes, and applies the independence classifiers to these m features. Interesting questions include how to choose the optimal m, or equivalently, the threshold value of t-statistic, such that the classification

error is minimized, and how this classifier performs compared with the independence rule without feature selection and the oracle-assisted FAIR. All these questions will be formally answered in this paper. Surprisingly, these results are similar to those for the adaptive Neyman test in Fan (1996). The theoretical results also indicate that FAIR without oracle information performs worse than the one with oracle information, and the difference of classification error depends on the threshold value, which is consistent with the common sense.

There is a huge literature on classification. To name a few in addition to those mention before, Bai and Saranadasa (1996) dealt with the effect of high dimensionality in a two-sample problem from a hypothesis testing viewpoint; Friedman (1989) proposed a regularized discriminant analysis to deal with the problems associated with high dimension while performing computations in the regular way; Dettling and Bühlmann (2003) and Bühlmann and Yu (2003) study boosting with logit loss and L_2 loss, respectively, and demonstrate the good performances of these methods in high-dimensional setting; Greenshtein and Ritov (2004), Greenshtein (2006) and Meinshausen (2005) introduced and studied the concept of persistence, which places more emphasis on misclassification rates or expected loss rather than the accuracy of estimated parameters.

This article is organized as follows. In Section 2, we demonstrate the impact of dimensionality on the independence classification rule, and show that discrimination based on projecting observations onto almost all linear directions is nearly the same as random guessing. We establish, in Section 3, the conditions under which two sample t-test can identify all the important features with probability tending to 1. In Section 4, we propose FAIR and give an upper bound of its classification error. Simulation studies and real data analyses are conducted in Section 5. The conclusion of our study is summarized in Section 6. All proofs are given in the Appendix.

2 Impact of High Dimensionality

Consider the p-dimensional classification problem between two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Suppose that from class $\mathcal{C}_k$, we have n_k observations $\boldsymbol{Y}_{k1}, \cdots, \boldsymbol{Y}_{kn_k}$ in $\mathbb{R}^p$. The j-th feature of the i-th sample from class $\mathcal{C}_k$ satisfies the model

$$Y_{kij} = \mu_{kj} + \epsilon_{kij},\ k = 1, 2,\ i = 1, \cdots, n_k,\ j = 1, \cdots, p, \tag{2.1}$$

where μ_{kj} is the mean effect of the j-th feature in class $\mathcal{C}_k$ and ϵ_{kij} is the corresponding Gaussian random noise for i-th observation. In matrix notation, the above model can be written as

$$\boldsymbol{Y}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_{ki},\ k = 1, 2,\ i = 1, \cdots, n_k,$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \cdots, \mu_{kp})'$ is the mean vector of class $\mathcal{C}_k$ and $\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, \cdots, \epsilon_{kip})'$ has the distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_k)$. We assume that all observations are independent across samples and in addition, within class $\mathcal{C}_k$, observations $\boldsymbol{Y}_{k1}, \cdots, \boldsymbol{Y}_{kn_k}$ are also identically distributed. Throughout this paper, we make the assumption that the two classes have compatible sample sizes, i.e., $c_1 \leq n_1/n_2 \leq c_2$ with c_1 and c_2 some positive constants.

We first investigate the impact of high dimensionality on classification. For simplicity, we temporarily assume that the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same covariance matrix $\boldsymbol{\Sigma}$. To illustrate our idea, we consider the independence classification rule, which classifies the new feature vector $\boldsymbol{x}$ into class $\mathcal{C}_1$ if

$$\delta(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})'\mathbf{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0,$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ and $\mathbf{D} = \text{diag}(\boldsymbol{\Sigma})$. This classifier has been thoroughly studied in Bickel and Levina (2004). They showed that in the classification of two normal populations, this independence rule greatly outperforms the Fisher linear discriminant rule under broad conditions when the number of variables is large.

The independence rule depends on the marginal parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{D} = \text{diag}\{\sigma_1^2, \cdots, \sigma_p^2\}$. They can easily be estimated from the samples:

$$\widehat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n_k} \boldsymbol{Y}_{ki}/n_k, \ k = 1, 2, \quad \widehat{\boldsymbol{\mu}} = (\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)/2$$

and

$$\widehat{\mathbf{D}} = \text{diag}\{(S_{1j}^2 + S_{2j}^2)/2, \ j = 1, \cdots, p\},$$

where $S_{kj}^2 = \sum_{i=1}^{n_k}(Y_{kij} - \bar{Y}_{kj})^2/(n_k - 1)$ is the sample variance of the j-th feature in class k and $\bar{Y}_{kj} = \sum_{i=1}^{n_k} Y_{ki}/n_k$. Hence, the plug-in discrimination function is

$$\hat{\delta}(\boldsymbol{x}) = (\boldsymbol{x} - \widehat{\boldsymbol{\mu}})'\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2).$$

Denote the parameter by $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. If we have a new observation $\boldsymbol{X}$ from class $\mathcal{C}_1$, then the misclassification rate of $\hat{\delta}$ is

$$W(\hat{\delta}, \boldsymbol{\theta}) = P(\hat{\delta}(\boldsymbol{X}) \leq 0 | \boldsymbol{Y}_{ki}, i = 1, \cdots, n_k, \ k = 1, 2) = 1 - \Phi(\Psi), \tag{2.2}$$

where

$$\Psi = \frac{(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})'\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{\sqrt{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)'\widehat{\mathbf{D}}^{-1}\boldsymbol{\Sigma}\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}},$$

and $\Phi(\cdot)$ is the standard Gaussian distribution function. The worst case classification error is

$$W(\hat{\delta}) = \max_{\boldsymbol{\theta} \in \Gamma} W(\hat{\delta}, \boldsymbol{\theta}),$$

where Γ is some parameter space to be defined. Let $n = n_1 + n_2$. In our asymptotic analysis, we always consider the misclassification rate of observations from $\mathcal{C}_1$, since the misclassification rate of observations from $\mathcal{C}_2$ can be easily obtained by interchanging n_1 with n_2 and $\boldsymbol{\mu}_1$ with $\boldsymbol{\mu}_2$. The high dimensionality is modeled through its dependence

on n, namely $p_n \to \infty$. However, we will suppress its dependence on n whenever there is no confusion.

Let $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$ be the correlation matrix, and $\lambda_{\max}(\mathbf{R})$ be its largest eigenvalue, and $\boldsymbol{\alpha} \equiv (\alpha_1, \cdots, \alpha_p)' = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Consider the parameter space

$$\Gamma = \{(\boldsymbol{\alpha}, \mathbf{\Sigma}) : \boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha} \geq C_p, \lambda_{\max}(\mathbf{R}) \leq b_0, \min_{1\leq j\leq p, k=1,2} \sigma_{kj}^2 > 0\}$$

where C_p is a deterministic positive sequence that depends only on the dimensionality p, and b_0 is a positive constant. Note that $\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}$ corresponds to the overall strength of signals, and the first condition $\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha} \geq C_p$ imposes a lower bound on the strength of signals. The second condition $\lambda_{\max}(\mathbf{R}) \leq b_0$ requires that the maximum eigenvalue of $\mathbf{R}$ should not exceed a positive constant. But since there are no restrictions on the smallest eigenvalue of $\mathbf{R}$, the condition number can still diverge. The third condition $\min_{1\leq j\leq p, k=1,2} \sigma_{kj}^2 > 0$ ensures that there are no deterministic features that make classification trivial and the diagonal matrix $\mathbf{D}$ is always invertible. We will consider the asymptotic behavior of $W(\hat{\delta}, \boldsymbol{\theta})$ and $W(\hat{\delta})$.

Theorem 1 *Suppose that* $\log p = o(n)$, $n = o(p)$ *and* $nC_p \to \infty$. *Then*

(i) The classification error $W(\delta, \boldsymbol{\theta})$ *with* $\boldsymbol{\theta} \in \Gamma$ *is bounded from above as*

$$W(\hat{\delta}, \boldsymbol{\theta}) \leq 1 - \Phi\left(\frac{[n_1n_2/(pn)]^{1/2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}(1 + o_P(1)) + \sqrt{p/(nn_1n_2)}(n_1 - n_2)}{2\sqrt{\lambda_{\max}(\mathbf{R})}\big\{1 + n_1n_2/(pn)\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}(1 + o_P(1))\big\}^{1/2}}\right).$$

(ii) Suppose $p/(nC_p) \to 0$. *For the worst case classification error* $W(\delta)$, *we have*

$$W(\hat{\delta}) = 1 - \Phi\Big(\frac{1}{2}[n_1n_2/(pnb_0)]^{1/2}C_p\{1 + o_P(1)\}\Big).$$

Specifically, when $\{\frac{n_1n_2}{pn}\}^{1/2}C_p \to C_0$ *with* C_0 *a nonnegative constant, then*

$$W(\hat{\delta}) \xrightarrow{P} 1 - \Phi(C_0/(2\sqrt{b_0})).$$

In particular, if $C_0 = 0$, *then* $W(\hat{\delta}) \xrightarrow{P} \frac{1}{2}$.

Theorem 1 reveals the trade-off between the signal strength C_p and the dimensionality, reflected in the term $C_p/\sqrt{p}$ when all features are used for classification. It states that the independence rule $\hat{\delta}$ would be no better than the random guessing due to noise accumulation, unless the signal levels are extremely high, say, $\{\frac{n}{p}\}^{1/2}C_p \geq B$ for some $B > 0$. Indeed, discrimination based on linear projections to almost all directions performs nearly the same as random guessing, as shown in the theorem below. The poor performance is caused by noise accumulation in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Theorem 2 *Suppose that* $\mathbf{a}$ *is a* p*-dimensional uniformly distributed unit random vector on a* $(p-1)$*-dimensional sphere. Let* $\lambda_1, \cdots, \lambda_p$ *be the eigenvalues of the covariance matrix* $\boldsymbol{\Sigma}$*. Suppose* $\lim_p \frac{1}{p^2}\sum_{j=1}^p \lambda_j^2 < \infty$ *and* $\lim_p \frac{1}{p}\sum_{j=1}^p \lambda_j = \tau$ *with* τ *a positive constant. Moveover, assume that* $p^{-1}\boldsymbol{\alpha}'\boldsymbol{\alpha} \to 0$*. Then if we project all the observations onto the vector* $\mathbf{a}$ *and use the classifier*

$$\hat{\delta}_{\mathbf{a}}(\boldsymbol{x}) = (\mathbf{a}'\boldsymbol{x} - \mathbf{a}'\widehat{\boldsymbol{\mu}})(\mathbf{a}'\widehat{\boldsymbol{\mu}}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}_2), \tag{2.3}$$

the misclassification rate of $\hat{\delta}_{\mathbf{a}}$ *satisfies*

$$P(\hat{\delta}_{\mathbf{a}}(\boldsymbol{X}) \leq 0 | \boldsymbol{Y}_{ki},\ i = 1, \cdots, n_k,\ k = 1, 2) \xrightarrow{P} \frac{1}{2},$$

where the probability is taken with respect to $\mathbf{a}$ *and* $\boldsymbol{X} \in \mathcal{C}_1$*.*

3 Feature Selection by Two-Sample t-Test

To extract salient features, we appeal to the two sample t-test statistics. Other componentwise tests such as the rank sum test can also be used, but we do not pursue those in detail. The two-sample t-statistic for feature j is defined as

$$T_j = \frac{\bar{Y}_{1j} - \bar{Y}_{2j}}{\sqrt{S_{1j}^2/n_1 + S_{2j}^2/n_2}},\ j = 1, \cdots, p, \tag{3.1}$$

where $\bar{Y}_{kj}$ and S_{kj}^2 are the same as those defined in Section 1. We work under more relaxed technical conditions: the normality assumption is not needed. Instead, we assume merely that the noise vectors $\boldsymbol{\epsilon}_{ki},\ i = 1, \cdots, n_k$ are i.i.d. within class $\mathcal{C}_k$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_k$, and are independent between classes. The covariance matrix $\boldsymbol{\Sigma}_1$ can also differ from $\boldsymbol{\Sigma}_2$.

To show that the t-statistic can select all the important features with probability 1, we need the following condition.

Condition 1:

(a) Assume that the vector $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse and without loss of generality, only the first s entries are nonzero.

(b) Suppose that ϵ_{kij} and $\epsilon_{kij}^2 - 1$ satisfy the Cramér's condition, i.e., there exist constants ν_1, ν_2, M_1 and M_2, such that $E|\epsilon_{kij}|^m \leq m! M_1^{m-2}\nu_1/2$ and $E|\epsilon_{kij}^2 - \sigma_{kj}^2|^m \leq m! M_2^{m-2}\nu_2/2$ for all $m = 1, 2, \cdots$.

(c) Assume that the diagonal elements of both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are bounded away from 0.

The following theorem describes the situation under which the two sample t-test can pick up all important features by choosing an appropriate critical value. Recall that $c_1 \leq n_1/n_2 \leq c_2$ and $n = n_1 + n_2$.

Theorem 3 *Let s be a sequence such that $\log(p-s) = o(n^\gamma)$ and $\log s = o(n^{\frac{1}{2}-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < \frac{1}{3}$. Suppose that $\min_{1\leq j\leq s} \frac{|\alpha_j|}{\sqrt{\sigma_{1j}^2+\sigma_{2j}^2}} = n^{-\gamma}\beta_n$. Then under Condition 1, for $x \sim cn^{\gamma/2}$ with c some positive constant, we have*

$$P(\min_{j\leq s} |T_j| \geq x \text{ and } \max_{j>s} |T_j| < x) \to 1.$$

In the proof of Theorem 3, we used the moderate deviation results of the two-sample t-statistic (see Cao, 2007 or Shao, 2005). Theorem 3 allows the lowest signal level to decay with sample size n. As long as the rate of decay is not too fast and the sample size is not too small, the two sample t-test can pick up all the important features with probability tending to 1.

4 Features Annealed Independence Rules

We apply the independence classifier to the selected features, resulting in a Features Annealed Independence Rule (FAIR). In many applications such as tumor classification using gene expression data, we would expect that elements in the population mean difference vector $\boldsymbol{\alpha}$ are sparse: most entries are small. Thus, even if we could use t-test to correctly extract out all these features, the resulting choice is not necessarily optimal, since the noise accumulation can even exceed the signal accumulation for faint features. This can be seen from Theorem 1. Therefore, it is necessary to further single out the most important features that help reduce misclassification rate.

To help us select the number of features, or the critical value of the test statistic, we first consider the ideal situation that the important features are located at the first m coordinates and our task is to merely select m to minimize the misclassification rate. This is the case when we have the ideal information about the relative importance of features, as measured by $|\alpha_j|/\sigma_j$, say. When such an oracle information is unavailable, we will learn it from the data. In the situation that we have vague knowledge about the importance of features such as tumor classification using gene expression data, we can give high ranks to features with large $|\alpha_j|/\sigma_j$.

In the presentation below, unless otherwise specified, we assume that the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ are both from Gaussian distributions and the common covariance matrix

is the identity, i.e., $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$. If this common covariance matrix is known, the independence classifier $\hat{\delta}$ becomes the nearest centroids classifier

$$\hat{\delta}_{\mathrm{NC}}(\boldsymbol{x}) = (\boldsymbol{x} - \widehat{\boldsymbol{\mu}})'(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2).$$

If only the first m dimensions are used in the classification, the corresponding features annealed independence classifier becomes

$$\hat{\delta}^m_{\mathrm{NC}}(\boldsymbol{x}) = (\boldsymbol{x}^m - \widehat{\boldsymbol{\mu}}^m)'(\widehat{\boldsymbol{\mu}}^m_1 - \widehat{\boldsymbol{\mu}}^m_2),$$

where the superscript m means that the vector is truncated after the first m entries. This is indeed the same as the nearest shrunken centroids method of Tibshirani *et al.*(2002).

Theorem 4 *Consider the truncated classifier $\hat{\delta}^{m_n}_{NC}$ for a given sequence m_n. Suppose that $\frac{n}{\sqrt{m_n}} \sum_{j=1}^{m_n} \alpha_j^2 \to \infty$ as $m_n \to \infty$. Then the classification error of $\hat{\delta}^{m_n}_{NC}$ is*

$$W(\hat{\delta}^{m_n}_{NC}, \boldsymbol{\theta}) = 1 - \Phi\Big(\frac{(1+o_P(1))\sum_{j=1}^{m_n}\alpha_j^2 + m_n(n_1-n_2)/(n_1n_2)}{2\{(1+o_P(1))\sum_{j=1}^{m_n}\alpha_j^2 + nm_n/n_1n_2\}^{1/2}}\Big),$$

where $n = n_1 + n_2$ as defined in Section 2.

In the following, we suppress the dependence of m on n when there is no confusion. The above theorem reveals that the ideal choice on the number of features is

$$m_0 = \mathrm{argmax}_{1\le m\le p} \frac{[\sum_{j=1}^m \alpha_j^2 + m(n_1-n_2)/(n_1n_2)]^2}{nm/(n_1n_2) + \sum_{j=1}^m \alpha_j^2}.$$

It can be estimated as

$$\hat{m}_0 = \mathrm{argmax}_{1\le m\le p} \frac{[\sum_{j=1}^m \hat{\alpha}_j^2 + m(n_1-n_2)/(n_1n_2)]^2}{nm/(n_1n_2) + \sum_{j=1}^m \hat{\alpha}_j^2},$$

where $\hat{\alpha}_j = \hat{\mu}_{1j} - \hat{\mu}_{2j}$. The expression for m_0 quantifies how the signal and the noise affect the misclassification rates as the dimensionality m increases. In particular, when $n_1 =$

n_2, the express reduces to $m_0 = \text{argmax}_{1\le m\le p} \frac{[m^{-1/2}\sum_{j=1}^m \alpha_j^2]^2}{2/n+\sum_{j=1}^m \alpha_j^2/m}$. The term $m^{-1/2}\sum_{j=1}^m \alpha_j^2$ reflects the trade-off between the signal and noise as dimensionality m increases.

The good performance of the classifier $\hat{\delta}_{\text{NC}}^m$ depends on the assumption that the largest entries of $\boldsymbol{\alpha}$ cluster at the first m dimensions. An ideal version of the classifier $\hat{\delta}_{\text{NC}}$ is to select a subset $\mathcal{A} = \{j : |\alpha_j| > a\}$ and use this subset to construct independence classifier. Let m be the number of elements in $\mathcal{A}$. The oracle classifier can be written as

$$\hat{\delta}_{\text{orc}}(\boldsymbol{x}) = \sum_{j=1}^p \hat{\alpha}_j(x_j - \hat{\mu}_j)1_{\{|\alpha_j|>a\}}.$$

The misclassification rate is approximately

$$1 - \Phi\Big(\frac{\sum_{j\in\mathcal{A}} \alpha_j^2 + m(n_1-n_2)/(n_1n_2)}{2\{nm/(n_1n_2) + \sum_{j\in\mathcal{A}} \alpha_j^2\}^{1/2}}\Big), \tag{4.1}$$

when $\frac{n}{\sqrt{m}}\sum_{j\in\mathcal{A}} \alpha_j^2 \to \infty$ and $m \to \infty$. This is straightforward from Theorem 4. In practice, we do not have such an oracle, and selecting the subset $\mathcal{A}$ is difficult. A simple procedure is to use the feature annealed independence rule based on the hard thresholding:

$$\hat{\delta}_{\text{FAIR}}^b(\boldsymbol{x}) = \sum_{j=1}^p \hat{\alpha}_j(x_j - \hat{\mu}_j)1_{\{|\hat{\alpha}_j|>b\}}.$$

We study the classification error of FAIR and the impact of the threshold b on the classification result in the following theorem.

Theorem 5 *Suppose that* $\max_{j\in\mathcal{A}^c}|\alpha_j| < b_n$ *and* $\log(p-m)/[n(b_n - \max_{j\in\mathcal{A}^c}|\alpha_j|)^2] \to 0$ *with* $m = |\mathcal{A}|$. *Moreover, assume that* $\frac{n}{\sqrt{m}}\sum_{j\in\mathcal{A}} \alpha_j^2 \to \infty$ *and* $\sum_{j\in\mathcal{A}}|\alpha_j|/[\sqrt{n}\sum_{j\in\mathcal{A}} \alpha_j^2] \to 0$. *Then*

$$W(\hat{\delta}_{FAIR}^{b_n}, \boldsymbol{\theta}) \le 1 - \Phi\Big(\frac{(1+o_P(1))\sum_{j\in\mathcal{A}} \alpha_j^2 + nm(n_1n_2)^{-1} - mb_n^2}{2\{(1+o_P(1))\sum_{j\in\mathcal{A}} \alpha_j^2 + nm(n_1n_2)^{-1}\}^{1/2}}\Big).$$

Notice that the upper bound of $W(\hat{\delta}_{\text{FAIR}}^{b_n}, \boldsymbol{\theta})$ in Theorem 5 is greater than the classification error in Theorem 4, and the magnitude of difference depends on mb_n^2. This

is expected as estimating the set $\mathcal{A}$ increases the classification error. These results are similar to those in Fan (1996) for high-dimensional hypothesis testing.

When the common covariance matrix is different from the identity, FAIR takes a slightly different form to adapt to the unknown componentwise variance:

$$\hat{\delta}_{\mathrm{FAIR}}(\boldsymbol{x}) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j)/\hat{\sigma}_j^2 1_{\{\sqrt{n/(n_1 n_2)}|T_j|>b\}}, \tag{4.2}$$

where T_j is the two sample t-statistic. It is clear from (4.2) that FAIR works the same way as that we first sort the features by the absolute values of their t-statistics in the descending order, and then take out the first m features to classify the data. The number of features can be selected by minimizing the upper bound of the classification error given in Theorem 1. The optimal m in this sense is

$$m_1 = \mathrm{argmax}_{1\le m\le p} \frac{1}{\lambda_{\max}^m} \frac{[\sum_{j=1}^m \alpha_j^2/\sigma_j^2 + m(1/n_2 - 1/n_1)]^2}{nm/(n_1 n_2) + \sum_{j=1}^m \alpha_j^2/\sigma_j^2},$$

where $\lambda_{\max}^m$ is the largest eigenvalue of the correlation matrix $\mathbf{R}^m$ of the truncated observations. It can be estimated from the samples:

$$\begin{aligned} \hat{m}_1 &= \mathrm{argmax}_{1\le m\le p} \frac{1}{\hat{\lambda}_{\max}^m} \frac{[\sum_{j=1}^m \hat{\alpha}_j^2/\hat{\sigma}_j^2 + m(1/n_2 - 1/n_1)]^2}{nm/(n_1 n_2) + \sum_{j=1}^m \hat{\alpha}_j^2/\hat{\sigma}_j^2} \\ &= \mathrm{argmax}_{1\le m\le p} \frac{1}{\hat{\lambda}_{\max}^m} \frac{n[\sum_{j=1}^m T_j^2 + m(n_1 - n_2)/n]^2}{mn_1 n_2 + n_1 n_2 \sum_{j=1}^m T_j^2}. \end{aligned} \tag{4.3}$$

Note that the factor $\lambda_{\max}^m$ in (4.3) increases with m, which makes $\hat{m}_1$ usually smaller than $\hat{m}_0$.

5 Numerical Studies

In this section we use a simulation study and three real data analyses to illustrate our theoretical results and to verify the performance of our newly proposed classifier FAIR.

5.1 Simulation Study

We first introduce the model. The covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ for the two classes are chosen to be the same. For the distribution of the error ϵ_{ij} in (2.1), we use the same model as that in Fan, Hall and Yao (2006). Specifically, features are divided into three groups. Within each group, features share one unobservable common factor with different factor loadings. In addition, there is an unobservable common factor among all the features across three groups. For simplicity, we assume that the number of features p is a multiple of 3. Let Z_{ij} be a sequence of independent $N(0,1)$ random variables, and χ^2_{ij} be a sequence of independent random variables of the same distribution as $(\chi^2_d - d)/\sqrt{2d}$ with χ^2_d the Chi-square distribution with degrees of freedom d. In the simulation we set $d = 6$.

Let $\{a_j\}$ and $\{b_j\}$ be factor loading coefficients. Then the error in (2.1) is defined as

$$\epsilon_{ij} = \frac{Z_{ij} + a_{1j}\chi_{1i} + a_{2j}\chi_{2i} + a_{3j}\chi_{3i} + b_j\chi_{4i}}{(1 + a_{1j}^2 + a_{2j}^2 + a_{3j}^2 + b_j^2)^{1/2}},\ i = 1, \cdots, n_k,\ j = 1, \cdots, p,$$

where $a_{ij} = 0$ except that $a_{1j} = a_j$ for $j = 1, \cdots, p/3$, $a_{2j} = a_j$ for $j = (p/3) + 1, \cdots, 2p/3$, and $a_{3j} = a_j$ for $j = (2p/3)+1, \cdots, p$. Therefore, $E\epsilon_{ij} = 0$ and $\mathrm{var}(\epsilon_{ij}) = 1$, and in general, within group correlation is greater than the between group correlation. The factor loadings a_j and b_j are independently generated from uniform distributions $U(0, 0.4)$ and $U(0, 0.2)$. The mean vector $\boldsymbol{\mu}_1$ for class $\mathcal{C}_1$ is taken from a realization of the mixture of a point mass at 0 and a double-exponential distribution:

$$(1 - c)\delta_0 + \frac{1}{2}c\exp(-2|x|),$$

where $c \in (0, 1)$ is a constant. In the simulation, we set $p = 4,500$ and $c = 0.02$. In other words, there are around 90 signal features on an average, many of which are weak signals. Without loss of generality, $\boldsymbol{\mu}_2$ is set to be 0. Figure 1 shows the true mean

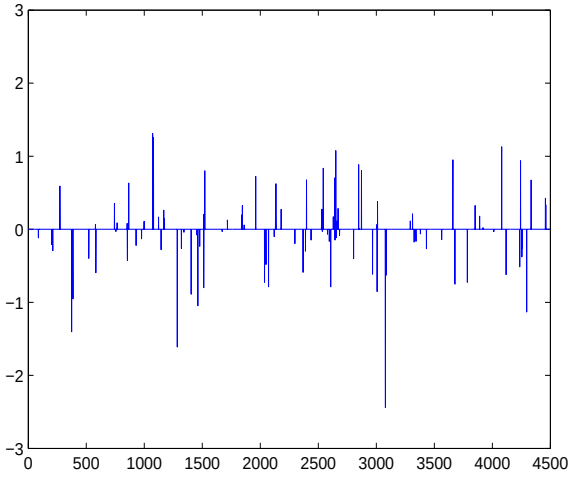

Figure 1: True mean difference vector $\boldsymbol{\alpha}$. x-axis represents the dimensionality, and y-axis shows the values of corresponding entries of $\boldsymbol{\alpha}$.

difference vector $\boldsymbol{\alpha}$, which is fixed across all simulations. It is clear that there are only very few features with signal levels exceeding 1 standard deviation of the noise.

With the parameters and model above, for each simulation, we generate $n_1 = 30$ training data from class $\mathcal{C}_1$ and $n_2 = 30$ training data from $\mathcal{C}_2$. In addition, separate 200 samples are generated from each of the two classes in each simulation, and these 400 vectors are used as test samples. We apply our newly proposed classifier FAIR to the simulated data. Specifically, for each feature, the t-test statistic in (3.1) is calculated using the training sample. Then the features are sorted in the decreasing order of the absolute values of their t-statistics. We then examine the impact of the number of features m on the misclassification rate. In each simulation, with m ranging from 1 to 4500, we construct the feature annealed independence classifiers using the training samples, and then apply these classifiers to the 400 test samples. The classification errors are compared to those of the independence rule with the oracle ordering information,

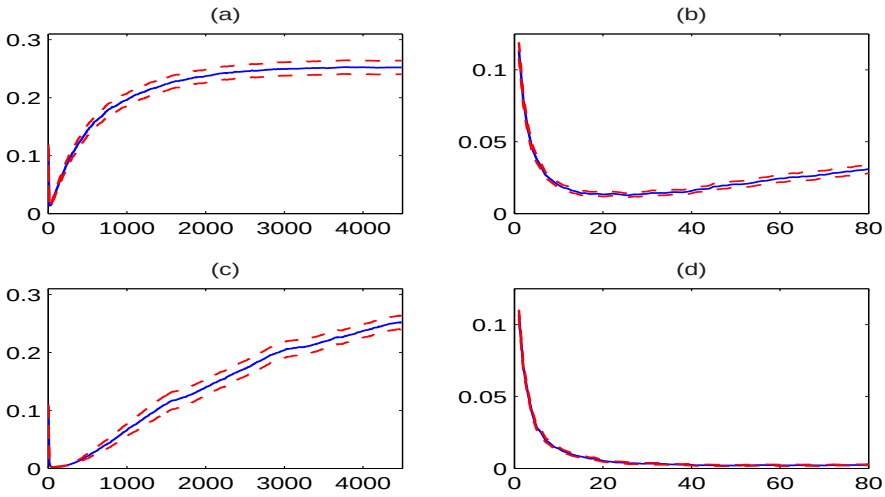

Figure 2: Number of features versus misclassification rates. The solid curves represent the averages of classification errors across 100 simulations. The dashed curves are 2 standard errors away from the solid curves. The x-axis represents the number of features used in the classification, and the y-axis shows the misclassification rates. (a) The features are ordered in a way such that the corresponding t-statistics are decreasing in absolute values. (b) The amplified plot of the first 80 values of x-axis in plot (a). (c) The same as (a) except that the features are arranged in a way such that the corresponding true mean differences are decreasing in absolute values. (d) The amplified plot of the first 80 values of x-axis in plot (c).

which is constructed by repeating the above procedure except that in the first step the features are ordered by their true signal levels, $|\boldsymbol{\alpha}|$, instead of by their t-statistics.

The above procedure is repeated 100 times, and averages and standard errors of the misclassification rates (based on 400 test samples in each simulation) are calculated across the 100 simulations. Note that the average of the 100 misclassification rates is indeed computed based on 100×400 testing samples.

Figure 2 depicts the misclassification rate as a function of the number of features m. The solid curves represent the average of classification rates across the 100 simulations, and the corresponding dashed curves are 2 standard errors (i.e. the standard deviation of 100 misclassification rates divided by 10) away from the solid one. The misclassification

rates using the first 80 features in Figure 2(a) are zoomed in Figure 2(b). Figures 2(c) and 2(d) are the same as 2(a) and 2(b) except that the features are arranged in the decreasing order of $|\boldsymbol{\alpha}|$, i.e., the results are based on the oracle-assisted feature annealed independence classifier. From these plots we see that the classification results of FAIR are close to those of the oracle-assisted independence classifier. Moreover, as the dimensionality m grows, the misclassification rate increases steadily due to the noise accumulation. When all the features are included, i.e. $m = 4500$, the misclassification rate is 0.2522, whereas the minimum classification errors are 0.0128 in plot 2(b) and 0.0020 in plot 2(d). These results are consistent with Theorem 1. We also tried to decrease the signal levels, i.e., the mean of the double exponential distribution, or to increase the dimensionality p, and found that the classification error tend to 0.5 when all the dimensions are included. Comparing Figures 2(a) and 2(b) to Figures 2(c) and 2(d), we see that the features ordered by t-statistics has higher misclassification rates than those ordered by the oracle. Also, using t-statistics results in larger minimum classification errors (see plots 2(b) and 2(d)), but the differences are not very large.

Figure 3 shows the classification errors of the independence rule based on projected samples onto randomly chosen directions across 100 simulations. Specifically, in each of the simulations in Figure 2, we generate a direction vector $\mathbf{a}$ randomly from the $(p-1)$-dimensional unit sphere, then project all the data in that simulation onto the direction $\mathbf{a}$, and finally apply the Fisher discriminant to the projected data (see (2.3)). The average of these misclassification rates is 0.4986 and the corresponding standard deviation is 0.0318. These results are consistent with our Theorem 2.

Finally, we examine the effectiveness of our proposed method (4.3) for selecting features in FAIR. In each of the 100 simulations, we apply (4.3) to choose the number of features and compute the resulting misclassification rate based on 400 test samples. We

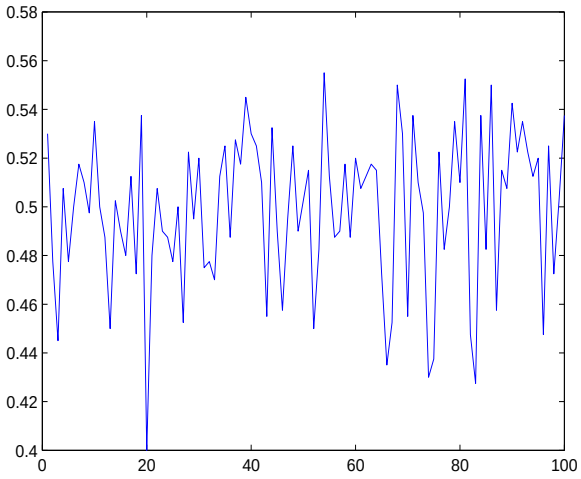

Figure 3: Classification errors of the independence rule based on projected samples onto randomly chosen directions over 100 simulations.

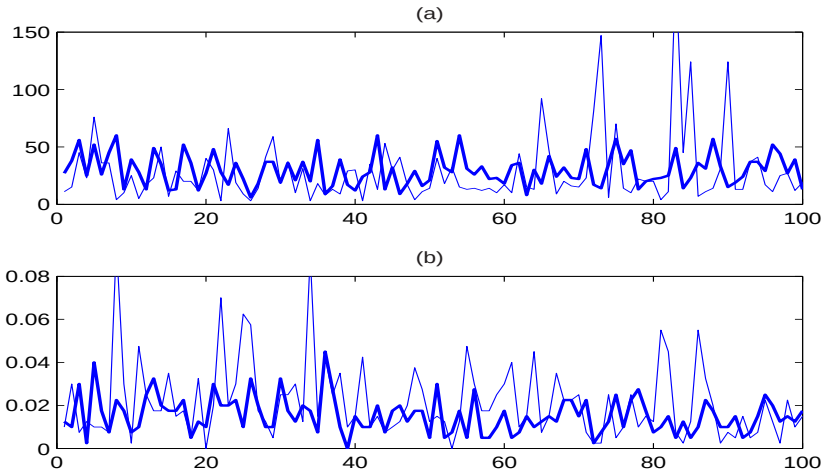

Figure 4: The thick curves correspond to FAIR, while the thin curves correspond to the nearest shrunken centroids method. (a) The numbers of features chosen by (4.3) and by the nearest shrunken centroids method over 100 simulations. (b) Corresponding classification errors based on the optimal number of features chosen in (a) over 100 simulations.

also use the nearest shrunken centroids of Tibshirani *et al.* (2003) to select the important features. Figure 4 summarizes these results. The thin curves correspond to the nearest shrunken centroids method, and the thick curves correspond to FAIR. Figure 4(a) presents the number of features calculated from these two methods, and Figure 4(b) shows the corresponding misclassification rates. For our newly proposed classifier FAIR, the average of the optimal number of features over 100 simulations is 29.71, which is very close to the smallest number of features with the minimum misclassification rate in Figure 2(d). The misclassification rates of FAIR in Figure 4(b) have average 0.0154 and standard deviation 0.0085, indicating the outstanding performance of FAIR. Nearest shrunken centroids method is unstable in selecting features. Over the 100 simulations, there are several realizations in which it chooses plenty of features. We truncated Figure 4 to make it easier to view. The average number of features chosen by the nearest shrunken centroids is 28.43, and the average classification error is 0.0216, with corresponding standard deviation 0.0179. It is clear that nearest shrunken centroids method tends to choose less features than FAIR, but the misclassification rates are larger.

5.2 Real Data Analysis

5.2.1 Leukemia Data

Leukemia data from high-density Affymetrix oligonucleotide arrays were previously analyzed in Golub *et al.*(1999), and are available at `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`. There are 7129 genes and 72 samples coming from two classes: 47 in class ALL (acute lymphocytic leukemia) and 25 in class AML (acute mylogenous leukemia). Among these 72 samples, 38 (27 in class ALL and 11 in class AML) are set to be training samples and 34 (20 in class ALL and 14 in class AML) are set as test samples.

Before classification, we standardize each sample to zero mean and unit variance as done by Dudoit *et al.*(2002). The classification results from the nearest shrunken centroids (NSC hereafter) method and FAIR are shown in Table 1. The nearest shrunken centroids method picks up 21 genes and makes 1 training error and 3 test errors, while our method chooses 11 genes and makes 1 training error and 1 test error. Tibshirani *et al.*(2002) proposed and applied the nearest shrunken centroids method to the unstandardized Leukemia dataset. They chose 21 genes and made 1 training error and 2 test errors. Our results are still superior to theirs.

To further evaluate the performance of the two classifiers, we randomly split the 72 samples into training and test sets. Specifically, we set approximately $100\gamma\%$ of the observations from class ALL and $100\gamma\%$ of the observations from class AML as training samples, and the rest as test samples. FAIR and NSC are applied to the training data, and their performances are evaluated by the test samples. The above procedure is repeated 100 times for $\gamma = 0.4,\ 0.5$ and 0.6, respectively, and the distributions of test errors of FAIR, NSC and the independence rule without feature selection are summarized in Figure 5. In each of the splits, we also calculated the difference of test errors between NSC and FAIR, i.e., the test error of FAIR minus that of NSC, and the distribution is summarized in Figure 5. The top panel of Figure 6 shows the number of features selected by FAIR and NSC for $\gamma = 0.4$. The results for the other two values of γ are similar so we do not present here to save the space. From these figures we can see that the performance of independence rule improves significantly after feature selection. The classification errors of NSC and FAIR are approximately the same. As we have already noticed in the simulation study, NSC is not good with feature selection, that is, the number of features selected by NSC is very large and unstable, while the number of features selected by FAIR is quite reasonable and stable over different random splits.

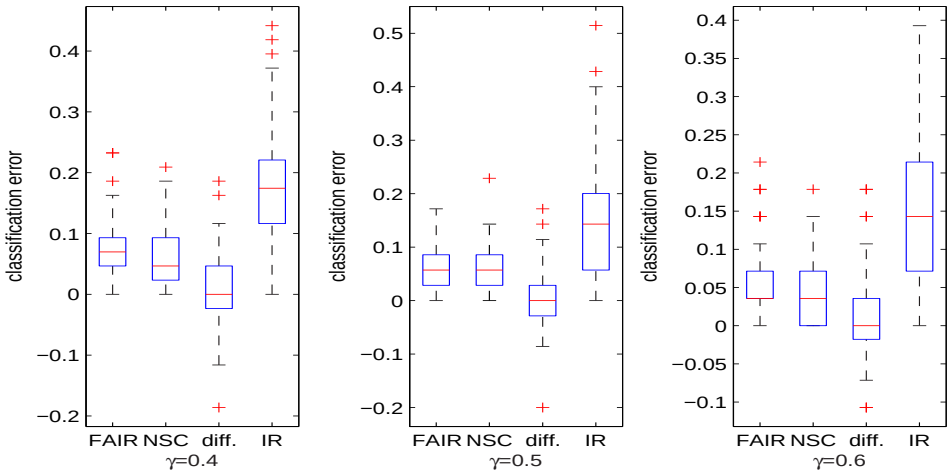

Figure 5: Leukemia data. Boxplots of test errors of FAIR, NSC and the independence rule without feature selection over 100 random splits of 72 samples, where $100\gamma\%$ of the samples from both classes are set as training samples. The three plots from left to right correspond to $\gamma = 0.4, 0.5$ and 0.6, respectively. In each boxplot above, "FAIR" refers to the test errors of the feature annealed independent rule; "NSC" corresponds to the test errors of nearest shrunken centroids method; "diff." means the difference of the test errors of FAIR and those of NSC; and "IR" corresponds the test errors of independence rule without feature selection.

Clearly, the independent rule without feature selection performs poorly.

Table 1: Classification errors of Leukemia data set

Method	Training error	Test error	No. of selected genes
Nearest shrunken centroids	1/38	3/34	21
FAIR	1/38	1/34	11

5.2.2 Lung Cancer Data

We evaluate our method by classifying between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. Lung cancer data were analyzed by Gordon

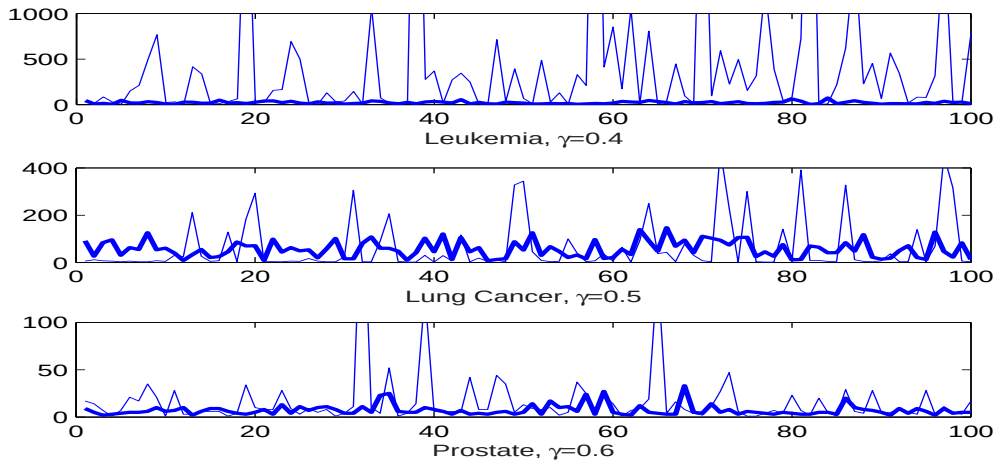

Figure 6: Leukemia, Lung Cancer, and Prostate data sets. The number of features selected by FAIR and NSC over 100 random splits of the total samples. In each split, $100\gamma\%$ of the samples from both class are set as training samples, and the rest are used as test samples. The three plots from top to bottom correspond to the Leukemia data with $\gamma = 0.4$, the Lung Cancer data with $\gamma = 0.5$ and the Prostate cancer data with $\gamma = 0.6$, respectively. The thin curves show the results from NSC, and the thick curves correspond to FAIR. The plots are truncated to make them easy to view.

et al.(2002) and are available at `http://www.chestsurg.org`. There are 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 of them, with 16 from MPM and 16 from ADCA. The rest 149 samples are used for testing (15 from MPM and 134 from ADCA). Each sample is described by 12533 genes.

As in the Leukemia data set, we first standardize the data to zero mean and unit variance, and then apply the two classification methods to the standardized data set. Classification results are summarized in Table 2. Although FAIR uses 5 more genes than the nearest shrunken centroids method, it has better classification results: both methods perfectly classify the training samples, while our classification procedure has smaller test error.

We follow the same procedure as that in Leukemia example to randomly split the 181 samples into training and test sets. FAIR and NSC are applied to the training data, and the test errors are calculated using the test data. The procedure is repeated 100 times with $\gamma = 0.4, 0.5$ and 0.6, respectively, and the test error distributions of FAIR, NSC and the independence rule without feature selection can be found in Figure 7. We also present the difference of the test errors between FAIR and NSC in Figure 7. The numbers of features used by FAIR and NSC with $\gamma = 0.5$ are shown in the middle panel of Figure 6. Figure 7 shows again that feature selection is very important in high dimensional classification. The performance of FAIR is close to NSC in terms of classification error (Figure 7), but FAIR is stable in feature selection, as shown in the middle panel of Figure 6. One possible reason of Figure 7 might be that the signal strength in this Lung Cancer dataset is relatively weak, and more features are needed to obtain the optimal performance. However, the estimate of the largest eigenvalue is not accurate anymore when the number of features is large, which results in inaccurate estimates of m_1 in (4.3).

Table 2: Classification errors of Lung Cancer data

Method	Training error	Test error	No. of selected genes
Nearest shrunken centroids	0/32	11/149	26
FAIR	0/32	7/149	31

5.2.3 Prostate Cancer Data

The last example uses the prostate cancer data studied in Singh *et al.*(2002). The data set is available at `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`. The training data set contains 102 patient samples, 52 of which (labeled as "tumor") are

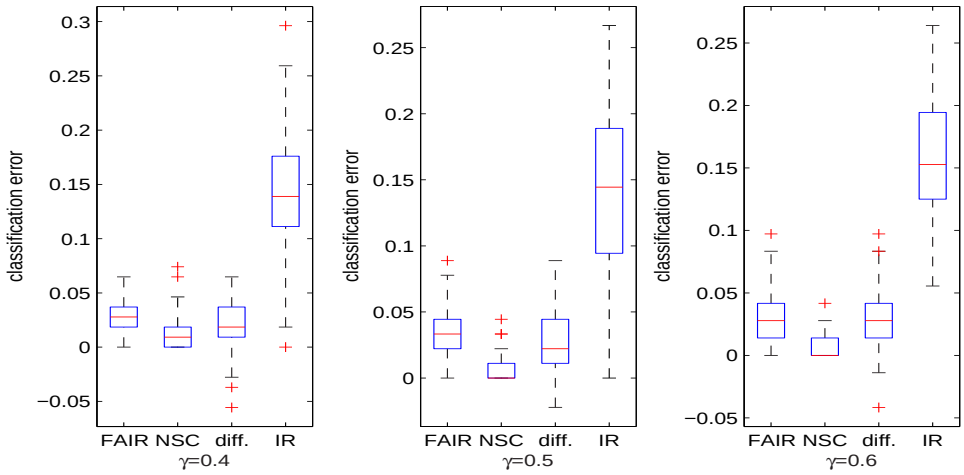

Figure 7: Lung cancer data. The same as Figure 5 except that the data set is different.

prostate tumor samples and 50 of which (labeled as "Normal") are prostate samples. There are around 12600 genes. An independent set of test samples is from a different experiment and has 25 tumor and 9 normal samples.

We preprocess the data by standardizing the gene expression data as before. The classification results are summarized in Table 3. We make the same test error as and a bit larger training error than the nearest shrunken centroids method, but the number of selected genes we use is much less.

The samples are randomly split into training and test sets in the same way as before, the test errors are calculated, and the number of features used by these two methods are recorded. Figure 8 shows the test errors of FAIR, NSC and the independence rule without feature selection, and the difference of the test errors of FAIR and NSC. The bottom panel of Figure 6 presents the numbers of features used by FAIR and NSC in each random split for $\gamma = 0.6$. As we mentioned before, the plots for $\gamma = 0.4$ and 0.5 are

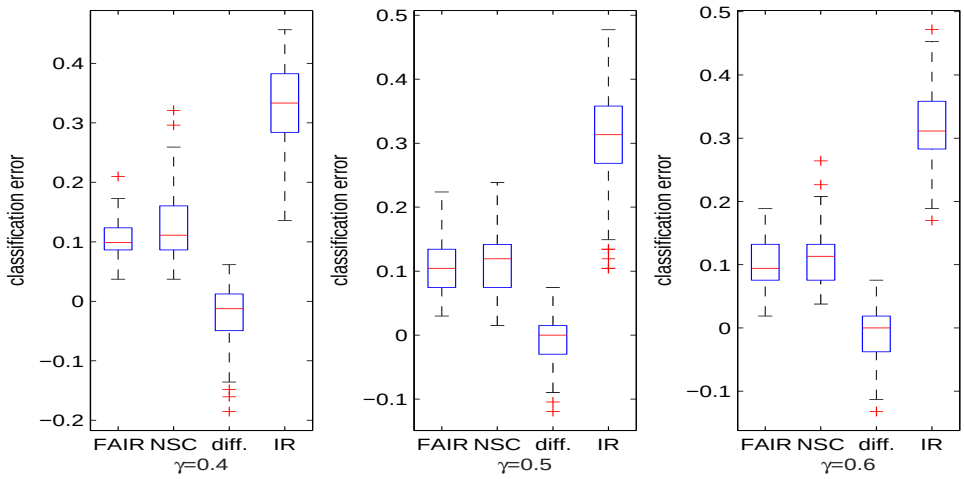

Figure 8: Prostate cancer data. The same as Figure 5 except that the data set is different.

similar so we omit them in the paper. The performance of FAIR is better than that of NSC both in terms of classification error and in terms of the selection of features. The good performance of FAIR might be caused by the strong signal level of few features in this data set. Due to the strong signal level, FAIR can attain the optimal performance with small number of features. Thus, the estimate of m_1 in (4.3) is accurate and hence the actual performance of FAIR is good.

Table 3: Classification errors of Prostate Cancer data set

Method	Training error	Test error	No. of selected genes
Nearest shrunken centroids	8/102	9/34	6
FAIR	10/102	9/34	2

6 Conclusion

This paper studies the impact of high dimensionality on classifications. To illustrate the idea, we have considered the independence classification rule, which avoids the difficulty of estimating large covariance matrix and the diverging condition number frequently associated with the large covariance matrix. When only a subset of the features capture the characteristics of two groups, classification using all dimensions would intrinsically classify the noises. We prove that classification based on linear projections onto almost all directions performs nearly the same as random guessing. Hence, it is necessary to choose direction vectors which put more weights on important features.

The two-sample t-test can be used to choose the important features. We have shown that under mild conditions, the two sample t-test can select all the important features with probability one. The features annealed independence rule using hard thresholding, FAIR, is proposed, with the number of features selected by a data-driven rule. An upper bound of the classification error of FAIR is explicitly given. We also give suggestions on the optimal number of features used in classification. Simulation studies and real data analysis support our theoretical results convincingly.

7 Appendix

Proof of Theorem 1. For $\boldsymbol{\theta} \in \Gamma$, Ψ defined in (2.2) can be bounded as

$$\Psi \geq \frac{(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})' \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{\sqrt{b_0 (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)' \widehat{\mathbf{D}}^{-1} \mathbf{D} \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}}, \tag{7.1}$$

where we have used the assumption that $\lambda_{\max}(\mathbf{R}) \leq b_0$. Denote by

$$\widetilde{\Psi} = \frac{(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})' \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{\sqrt{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)' \widehat{\mathbf{D}}^{-1} \mathbf{D} \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}}.$$

We next study the asymptotic behavior of $\widetilde{\Psi}$.

Since Condition 1(b) in Section 3 is satisfied automatically for normal distribution, by Lemma 2 below we have $\widehat{\mathbf{D}} = \mathbf{D}(1 + o_P(1))$, where $o_P(1)$ holds uniformly across all diagonal elements. Thus, the right hand side of (7.1) can be written as

$$\frac{1}{\sqrt{b_0}} \frac{(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})' \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{\sqrt{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)' \mathbf{D}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}} (1 + o_P(1)).$$

We first consider the denominator. Notice that it can be decomposed as

$$\begin{aligned} (7.2) \qquad & (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)' \mathbf{D}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) \\ =& \boldsymbol{\alpha}' \mathbf{D}^{-1} \boldsymbol{\alpha} + 2 \sum \alpha_j \frac{\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j}}{\sigma_j^2} + \sum \frac{(\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j})^2}{\sigma_j^2} \\ =& \boldsymbol{\alpha}' \mathbf{D}^{-1} \boldsymbol{\alpha} + 2 \sum \frac{\alpha_j}{\sigma_j^2} (\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j}) + \sum \frac{(\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j})^2}{\sigma_j^2} \\ \equiv& \boldsymbol{\alpha}' \mathbf{D}^{-1} \boldsymbol{\alpha} + 2(1 + o_P(1)) I_1 + I_2, \end{aligned}$$

where σ_j^2 is the j-th diagonal entry of $\mathbf{D}$, $\hat{\sigma}_j^2$ is the j-th diagonal entry of $\widehat{\mathbf{D}}$, and $\hat{\epsilon}_{kj} = \sum_{i=1}^{n_k} \epsilon_{kij}/n_k,\ k = 1, 2$. Notice that $\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2 \sim N(0, \frac{n}{n_1 n_2} \boldsymbol{\Sigma})$. By singular value decomposition we have

$$\mathbf{R} = \mathbf{Q}_R \mathbf{V}_R \mathbf{Q}_R',$$

where $\mathbf{Q}_R$ is orthogonal matrix and $\mathbf{V}_R = \text{diag}\{\lambda_{R,1}, \cdots, \lambda_{R,p}\}$ be the eigenvalues of the correlation matrix $\mathbf{R}$. Define $\tilde{\boldsymbol{\epsilon}} = \sqrt{n_1 n_2/n} \mathbf{V}_R^{-1/2} \mathbf{Q}_R' \mathbf{D}^{-1/2} (\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)$, then $\tilde{\boldsymbol{\epsilon}} \sim N(0, \mathbf{I})$. Hence,

$$I_2 = (\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)' \mathbf{D}^{-1} (\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2) = \frac{n}{n_1 n_2} \tilde{\boldsymbol{\epsilon}}' \mathbf{V}_R \tilde{\boldsymbol{\epsilon}}.$$

Since $\sum_{i=1}^p \lambda_{R,i} = p$ and $\lambda_{R,i} \geq 0$ for all $i = 1, \cdots, p$, we have $\frac{1}{p^2} \sum_{i=1}^p \lambda_{R,i}^2 < \infty$. By weak law of large number we have

$$(7.3) \qquad n_1 n_2 I_2 / [pn] \xrightarrow{\text{P}} 1 \quad \text{as } n \to \infty, p \to \infty.$$

Next, we consider I_1. Note that I_1 has the distribution $I_1 \sim N(0, \frac{n}{n_1 n_2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\boldsymbol{\alpha})$. Since $\lambda_{\max} \leq b_0$, $n\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha} \geq nC_p \to \infty$ and

$$\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2}\boldsymbol{\alpha} \leq \lambda_{\max}(R)\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha},$$

we have $I_1 = \boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}o_P(1)$. This together with (7.2) and (7.3) yields

$$\frac{n_1 n_2}{pn}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)'\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) = 1 + \frac{n_1 n_2}{pn}\sum_{j=1}^{p}\frac{\alpha_j^2}{\sigma_j^2}(1 + o_P(1)). \tag{7.4}$$

Now, we consider the numerator. It can be decomposed as

$$\begin{aligned}
&(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})'\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)\\
=&\frac{1}{2}\boldsymbol{\alpha}'\widehat{\mathbf{D}}^{-1}\boldsymbol{\alpha} - \sum\frac{\alpha_j}{\hat{\sigma}_j^2}(\hat{\epsilon}_{2j}) - \frac{1}{2}(1 + o_P(1))\sum\hat{\epsilon}_{1j}^2/\sigma_j^2 + \frac{1}{2}(1 + o_P(1))\sum\hat{\epsilon}_{2j}^2)/\sigma_j^2\\
\equiv&\frac{1}{2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}(1 + o_P(1)) - I_3 - \frac{1}{2}(1 + o_P(1))I_4 + \frac{1}{2}(1 + o_P(1))I_5.
\end{aligned}$$

Denote by $\tilde{I}_3 = \sum\frac{\alpha_j}{\sigma_j^2}(\hat{\epsilon}_{2j})$. Note that

$$\max|\frac{\alpha_j}{\sigma_j^2}(\hat{\epsilon}_{2j}) - \frac{\alpha_j}{\hat{\sigma}_j^2}(\hat{\epsilon}_{2j})| \leq \max|\frac{\sigma_j^2}{\hat{\sigma}_j^2} - 1|\max|\frac{\alpha_j}{\sigma_j^2}\hat{\epsilon}_{2j}| = o_P(1)\max|\frac{\alpha_j}{\sigma_j^2}\hat{\epsilon}_{2j}|. \tag{7.5}$$

Define $F_j = \sqrt{n_2}\frac{\alpha_j}{\sigma_j^2}\hat{\epsilon}_{2j}/\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}$, then $\sigma_{F_j}^2 \equiv \mathrm{var}(F_j) \leq 1$ for all j. For the normal distribution, we have the following tail probability inequality

$$1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}.$$

Since $F_j \sim N(0, \sigma_{F_j}^2)$, by the above inequality we have

$$P(|F_j| \geq x) \leq 2\exp\{-\frac{x^2}{2C}\}.$$

with C some positive constant, for all $x > 0$ and $j = 1, \cdots, p$. By Lemma 2.2.10 of van de Vaart and Wellner (1996, P102), we have

$$(\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha})^{-1}E\max|\frac{\alpha_j}{\sigma_j^2}\hat{\epsilon}_{2j}| = n_2^{-1}E\max_{j\leq p}|F_j| \leq K\sqrt{C\log(p+1)/n_2} \xrightarrow{\mathrm{P}} 0,$$

where K is some universal constant. This together with (7.5) ensures that

$$(\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha})^{-1}\max|\frac{\alpha_j}{\sigma_j^2}(\hat{\epsilon}_{2j})-\frac{\alpha_j}{\hat{\sigma}_j^2}(\hat{\epsilon}_{2j})|=o_P(1)$$

Hence,

$$I_3=\tilde{I}_3+\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}o_P(1). \tag{7.6}$$

Now we only need to consider $\tilde{I}_3$. Note that $\tilde{I}_3=\sum\frac{\alpha_j}{\sigma_j^2}\hat{\epsilon}_{2j}\sim N(0,\frac{1}{n_2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\boldsymbol{\alpha})$. Since the variance term can be bounded as

$$\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\boldsymbol{\alpha}\le\lambda_{\max}(\mathbf{R})\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha},$$

By the assumption that $n\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}\to\infty$ and $\lambda_{\max}(\mathbf{R})$ is bounded, we have $\tilde{I}_3=\frac{1}{2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}o_P(1)$. Combining this with (7.6) leads to

$$I_3=\frac{1}{2}\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}o_P(1)$$

We now examine I_4 and I_5. By the similar proof to (7.3) above we have

$$I_4=p/n_1+o_P\big(\sqrt{np/(n_1n_2)}\big)\text{ and }I_5=p/n_2+o_P\big(\sqrt{np/(n_1n_2)}\big).$$

Thus the numerator can be written as

$$(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}})'\widehat{\mathbf{D}}^{-1}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)=(1+o_P(1))\frac{1}{2}\sum\alpha_j^2/\sigma_j^2-(p/n_1-p/n_2)/2+o_P(\sqrt{np/(n_1n_2)}).$$

and by (7.4)

$$\tilde{\Psi}=\frac{\sqrt{\frac{n_1n_2}{pn}}\sum\alpha_j^2/\sigma_j^2(1+o_P(1))+\sqrt{p/(nn_1n_2)}(n_1-n_2)}{2\Big\{1+\frac{n_1n_2}{pn}\sum\alpha_j^2/\sigma_j^2(1+o_P(1))\Big\}^{1/2}}.$$

Since $\frac{ax}{\sqrt{1+a^2x}}$ is an increasing function of x and $\sum\frac{\alpha_j^2}{\sigma_j^2}\ge C_p$, in view of (7.1) and the definition of the parameter space Γ, we have

$$W(\hat{\delta})=1-\Phi\Big(\frac{[n_1n_2/(pn)]^{1/2}C_p\{1+o_P(1)+\frac{p(n_1-n_2)}{n_1n_2C_p}\}}{2\sqrt{b_0}\{1+n_1n_2/(pn)C_p(1+o_P(1))\}^{1/2}}\Big).$$

If $p/(nC_p) \to 0$, then $W(\hat{\delta}) = 1 - \Phi\Big(\frac{1}{2}[n_1 n_2/(pnb_0)]^{1/2} C_p\{1 + o_P(1)\}\Big)$. Furthermore, if $\big\{\frac{n_1 n_2}{pn}\big\}^{1/2} C_p \to C_0$ with C_0 some constant, then

$$W(\hat{\delta}) \xrightarrow{\text{P}} 1 - \Phi\Big(\frac{C_0}{2\sqrt{b_0}}\Big);$$

This completes the proof. ■

Proof of Theorem 2. Suppose we have a new observation $\boldsymbol{X}$ from class $\mathcal{C}_1$. Then the posterior classification error of using $\hat{\delta}_{\mathbf{a}}(\cdot)$ is

$$\begin{aligned} W(\delta_{\mathbf{a}}, \boldsymbol{\theta}) =& E^{\mathbf{a}}[P(\delta_{\mathbf{a}}(\boldsymbol{X}) < 0 | \boldsymbol{Y}_{ki}, i = 1, \cdots, n_k,\ k = 1, 2, \mathbf{a})] \\ =& 1 - E^{\mathbf{a}}\Phi(\Psi_{\mathbf{a}} \text{sign}(\mathbf{a}'\widehat{\boldsymbol{\mu}}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}_2)), \end{aligned}$$

where $\Psi_{\mathbf{a}} = \frac{\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}}$, $\Phi(\cdot)$ is the standard Gaussian distribution function, and $E^{\mathbf{a}}$ means expectation taken with respect to $\mathbf{a}$. We are going to show that

$$\Psi_{\mathbf{a}} \xrightarrow{\text{P}} 0, \tag{7.7}$$

which together with the continuity of $\Phi(\cdot)$ and the dominated convergence theorem gives

$$\lim_p E^{\mathbf{a}}\Phi(\Psi_{\mathbf{a}} \text{sign}(\mathbf{a}'\widehat{\boldsymbol{\mu}}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}_2)) = 1/2.$$

Therefore, the posterior error $W(\hat{\delta}_{\mathbf{a}}, \boldsymbol{\theta})$ is no better than the random guessing.

Now, let us prove (7.7). Note that the random vector $\mathbf{a}$ can be written as

$$\mathbf{a} = \mathbf{Z}/\|\mathbf{Z}\|,$$

where $\mathbf{Z}$ is a p-dimensional standard Gaussian distributed random vector, independent of all the observations $\boldsymbol{Y}_{ki}$ and $\boldsymbol{X}$. Therefore,

$$\Psi_{\mathbf{a}} = \frac{\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}} = \frac{\mathbf{Z}'\boldsymbol{\alpha}/\sqrt{p} - \sqrt{\frac{n}{n_1 n_2 p}}\mathbf{Z}'[\sqrt{\frac{n_1 n_2}{n}}(\hat{\boldsymbol{\epsilon}}_1 + \hat{\boldsymbol{\epsilon}}_2)]}{2\sqrt{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}/p}}, \tag{7.8}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\hat{\boldsymbol{\epsilon}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{\epsilon}_{ki}$, $k = 1, 2$. By the singular value decomposition we have

$$\boldsymbol{\Sigma} = \mathbf{Q}'\mathbf{V}\mathbf{Q},$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{V} = \text{diag}\{\lambda_1, \cdots, \lambda_p\}$ is a diagonal matrix. Let $\tilde{\mathbf{Z}} = \mathbf{Q}\mathbf{Z}$, then $\tilde{\mathbf{Z}}$ is also a p-dimensional standard Gaussian random vector. Hence the denominator of $\Psi_{\mathbf{a}}$ can be written as

$$2\sqrt{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}/p} = 2\big(\frac{1}{p}\sum_{j=1}^{p}\lambda_j\tilde{Z}_j^2\big)^{1/2},$$

where $\tilde{Z}_j$ is the j-th entry of $\tilde{\mathbf{Z}}$. Since it is assumed that $\lim_p \frac{1}{p^2}\sum_{j=1}^{p}\lambda_j^2 < \infty$ and $\lim_p \frac{1}{p}\sum_{j=1}^{p}\lambda_j = \tau$ for some positive constant τ, by the weak law of large numbers, we have

$$\frac{1}{p}\sum_{j=1}^{G}\lambda_j\tilde{Z}_j^2 \xrightarrow{\text{P}} \tau. \tag{7.9}$$

Next, we study the numerator of $\Psi_{\mathbf{a}}$ in (7.8). Since $\frac{1}{p}\sum_{j=1}^{p}\alpha_j^2 \to 0$, the first term of the numerator converges to 0 in probability, i.e.,

$$\mathbf{Z}'\boldsymbol{\alpha}/\sqrt{p} \xrightarrow{\text{P}} 0. \tag{7.10}$$

Let $\boldsymbol{\varepsilon} = \sqrt{\frac{n_1 n_2}{n}}(\hat{\boldsymbol{\epsilon}}_1 + \hat{\boldsymbol{\epsilon}}_2)$ and $\tilde{\boldsymbol{\varepsilon}} = \mathbf{V}^{-1/2}\mathbf{Q}\boldsymbol{\varepsilon}$, then $\tilde{\boldsymbol{\varepsilon}}$ has distribution $N(0, \mathbf{I})$ and is independent of $\tilde{\mathbf{Z}}$. The second term of the numerator can be written as

$$\mathbf{Z}'[\sqrt{n_1 n_2/n}(\hat{\boldsymbol{\epsilon}}_1 + \hat{\boldsymbol{\epsilon}}_2)] = \tilde{\mathbf{Z}}'\mathbf{V}^{1/2}\tilde{\boldsymbol{\varepsilon}} = \sum_{j=1}^{p}\sqrt{\lambda_j}\tilde{Z}_j\tilde{\varepsilon}_j.$$

Since $\frac{n}{n_1 n_2 p}\sum_{j=1}^{p}\lambda_j \to 0 < \infty$, it follows from the weak law of large number that

$$\sqrt{\frac{n}{n_1 n_2 p}}\sum_{j=1}^{p}\sqrt{\lambda_j}\tilde{Z}_j\tilde{\varepsilon}_j \xrightarrow{\text{P}} 0.$$

This together with (7.8), (7.9), and (7.10) completes the proof. ■

We need the the following two lemmas to prove Theorem 3.

Lemma 1 *[Cao(2005)] Let $n = n_1 + n_2$. Assume that there exist $0 < c_1 \le c_2 < 1$ such that $c_1 \le n_1/n_2 \le c_2$. Let $\tilde{T}_j = T_j - \frac{\mu_{j1}-\mu_{j2}}{\sqrt{S_{1j}^2/n_1+S_{1j}^2/n_2}}$. Then for any $x \equiv x(n_1, n_2)$ satisfying $x \to \infty$ and $x = o(n^{1/2})$,*

$$\log P(\tilde{T}_j \ge x) \sim -x^2/2, \text{ as } n_1, n_2 \to \infty.$$

If in addition, if we have only $E|Y_{1ij}|^3 < \infty$ and $E|Y_{2ij}|^3 < \infty$, then

$$\frac{P(\tilde{T}_j \ge x)}{1-\Phi(x)} = 1 + O(1)(1+x)^3 n^{-1/2} d^3, \text{ for } 0 \le x \le n^{1/6}/d,$$

where $d = (E|Y_{1ij}|^3 + E|Y_{2ij}|^3)/(\mathrm{var}(Y_{1ij}) + \mathrm{var}(Y_{2ij}))^{3/2}$ and $O(1)$ is a finite constant depending only on c_1 and c_2. In particular,

$$\frac{P(\tilde{T}_j \ge x)}{1-\Phi(x)} \to 1$$

uniformly in $x \in (0, o(n^{1/6}))$.

Lemma 2 *Suppose Condition 1(b) holds and $\log p = o(n)$. Let S_{kj}^2 be the sample variance defined in Section 1, and σ_{kj}^2 be the variance of the j-th feature in class $\mathcal{C}_k$. Suppose $\min \sigma_{kj}^2$ is bounded away from 0. Then we have the following uniform convergence result*

$$\max_{k=1,2,\ j=1,\cdots,p} |S_{kj}^2 - \sigma_{kj}^2| \xrightarrow{P} 0.$$

Proof of Lemma 2. For any $\varepsilon > 0$, we know when n_k is very large,

$$\begin{aligned} P(\max_{k=1,2,\ j=1,\cdots,p} |S_{kj}^2 - \sigma_{kj}^2| > \varepsilon) &\le \sum_{k=1,2} \sum_{j=1}^{p} P(|S_{kj}^2 - \sigma_{kj}^2| > \varepsilon) \\ &\le \sum_{k=1,2} \sum_{j=1}^{s} P\Big(|\sum_{i=1}^{n_k} (\epsilon_{kij}^2 - \sigma_{kj}^2)| > n_k\varepsilon/3\Big) + \sum_{k=1,2} \sum_{j=1}^{s} P\Big(|\sum_{i=1}^{n_k} \epsilon_{kij}| > n_k\sqrt{\varepsilon}/2\Big) \\ &\equiv I_1 + I_2. \end{aligned} \tag{7.11}$$

It follows from Bernstein's inequality that

$$P\Big(|\sum_{i=1}^{n_k}(\epsilon_{kij}^2-\sigma_{kj}^2)|>n_k\varepsilon/3\Big)\le 2\exp\{-\frac{1}{2}\frac{n_k^2\varepsilon^2}{9\nu_1+3M_1n_k\varepsilon}\},$$

and

$$P\Big(|\sum_{i=1}^{n_k}\epsilon_{kij}|>n_k\sqrt{\varepsilon}/2\Big)\le 2\exp\{-\frac{1}{2}\frac{n_k^2\varepsilon}{4\nu_2+2M_2n_k\sqrt{\varepsilon}}\},$$

Since $\log p = o(n)$, we have $I_1 = o(1)$ and $I_2 = o(1)$. These together with (7.11) completes the proof of Lemma 2. ■

Proof of Theorem 3. We devidte the proof into two parts. (a) Let us first look at the probability $P(\max_{j>s}|T_j|>x)$. Clearly,

$$P(\max_{j>s}|T_j|>x)\le\sum_{j=s+1}^{p}P(|T_j|\ge x). \tag{7.12}$$

Note that for all $j>s$, $\alpha_j=\mu_{j1}-\mu_{j2}=0$. By Condition 1(b) and Lemma 1, the following inequality holds for $0\le x\le n^{1/6}/d$,

$$P(T_j\ge x)=(1-\Phi(x))(1+C(1+x)^3n^{-1/2}d^3),$$

where C is a constant that only depends on c_1 and c_2, and

$$d=(E|Y_{1ij}|^3+E|Y_{2ij}|^3)/(\sigma_{1j}^2+\sigma_{2j}^2)^{3/2}$$

with σ_{kj}^2 the j-th diagonal element of $\mathbf{\Sigma}_k$. For the normal distribution, we have the following tail probability inequality

$$1-\Phi(x)\le\frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}.$$

This together with the symmetry of T_j gives

$$P(|T_j|\ge x)\le 2\frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}(1+C(1+x)^3n^{-1/2}d^3).$$

Combining the above inequality with (7.12), we have

$$\sum_{j>s} P(|T_j| \geq x) \leq (p-s)\frac{2}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}(1+C(1+x)^3 n^{-1/2}d^3).$$

Since $\log(p-s) = o(n^\gamma)$ with $0 < \gamma < \frac{1}{3}$, if we let $x \sim cn^{\gamma/2}$, then

$$\sum_{j>s} P(|T_j| \geq x) \to 0,$$

which along with (7.12) yields

$$P(\max_{j>s} |T_j| > x) \to 0.$$

(b) Next, we consider $P(\min_{j\leq s} |T_j| \leq x)$. Notice that for $j \leq s$, $\alpha_j = \mu_{1j} - \mu_{2j} \neq 0$. Let $\eta_j = \frac{\alpha_j}{\sqrt{S_{1j}^2/n_1 + S_{1j}^2/n_2}}$ and define

$$\tilde{T}_j = T_j - \eta_j.$$

Then following the same lines as those in (a), we have

$$\sum_{j\leq s} P(|\tilde{T}_j| \geq x) \leq s\frac{2}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}(1+C(1+x)^3 n^{-1/2}d^3) \to 0.$$

It follows from Lemma 2 that,

$$\max_{j\leq s} |S_{kj}^2 - \sigma_{kj}^2| \xrightarrow{\mathrm{P}} 0, \ k = 1, 2.$$

Hence, uniformly over $j = 1, \cdots, s$, we have

$$\eta_j = \frac{|\alpha_j|}{\sqrt{\sigma_{1j}^2/n_1 + \sigma_{2j}^2/n_2}}(1 + o_P(1)).$$

Therefore,

$$\min_{j\leq s} |\eta_j| = \min_{j\leq s} \frac{\sqrt{n_1}|\alpha_j|}{\sqrt{\sigma_{1j}^2 + \sigma_{2j}^2 n_1/n_2}}(1+o_P(1)) \geq \min_{j\leq s} \frac{\sqrt{n_1}|\alpha_j|}{\sqrt{\sigma_{1j}^2 + c_2\sigma_{2j}^2}}(1+o_P(1))$$

with c_2 defined in Theorem 3. Let $\alpha_0 = \min_{j\le s} |\mu_{j1}-\mu_{j2}|/\sqrt{\sigma_{1j}^2 + c_2\sigma_{2j}^2}$. Then it follows that

$$P(\min_{j\le s}|T_j| \le x) \le P(\max_{j\le s}|\tilde{T}_j| \ge \min_{j\le s}|\eta_j| - x) \le P(\max_{j\le s}|\tilde{T}_j| \ge \sqrt{n_1}\alpha_0(1+o_P(1)) - x).$$

By part (a), we know that $x \sim cn^{\gamma/2}$ and $\log(p-s) = o(n^\gamma)$. Thus if $\alpha_0 \sim \min_{j\le s} \frac{|\mu_{j1}-\mu_{j2}|}{\sqrt{\sigma_{j1}^2+\sigma_{j2}^2}} = n^{-\gamma}\beta_n$ for some $\beta_n \to \infty$, then similarly to part (a), we have

$$P(\min_{j\le s}|T_j| \le x) \to 0.$$

Combination of part (a) and part (b) completes the proof. ■

Proof of Theorem 4. The classification error of the truncated classifier $\hat{\delta}_{\mathrm{NC}}^m$ is

$$W(\hat{\delta}_{\mathrm{NC}}^m, \boldsymbol{\theta}) = 1 - \Phi\left(\frac{\sum_{j=1}^m \hat{\alpha}_j(\mu_{1j} - \hat{\mu}_j)}{\sum_{j=1}^m \hat{\alpha}_j^2}\right).$$

We first consider the denominator. Note that $\hat{\alpha}_j \sim N(\alpha_j, \frac{n}{n_1n_2})$. It can be shown that

$$\Big(\frac{4n}{n_1n_2}\sum_{j=1}^m \alpha_j^2 + \frac{2mn^2}{n_1^2n_2^2}\Big)^{-1/2}\sum_{j=1}^m(\hat{\alpha}_j^2 - \alpha_j^2 - \frac{n}{n_1n_2}) \xrightarrow{\mathrm{D}} N(0,1),$$

which together with the assumption $\frac{n}{\sqrt{m}}\sum_{j=1}^m \alpha_j^2 \to \infty$ gives

$$\begin{aligned}\sum_{j=1}^m \hat{\alpha}_j^2 &= \sum_{j=1}^m \alpha_j^2 + \frac{mn}{n_1n_2} + \{\frac{4n}{n_1n_2}\sum_{j=1}^m \alpha_j^2 + \frac{2mn^2}{n_1^2n_2^2}\}^{1/2}O_p(1)\\ &= (1+o_P(1))\sum_{j=1}^m \alpha_j^2 + \frac{mn}{n_1n_2}.\end{aligned}$$

Next, let us look at the numerator. We decompose it as

$$\sum_{j=1}^m \hat{\alpha}_j(\mu_{1j} - \hat{\mu}_j) = \frac{1}{2}\sum_{j=1}^m \alpha_j^2 - \sum_{j=1}^m \alpha_j\hat{\epsilon}_{2j} - \frac{1}{2}\sum_{j=1}^m(\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2). \tag{7.13}$$

Since the second term above has the distribution $N(0, \sum_{j=1}^m \alpha_j^2/n_2)$, it follows from the assumption $n\sum_{j=1}^m \alpha_j^2 \to \infty$ that

$$\sum_{j=1}^m \alpha_j \hat{\epsilon}_{2j} = o_P(1) \sum_{j=1}^m \alpha_j^2.$$

The third term in (7.13) can be written as

$$\sum_{j=1}^m (\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2) = \frac{m}{n_1} - \frac{m}{n_2} + O_p(\frac{nm}{n_1 n_2}) = \frac{m(n_2 - n_1)}{n_1 n_2} + o_P(1) \sum_{j=1}^m \alpha_j^2.$$

Hence the numerator is

$$\sum_{j=1}^m \hat{\alpha}_j(\mu_{1j} - \hat{\mu}_j) = \frac{m(n_2 - n_1)}{n_1 n_2} + (1 + o_P(1)) \sum_{j=1}^m \alpha_j^2.$$

Therefore, the classification error is

$$W(\hat{\delta}_{\mathrm{NC}}^m, \boldsymbol{\theta}) = 1 - \Phi\Big(\frac{(1 + o_P(1)) \sum_{j=1}^m \alpha_j^2 + m(n_1 - n_2)/(n_1 n_2)}{2\{(1 + o_P(1)) \sum_{j=1}^m \alpha_j^2 + mn/(n_1 n_2)\}^{1/2}} \Big).$$

This concludes the proof. ■

Proof of Theorem 5. Note that the classification error of $\hat{\delta}_{\mathrm{FAIR}}^{b_n}$ is

$$W(\hat{\delta}_{\mathrm{FAIR}}^{b_n}(\boldsymbol{x}), \boldsymbol{\theta}) = 1 - \Phi\Big(\frac{\sum_j (\mu_{1j} - \hat{\mu}_j)\hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\}}{\sum_j \hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\}} \Big) \equiv 1 - \Phi(\Psi^H).$$

We divide the proof into two parts: the numerator andthe denominator.

(a) First, we study the numerator of Ψ^H. It can be decomposed as

$$\sum_j (\mu_{1j} - \hat{\mu}_j)\hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\} = I_1 + I_2,$$

where $I_1 = \sum_{j \in \mathcal{A}} (\mu_{1j} - \hat{\mu}_j)\hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\}$ and $I_2 = \sum_{j \in \mathcal{A}^c} (\mu_{1j} - \hat{\mu}_j)\hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\}$ with $\mathcal{A}^c$ the complementary of the set $\mathcal{A}$. Note that

$$\begin{aligned} I_2 &= \frac{1}{2} \sum_{j \in \mathcal{A}^c} \alpha_j^2 1\{|\hat{\alpha}_j| \geq b_n\} - \sum_{j \in A^c} \alpha_j \hat{\epsilon}_{2j} 1\{|\hat{\alpha}_j| \geq b_n\} - \frac{1}{2} \sum_{j \in A^c} (\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2) 1\{|\hat{\alpha}_j| \geq b_n\} \\ &\equiv \frac{1}{2} I_{2,1} - I_{2,2} - \frac{1}{2} I_{2,3}. \end{aligned}$$

Since $\hat{\alpha}_j \sim N(\alpha_j, \frac{n}{n_1 n_2})$, it follows from the normal tail probability inequality that for every $j \in \mathcal{A}^c$ and $b_n > \max_{j\in\mathcal{A}^c} |\alpha_j|$,

$$
\begin{aligned}
(7.14) \qquad P(|\hat{\alpha}_j| \geq b_n) &\leq P(|\hat{\alpha}_j - \alpha_j| \geq b_n - \max_{j\in\mathcal{A}^c} |\alpha_j|) \\
&\leq M \frac{\exp\{-n_1 n_2 (b_n - \max_{j\in A^c} |\alpha_j|)^2/(2n)\}}{\sqrt{n_1 n_2 n^{-1}}(b_n - \max_{j\in\mathcal{A}^c} |\alpha_j|)},
\end{aligned}
$$

where M is a generic constant. Thus for every $\varepsilon > 0$, if $\log(p-m)/[n(b_n - \max_{j\in\mathcal{A}^c} |\alpha_j|)^2] \to 0$ and $\max_{j\in\mathcal{A}^c} |\alpha_j| < b_n$, we have

$$
\begin{aligned}
P(|I_{2,1}| \geq \varepsilon) &\leq \varepsilon^{-1} \sum_{j\in A^c} \alpha_j^2 P(|\hat{\alpha}_j| \geq b_n) \\
&\leq M \max_{j\in\mathcal{A}^c} \alpha_j^2 \frac{(p-m)}{\varepsilon} \frac{\exp\{-n_1 n_2 (b_n - \max_{j\in A^c} |\alpha_j|)^2/(2n)\}}{\sqrt{n_1 n_2 n^{-1}}(b_n - \max_{j\in\mathcal{A}^c} |\alpha_j|)},
\end{aligned}
$$

which tends to zero. Hence,

$$
(7.15) \qquad I_{2,1} \xrightarrow{\mathrm{P}} 0
$$

We next consider $I_{2,2}$. Since $E(\hat{\epsilon}_{2j})^2 = \frac{1}{n_2}$, $\log(p-m)/[n(b_n - \max_{j\in A^c} |\alpha_j|)^2] \to 0$, and $\max_{j\in\mathcal{A}^c} |\alpha_j| < b_n$, we have

$$
\begin{aligned}
P(|I_{2,2}| \geq \varepsilon) \leq & \varepsilon^{-1} \sum_{j\in A^c} E|\hat{\epsilon}_{2j} \alpha_j 1\{|\hat{\alpha}_j| \geq b_n\}| \\
\leq & \varepsilon^{-1} \sum_{j\in A^c} \{E(\hat{\epsilon}_{2j})^2\}^{1/2} \{E\alpha_j^2 1\{|\hat{\alpha}_j| \geq b_n\}|\}^{1/2} \\
\leq & M \frac{(p-m)\max_{j\in\mathcal{A}^c} |\alpha_j|}{\varepsilon\sqrt{n_2}} \frac{\exp\{-n_1 n_2 (b_n - \max_{j\in A^c} |\alpha_j|)^2/(4n)\}}{\sqrt{n_1 n_2 n^{-1}}(b_n - \max_{j\in\mathcal{A}^c} |\alpha_j|)},
\end{aligned}
$$

which converges to 0. Therefore,

$$
(7.16) \qquad I_{2,2} \xrightarrow{\mathrm{P}} 0.
$$

Then, we consider $I_{2,3}$. Since $c_1 \leq n_1/n_2 \leq c_2$ and $E(\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2)^2 = \frac{3n_1^2 + 3n_2^2 - 2n_1 n_2}{n_1^2 n_2^2} \leq$

$\frac{3c_2+3-2c_1}{c_1 n_2^2}$, by (7.14) we have for every $\varepsilon > 0$,

$$\begin{aligned} P(|I_{2,3}| \geq \varepsilon) &\leq \varepsilon^{-1} E\Big| \sum_{j \in A^c} (\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2) 1\{|\hat{\alpha}_j| \geq b_n\} \Big| \\ &\leq \varepsilon^{-1} \sum_{j \in A^c} \{E(\hat{\epsilon}_{1j}^2 - \hat{\epsilon}_{2j}^2)^2\}^{1/2} P(|\hat{\alpha}_j| \geq b_n)^{1/2} \\ &\leq M \frac{\sum_{j \in \mathcal{A}^c} P(|\hat{\alpha}_j| \geq b_n)}{n_2 \varepsilon} \longrightarrow 0, \end{aligned}$$

where M is some generic constant. Thus, $I_{2,3} \xrightarrow{\text{P}} 0$. Combination of this with (7.15) and (7.16) entails

$$I_2 = o_P(1).$$

We now deal with I_1. Decompose I_1 similarly as

$$\begin{aligned} I_1 &= \sum_{j \in \mathcal{A}} (\mu_{1j} - \hat{\mu}_{1j}) \hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\} + \frac{1}{2} \sum_{j \in \mathcal{A}} \hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\} \\ &\geq \sum_{j \in \mathcal{A}} (\mu_{1j} - \hat{\mu}_{1j}) \hat{\alpha}_j 1\{|\hat{\alpha}_j| \geq b_n\} + \frac{1}{2} \sum_{j \in \mathcal{A}} (\hat{\alpha}_j^2 - b_n^2) \\ &\equiv I_{1,1} + \frac{1}{2} I_{1,2}. \end{aligned}$$

We first study $I_{1,2}$. By using $\hat{\alpha}_j \sim N(\alpha_j, \frac{n}{n_1 n_2})$, it can be shown that

$$\Big(\frac{4n}{n_1 n_2} \sum_{j \in \mathcal{A}} \alpha_j^2 + \frac{2mn^2}{n_1^2 n_2^2}\Big)^{-1/2} \sum_{j \in \mathcal{A}} (\hat{\alpha}_j^2 - \alpha_j^2 - \frac{n}{n_1 n_2}) \xrightarrow{\text{D}} N(0,1). \tag{7.17}$$

Since $\frac{n}{\sqrt{m}} \sum_{j=1}^m \alpha_j^2 \to \infty$, we have $\Big(\frac{4n}{n_1 n_2} \sum_{j \in \mathcal{A}} \alpha_j^2 + \frac{2mn^2}{n_1^2 n_2^2}\Big)^{1/2} \Big/ \sum_{j \in \mathcal{A}} \alpha_j^2 \to 0$. Therefore,

$$\begin{aligned} I_{1,2} &= \sum_{j \in \mathcal{A}} (\alpha_j^2 - b_n^2) + \frac{nm}{n_1 n_2} + \Big(\frac{4n}{n_1 n_2} \sum_{j \in \mathcal{A}} \alpha_j^2 + \frac{2mn^2}{n_1^2 n_2^2}\Big)^{1/2} O_p(1) \\ &= (1 + o_P(1)) \sum_{j \in \mathcal{A}} \alpha_j^2 + \frac{nm}{n_1 n_2} - mb^2. \end{aligned}$$

Next, we look at $I_{1,1}$. For any $\varepsilon > 0$,

$$P(|I_{1,1}| \geq \varepsilon) \leq \frac{1}{\varepsilon} E|I_{1,1}| \leq \frac{1}{\varepsilon} \sum_{j \in \mathcal{A}} \{E|\mu_{1j} - \hat{\mu}_{1j}|^2 E|\hat{\alpha}_j|^2\}^{1/2} = \frac{1}{\sqrt{n_1}\varepsilon} \sum_{j \in \mathcal{A}} \sqrt{\alpha_j^2 + n/(n_1 n_2)}.$$

When n is large enough, the above probability can be bounded by

$$P(|I_{1,1}| \geq \varepsilon) \leq \sqrt{2/(n_1\varepsilon^2)} \sum_{j\in\mathcal{A}} |\alpha_j|,$$

which along with the assumption $\sum_{j\in\mathcal{A}} |\alpha_j|/[\sqrt{n}\sum_{j\in\mathcal{A}} \alpha_j^2] \to 0$ gives

$$I_{1,1} = o_P(1) \sum_{j\in\mathcal{A}} \alpha_j^2.$$

It follows that the numerator is bounded from below by

$$(1 + o_P(1))\frac{1}{2}\sum_{j\in\mathcal{A}} \alpha_j^2 + \frac{mn}{2n_1n_2} - \frac{1}{2}mb^2.$$

(b) Now, we study the denominator of Ψ. Let

$$\sum_j \hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\} = \sum_{j\in\mathcal{A}} \hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\} + \sum_{j\in\mathcal{A}^c} \hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\} \equiv J_1 + J_2.$$

We first show that $J_2 \xrightarrow{\mathrm{P}} 0$. Note that $E\hat{\alpha}_j^4 = \alpha_j^4 + 6n(n_1n_2)^{-1}\alpha_j^2 + 3n^2(n_1n_2)^{-2}$. Thus,

$$\begin{aligned} P(|J_2| \geq \varepsilon) &\leq \frac{1}{\varepsilon}E|J_2| = \sum_{j\in A^c} E\hat{\alpha}_j^2 1\{|\hat{\alpha}_j| \geq b_n\}/\varepsilon \leq \frac{1}{\varepsilon}\sum_{j\in A^c} \{E\hat{\alpha}_j^4 P(|\hat{\alpha}_j| \geq b_n)\}^{1/2} \\ &\leq \frac{1}{\varepsilon}\sum_{j\in A^c} \{(\alpha_j^4 + 6n(n_1n_2)^{-1}\alpha_j^2 + 3n^2(n_1n_2)^{-2})P(|\hat{\alpha}_j| \geq b_n)\}^{1/2}. \end{aligned}$$

This together with (7.14) and the assumption that $\log(p-m)/[n(b_n - \max_{j\in\mathcal{A}^c}|\alpha_j|)^2] \to 0$ yields $J_2 \xrightarrow{\mathrm{P}} 0$ as $n \to \infty,\ p \to \infty$. Now we study term J_1. By (7.17), we have

$$J_1 \leq \sum_{j\in\mathcal{A}} \hat{\alpha}_j^2 = (1 + o_P(1)) \sum_{j\in\mathcal{A}} \alpha_j^2 + \frac{mn}{n_1n_2}.$$

Hence the denominator is bounded from above by $(1+o_P(1))\sum_{j\in\mathcal{A}} \alpha_j^2 + \frac{mn}{n_1n_2}$. Therefore,

$$\Psi^H \geq \frac{(1 + o_P(1))\sum_{j\in\mathcal{A}} \alpha_j^2 + \frac{mn}{n_1n_2} - mb^2}{2\sqrt{(1 + o_P(1))\sum_{j\in\mathcal{A}} \alpha_j^2 + \frac{mn}{n_1n_2}}}.$$

It follows that the classification error is bounded from above by

$$1 - \Phi\Big(\frac{(1+o_P(1))\sum_{j\in\mathcal{A}}\alpha_j^2 + \frac{mn}{n_1 n_2} - mb^2}{2\sqrt{(1+o_P(1))\sum_{j\in\mathcal{A}}\alpha_j^2 + \frac{mn}{n_1 n_2}}}\Big).$$

This completes the proof. ■

REFERENCES

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19** 563–570.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension : by an example of a two sample problem. *Statistica Sinica* **6**, 311-329.

Bair, E., Hastie, T., Debashis, P., and Tibshirani, R. (2007) Prediction by supervised principal components. *The Annals of Statistics*, to appear.

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

Boulesteix, A. (2004). PLS Dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology* **3** 1–33.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association* **98**, 324-339.

Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, 19, 1252–1258.

Cao, H.Y. (2007). Moderate deviations for two sample t-statistics. *Probability and Statistics*, Forthcoming.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences* **176** 123–144.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** No. 9, 1061-1069.

Dudoit, S., Fridlyard, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97** 77–87.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91** 674–688.

Fan, J., Hall, P. and Yao, Q. (2006). To how many simultaneous hypothesis tests can normal, student's t or Bootstrap calibration be applied? *Manuscript.*

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, 595–622.

Fan, J and Ren, Y. (2006). Statistical analysis of DNA microarray data. *Clinical Cancer Research* **12** 4469–4473.

Fan, J. and Lv, J. (2007). Sure independence screening for ultra-high dimensional feature space. *Manuscript.*

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165-175.

Ghosh, D. (2002). Singular value decomposition regression modeling for classification of tumors from microarray experiments. *Proceedings of the Pacific Symposium on Biocomputing*, 11462–11467.

Greenshtein, E. (2006). Best subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, to appear.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988.

Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics* **19** 2072–2978.

Lin, Z. and Lu, C. (1996). *Limit Theory for Mixing Dependent Random Variables.* Kluwer Academic Publishers, Dordrecht.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, to appear.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18** 39–50.

Shao, Q. M. (2005). Self-normalized Limit Theorems in Probability and Statistics. *Manuscript*.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

West, M., Blanchette, C., Fressman, H., Huang, E., Ishida, S., Spang, R., Zuan, H., Marks, J. R. and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Natl. Acad. Sci.* **98** 11462–11467.

Zou, H., Hastie, T., and Tibshirani. R. (2004). Sparse principal component analysis. *Technical report*.

Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu

Yingying Fan
Department of Statistics
Harvard University
Cambridge, Massachusetts 02138
USA
E-mail: yingying@princeton.edu