Pattern recognition based on canonical correlations in a high dimension low sample size context

Mitsuru Tamatani\textsuperscript{a}, Inge Koch\textsuperscript{b}, Kanta Naito\textsuperscript{c,*}

\textsuperscript{a} Graduate School of Science and Engineering, Shimane University, Japan
\textsuperscript{b} School of Mathematical Sciences, The University of Adelaide, Australia
\textsuperscript{c} Department of Mathematics, Shimane University, Japan

\textbf{A R T I C L E  I N F O}

\textbf{Article history:}
Received 5 October 2011
Available online 3 May 2012

\textbf{AMS subject classifications:}
62H30
62H20
62E20

\textbf{Keywords:}
Canonical Correlations
Consistency
High Dimension Low Sample Size
Misclassification
Naive Bayes rule

\textbf{A B S T R A C T}

This paper is concerned with pattern recognition for 2-class problems in a High Dimension Low Sample Size (HDLSS) setting. The proposed method is based on canonical correlations between the predictors $X$ and responses $Y$. The paper proposes a modified version of the canonical correlation matrix $\Sigma^{-1/2} X \Sigma^{-1/2} Y$ which is suitable for discrimination with class labels $Y$ in a HDLSS context. The modified canonical correlation matrix yields ranking vectors, a discriminant direction and a rule which is essentially equivalent to the naive Bayes rule. The paper examines the asymptotic behavior of the ranking vectors and the discriminant direction and gives precise conditions for HDLSS consistency in terms of the growth rates of the dimension and sample size. The feature selection induced by the discriminant direction as ranking vector is shown to work efficiently in simulations and in applications to real HDLSS data.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

We consider the problem of classifying $d$-dimensional random vectors $X$ into one of two classes or populations. Following Devroye et al. [2], we use the notion Pattern Recognition synonymously with Classification and Discrimination.

We assume that the two populations, $C_1$ and $C_0$, have multivariate normal distributions which differ in their means $\mu_1$ and $\mu_0$, but share a common covariance matrix $\Sigma$. For $X$ from one of the two classes, Fisher’s discriminant function

$$\delta(X) = \left( X - \frac{1}{2}(\mu_1 + \mu_0) \right)^T \Sigma^{-1}(\mu_1 - \mu_0)$$

assigns $X$ to $C_1$ if $\delta(X) > 0$, and to $C_0$ otherwise.

If the random vectors have equal probability of belonging to either of the two classes, then the probability of misclassification, based on (1), is $\Phi(-\Delta^2/2) + \{1 - \Phi(\Delta^2/2)\}/2$, where $\Phi$ is the standard normal distribution function, and $\Delta$ is the Mahalanobis distance between the two populations:

$$\Delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0)}.$$

* Corresponding author.
E-mail address: naito@riko.shimane-u.ac.jp (K. Naito).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
A sample version of (1) is
\[
\hat{\delta}(X) = \left( X - \frac{1}{2} (\hat{\mu}_1 + \hat{\mu}_0) \right)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0),
\]
where \(\hat{\mu}_1, \hat{\mu}_0\) and \(\hat{\Sigma}\) are appropriate estimates for the corresponding population parameters of the two classes.

Throughout this paper we consider data in a high dimension low sample size (HDLSS) setting, that is, we assume that the dimension \(d\) of the data is bigger than the sample size \(n\). In this framework, the inverse of the sample covariance matrix does not exist. Instead of using a generalized inverse, a natural and simple remedy is the replacement of \(\Sigma\) by the diagonal matrix \(\hat{\Sigma} = \text{diag}\hat{\Sigma}\) in (2), which results in the discriminant function for the naive Bayes rule or classifier. Bickel and Levina [1] investigate asymptotic error probabilities of the naive Bayes classifier and include \(\hat{\Sigma}\) in a HDLSS setting. Fan and Fan [3] extend their results: they derive bounds for the asymptotic error probabilities of the naive Bayes rule and they propose to integrate a reliable and efficient feature selection method, called FAIR, into the naive Bayes rule.

Less attention has been devoted to consistency properties of the naive Bayes rule. In principal component analysis, Johnstone [6] proposes the notion of spiked covariance matrices and Jung and Marron [7] combine the spikiness of the covariance matrix with HDLSS consistency, which is a measure of the closeness of two vectors in a HDLSS setting, and is given in terms of the angle between the vectors. Jung and Marron [7] derive precise conditions under which the first eigenvector of the sample covariance matrix is HDLSS consistent. These conditions include the asymptotic rate of growth of the first eigenvalue which exceeds the dimension \(d\). Like Jung and Marron [7], we consider the asymptotic behavior of the first eigenvector and eigenvalue in a HDLSS classification context. In our setting, however, their spiked covariance model is not appropriate, since we are dealing with inverses of the covariance matrix. As a result, our conditions for consistency differ considerably from theirs.

The aims of this paper are to relate linear discriminant rules and canonical correlation analysis, and to analyze asymptotic properties of discriminant directions. We start with an adaptation to the classification of the canonical correlation matrix \(C = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}\), which Koch and Naito [8] use for variable ranking in regression with predictors \(X\) and responses \(Y\), and we show that our modified canonical correlation matrix \(\hat{C}\) naturally leads to the naive Bayes rule, and provides a justification for the variable selection used in [3]. This canonical correlation based framework allows us to give precise conditions for the HDLSS consistency of the discriminant direction, and leads to a feature selection algorithm which provides an alternative to FAIR of Fan and Fan [3].

The paper is organized as follows. In Section 2, we derive a relationship between Fisher’s linear discriminant function and canonical correlation analysis for data from two classes. This relationship explicitly exhibits Fisher’s ratio of the within-class and between-class variances, and its maximizer provides a criterion for selecting a discriminant rule. In Section 3, we present a similar discussion for a HDLSS setting; we replace \(\Sigma\) by \(\text{diag}\Sigma\), and obtain a corresponding criterion, which leads to a natural derivation of the naive Bayes rule. In Section 4, we analyze the asymptotic behavior of the first singular value and the corresponding canonical correlation vector of the modified matrix \(C\) in an HDLSS setting. A discussion of the consistency of the naive Bayes direction is given in Section 5, which requires a restriction of the parameter space. In Section 6, we use the conditions, required for the consistency results of the naive Bayes rule, and show how these conditions lead to a smaller upper bound for the asymptotic error probability of the naive Bayes rule than that obtained in [3]. In Section 7, we propose a method for feature selection, which naturally follows from our analysis of the matrix \(C\). Section 8 applies our approach to simulated and real data. We illustrate the performance of the naive Bayes rule on these data. Our numerical results confirm the theoretical results of Section 5, and show the good performance of our feature selection for the naive Bayes rule. The conclusions are found in Section 9. The Appendix contains proofs of the theoretical results.

2. Fisher’s rule and canonical correlations

In this section we review the relationship between canonical correlations and Fisher’s rule in pattern recognition.

Throughout this paper, we let \(C_1\) and \(C_0\) be two \(d\)-dimensional normal populations with different means \(\mu_1\) and \(\mu_0\) and the same covariance matrix \(\Sigma\). For a random vector \(X\) from one of these populations/classes, let \(\pi\) be the probability that \(X\) belongs to \(C_1\), and let \(1 - \pi\) be the probability that \(X\) belongs to \(C_0\).

We use linear discriminant functions \(\delta\) of the form
\[
\delta(X) = \left[ X - \frac{1}{2} (\mu_1 + \mu_0) \right]^T b,
\]
where \(X\) is a \(d\)-dimensional random vector from one of the two classes, and \(b\) is a suitably chosen direction vector. It is well known that the choice
\[
b = \Sigma^{-1} (\mu_1 - \mu_0)
\]
yields the optimal classifier, called the Bayes rule, see [2]. Further, choosing \(b\) as in (3), leads to Fisher’s rule, see [4]. Fisher’s idea is to obtain the vector \(b\) which maximizes the ratio of the between-class variance and within-class variance.
In a canonical correlation analysis of two subsets of variables $X^{[1]}$ and $X^{[2]}$ of a random vector $X$, the canonical correlation matrix

$$C = \Sigma^{-1/2}_1 \Sigma_{1,2} \Sigma^{-1/2}_2,$$

and the derived matrix $CC^T$ play important roles in multivariate analysis, see [10]. Here $\Sigma_k$ is the covariance matrix of $X^{[k]}$ ($k = 1, 2$), and $\Sigma_{1,2}$ is the between-covariance matrix of the two vectors.

For random vectors $X$ belonging to one of the classes $C_1$ and $C_0$, we replace $X^{[1]}$ in (4) by $X$, and $X^{[2]}$ by the vector of labels $Y$ defined by

$$Y = \begin{bmatrix} Y_1 \\ Y_0 \end{bmatrix}, \quad Y_1 = \begin{cases} 1 & \text{with probability } \pi, \\ 0 & \text{with probability } 1-\pi, \end{cases} \quad Y_0 = 1 - Y_1.$$  

(5)

For $X$ and $Y$ we define the matrix

$$C = \Sigma^{-1/2} E \left[ (X - \mu_X) Y^T \right] \left[ E \left[ YY^T \right] \right]^{-1/2},$$

(6)

where $\mu_X = \mu_1 + (1-\pi) \mu_0$. Strictly speaking, the centered $Y$ should be used for $C$ in (6); however, for the vector of labels $Y$, centering is not meaningful. Indeed, the matrix $E \left[ YY^T \right]$ and its sample counterpart, which we consider in Section 4.2, lead to natural and easily interpretable expressions. From (5) and (6) it follows that

$$E \left[ YY^T \right] = \begin{bmatrix} \pi & 0 \\ 0 & 1 - \pi \end{bmatrix},$$

$$E \left[ (X - \mu_X) Y^T \right] = (\mu_1 - \mu_0) [\pi (1-\pi) - \pi (1-\pi)],$$

and hence

$$CC^T = \rho \Sigma^{-1/2}(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \Sigma^{-1/2},$$

where $\rho = \pi (1-\pi)$.

Now, consider the eigenvalue problem $CC^T p = \lambda p$. Using the expression (7), we see that solving the eigenvalue problem is equivalent to maximizing $J(b)$ defined by

$$J(b) = \frac{b^T (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T b}{b^T \Sigma b} \left( -\frac{\lambda}{\rho} \right).$$

(8)

This expression is nothing other than the criterion which yields Fisher’s rule with $b = c_\Sigma \Sigma^{-1/2} p$, $p$ the eigenvector of $CC^T$ and $c_\Sigma = \|\Sigma^{-1/2}(\mu_1 - \mu_0)\|$.  

3. Naive Bayes and its derivation  

In a HDLSS setting the choice of $b$ as in (3) is not reliable since the natural sample based estimator includes an estimate of $\Sigma$, which becomes singular. Thus, Fisher’s rule experiences problems for HDLSS data.

To overcome these difficulties, the discriminant function based on

$$b = D^{-1}(\mu_1 - \mu_0),$$

(9)

with $D = \text{diag} \Sigma$, has been discussed in the literature, where it is known as the naive Bayes discriminant function from Bickel and Levina [1]. This $b$ applies to HDLSS settings, since it only takes into account the marginal variances of $X$.

Call $CC^T$ the naive canonical correlation matrix, where

$$\tilde{C} = D^{-1/2} E \left[ (X - \mu_X) Y^T \right] \left[ E \left[ YY^T \right] \right]^{-1/2}$$

is the naive version of $C$, which is obtained from $C$ in (6) by replacing $\Sigma$ by $D$. Let $\tilde{p}$ be the eigenvector belonging to the largest eigenvalue $\tilde{\lambda}$ of $CC^T$. As in (7) and (8), we obtain the analogous criterion

$$J(\tilde{b}) = \frac{\tilde{b}^T (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \tilde{b}}{\tilde{b}^T D \tilde{b}} \left( -\frac{\tilde{\lambda}}{\rho} \right),$$

(10)

where $\tilde{b} = c_\Sigma D^{-1/2} \tilde{p}$ with $c_\Sigma = \|D^{-1/2}(\mu_1 - \mu_0)\|$. Thus, the maximizer of (10) yields the naive Bayes discriminant function or the naive Bayes rule.

4. Asymptotics for the eigenvalue and eigenvector  

In this section we investigate the behavior of an estimator $\hat{p}$ of $\tilde{p}$, the eigenvector of $\tilde{C} \tilde{C}^T$ described in the previous section. Throughout this and later sections, we let $k = 0, 1$.

4.1. The empirical setting  

Consider the data:

$$(X_{11}, Y_{11}), \ldots, (X_{1n_1}, Y_{1n_1}), (X_{01}, Y_{01}), \ldots, (X_{0n_0}, Y_{0n_0}).$$
where the \( X_{ki} = (X_{k1}, \ldots, X_{kd})^\top \) are independently distributed as \( N_d(\mu_k, \Sigma) \). The vector valued labels \( Y_{ki} (i = 1, \ldots, n_k) \) are independent realizations of (5), and are defined by

\[
Y_{kj} = \begin{bmatrix} k \\ 1 - k \end{bmatrix}, \quad j = 1, \ldots, n_k.
\]

Let \( X \) and \( Y \) be matrices defined as

\[
X = [X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2n}], \quad Y = [Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots, Y_{2n}].
\]

We put \( n = n_1 + n_0 \), so the size of \( X \) is \( d \times n \) and the size of \( Y \) is \( 2 \times n \).

Next, we derive an empirical version of \( \tilde{C} \) and its eigenvector \( \tilde{p} \). Let

\[
P = I_n - \frac{1}{n} 1_n 1_n^\top
\]

be the centering matrix, where \( I_n \) is the \( n \)-dimensional identity matrix and \( 1_n \) is the \( n \)-dimensional vector of ones. Define estimators \( \hat{\mu}_k \) and \( \hat{\Sigma} \) of \( \mu_k \) and \( \Sigma \) by

\[
\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}, \quad \hat{\Sigma} = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{ki} - \hat{\mu}_k)(X_{ki} - \hat{\mu}_k)^\top.
\]

A natural estimator of \( \tilde{C} \) is

\[
\hat{C} = \hat{D}^{-1/2} \left( \frac{1}{n} (X P Y^\top) \right) \left( \frac{1}{n} YY^\top \right)^{-1/2}, \tag{11}
\]

where \( \hat{D} = \text{diag} \hat{\Sigma} \). Using \( \hat{\mu}_k \), (11) can be written as

\[
\hat{C} = \frac{1}{\sqrt{n}} \sqrt{n_1 n_0} \hat{D}^{-1/2} (\hat{\mu}_1 - \hat{\mu}_0) [\sqrt{n_0}, -\sqrt{n_1}],
\]

from which we obtain the expression

\[
\hat{C}^\top \hat{C} = \hat{\lambda} \left( \frac{1}{\sqrt{n}} \sqrt{n_1 n_0} \right) \left( \frac{1}{\sqrt{n}} \sqrt{n_1 n_0} \right)^\top, \tag{12}
\]

where

\[
\hat{\lambda} = \frac{n_0 n_1}{n^2} (\hat{\mu}_1 - \hat{\mu}_0)^\top \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0).
\]

From (12), it follows that the rank of \( \hat{C}^\top \hat{C} \) is one. The nonzero eigenvalue of \( \hat{C}^\top \hat{C} \) is \( \hat{\lambda} \) and its eigenvector is

\[
\hat{p}_0 \equiv \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{n_0} \\ -\sqrt{n_1} \end{bmatrix}.
\]

Hence the eigenvector \( \hat{p} \) of \( \hat{C} \hat{C}^\top \) is given by

\[
\hat{p} = \frac{\hat{D}^{-1/2} (\hat{\mu}_1 - \hat{\mu}_0) \hat{D}^{-1/2} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^\top \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}}. \tag{13}
\]

4.2. Asymptotic behavior of \( \hat{p} \) and \( \hat{\lambda} \)

Jung and Marron [7] analyze the asymptotic behavior of the principal component directions in a HDLSS setting. As done in [7], we make use of the notion of HDLSS consistency in our study of the asymptotic behavior of \( \hat{p} \).

For the remainder of this paper we use the notation \( a_{n,d} = O(b_{n,d}) \) to mean \( a_{n,d}/b_{n,d} \to M \) for some \( M > 0 \) as \( n, d \to \infty \).

We require \( M > 0 \), to distinguish this case from the case \( a_{n,d} = o(b_{n,d}) \), where \( M = 0 \).
Definition 1. Let \( a, b \in \mathbb{R}^d \) be vectors of length one, then the angle between \( a \) and \( b \) is defined by
\[
\angle(a, b) = \arccos \left( a^T b \right).
\]
Let \( x \in \mathbb{R}^d \) be a non-stochastic unit vector. Let \( \hat{x} \) be an estimate of \( x \) based on the sample of size \( n \), and assume that \( \hat{x} \) has unit length. Then, \( \hat{x} \) is HDLSS consistent with \( x \), if
\[
\hat{x}^T x \xrightarrow{p} 1 \quad \text{as} \ n, d \to \infty.
\]

The statements, Conditions A–C, list properties of the data \( X \) which we will refer to in our theorems.

Condition A. Let \( X = [X_{11}, \ldots, X_{in}, X_{01}, \ldots, X_{0n}] \equiv [X_1, \ldots, X_n] \) be a sequence of multivariate normal data from two classes \( C_1 \) and \( C_0 \), which is indexed by the dimension \( d \) and satisfies
\[
X_i = \mu_k + \epsilon_i, \quad \text{where} \quad \mu_k = \begin{cases} 
\mu_1 & \text{if } X_i \text{ belongs to class } C_1 \\
\mu_0 & \text{if } X_i \text{ belongs to class } C_0
\end{cases}
\]
\[
\epsilon_i \sim N(0, \Sigma), \quad \Sigma = (\sigma_{k\ell}) \,(k, \ell \leq d), \quad \text{and the } \epsilon_i \text{ are independent for } i \leq n.
\]

Condition B (Cramér’s Condition). The \( \epsilon_i = [\epsilon_{i1}, \ldots, \epsilon_{id}]^T \) of (14) satisfy: There exist constants \( v_1, v_2, M_1 \) and \( M_2 \) such that
\[
E[|\epsilon_{ij}|^m] \leq m!M_1^{m-2}v_1/2, \quad E[|\epsilon_{ij} - \sigma_{ij}|^m] \leq m!M_2^{m-2}v_2/2 \quad \text{for all } m \in \mathbb{N}.
\]

Condition C. \( \log d = o(n) \), \( n = o(d) \), as \( n \to \infty, \ d \to \infty. \)

Remarks on Condition B: Cramér’s condition states moment assumptions on centered univariate random variables which imply Bernstein’s inequality, Lemma A.2 of Fan and Fan [3] and finally \( D = D(1 + o_p(1)) \) in the proof of Theorem 1.

To proceed with the asymptotic calculations, we use the parameter space \( \Gamma \) of Fan and Fan [3]:
\[
\Gamma = \left\{(\mu_1, \mu_0, \Sigma) | (\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) \geq C_d, \lambda_{\max}(R) \leq b_0, \min_{1 \leq j \leq d} \sigma_{jj} > 0 \right\}
\]
where \( C_d \) is a positive sequence that depends only on \( d \), \( R \) is the correlation matrix \( R = D^{-1/2} \Sigma D^{-1/2} \), and \( \lambda_{\max}(R) \) is the largest eigenvalue of \( R \). For evaluating the behavior of \( \hat{p} \), the eigenvector
\[
\tilde{p} = \frac{D^{-1/2}(\mu_1 - \mu_0)}{\sqrt{(\mu_1 - \mu_0)^T D^{-1}(\mu_1 - \mu_0)}}
\]
of \( \tilde{C} \tilde{C}^T \), which we introduced just before (10), plays an important role. We have the following theorems:

Theorem 1. Suppose that the data \( X \) satisfy Conditions A–C, and that \( d = o(nC_d) \) and \( n_0/n_1 = O(1) \). Then, for all parameters \( \theta \in \Gamma \),
\[
\angle(\hat{p}, \tilde{p}) \xrightarrow{p} 0, \quad \text{as } d \to \infty.
\]

Theorem 2. Suppose that the data \( X \) satisfy Conditions A–C, and that \( d/(n_0C_d) \to k \,(k > 0) \) and \( n_1/n \to 1/\xi \,(\xi > 1) \). Then, for all parameters \( \theta \in \Gamma \),
\[
\angle(\hat{p}, \tilde{p}) \xrightarrow{p} \arccos \left( \frac{1}{\sqrt{1 + \xi k}} \right), \quad \text{as } d \to \infty.
\]

Proofs of Theorems 1 and 2 are given in the Appendix. We note that the rate of growth of \( d \) relative to \( nC_d \) is crucial in Theorems 1 and 2. Theorem 1 states that if \( d = o(nC_d) \) is satisfied, \( \hat{p} \) is HDLSS consistent with \( \tilde{p} \). On the other hand if \( d = O(nC_d) \) holds, then \( \hat{p} \) is not consistent, and the angle between \( \tilde{p} \) and \( \hat{p} \) converges to a nonzero degree.

Next, we consider the asymptotic behavior of the largest eigenvalue \( \lambda \) of \( \tilde{C} \tilde{C}^T \). The population matrix \( \tilde{C} \tilde{C}^T \) is
\[
\tilde{C} \tilde{C}^T = \rho D^{-1/2}(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T D^{-1/2}
\]
with \( \rho \) as in (7), and the largest eigenvalue \( \tilde{\lambda} \) of \( \tilde{C} \tilde{C}^T \) is
\[
\tilde{\lambda} = \rho(\mu_1 - \mu_0)^T D^{-1}(\mu_1 - \mu_0).
\]
The corresponding largest eigenvalue of \( \tilde{C} \tilde{C}^T \) is \( \tilde{\lambda} \) as in (12) since \( \tilde{C} \tilde{C}^T \) and \( \tilde{C} \tilde{C}^T \) have the same nonzero eigenvalue. We summarize the behavior of \( \tilde{\lambda} \) with respect to \( \tilde{\lambda} \) in Theorem 3.
Theorem 3. Suppose that the data $X$ satisfy Conditions A–C, and that $n_0/n_1 = O(1)$. Then, as $d \to \infty$, for any $(\mu_1, \mu_0, \Sigma) \in \Gamma$ satisfying $(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0)/C_d = O(1)$,

\[
\hat{\lambda} = \begin{cases} 
1 + o_P(1), & d = o(n_C), \\
1 + \frac{d}{n(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0)} + o_P(1), & d = O(n_C).
\end{cases}
\]

(16)

The rate of growth of $d$ relative to $n_C$ also plays a key role in Theorem 3: $\hat{\lambda}$ is a consistent estimator of $\tilde{\lambda}$, provided $d \ll n_C$; and $\lambda$ has a relative bias $d/\{n(\mu_1 - \mu_0)^T D^{-1}(\mu_1 - \mu_0)\}$ under the condition $d = O(n_C)$ and is therefore no longer consistent.

5. Asymptotics for the direction of discrimination

In (9) we introduce the vector $\tilde{b} = D^{-1}(\mu_1 - \mu_2)$ as the population direction for the naive Bayes discriminant function $\delta_{NB}(X) = [X - (\mu_1 + \mu_2)/2]^T \tilde{b}$. This $\tilde{b}$ is a scaled version of the eigenvector $\tilde{p}$ of the naive canonical correlation matrix $\tilde{C} \tilde{C}^T$, namely $\tilde{b} = c_D D^{-1/2} \tilde{p}$, and $c_D = ||D^{-1/2}(\mu_1 - \mu_2)||$. Section 4 establishes large sample properties of the estimators $\hat{C} \hat{C}^T$ of $\tilde{C} \tilde{C}^T$, and $\hat{p}$ of $\tilde{p}$. It is natural to define the sample based counterpart $\tilde{b}$ to $\tilde{b}$ by

\[
\tilde{b} = D^{-1}(\hat{\mu}_1 - \hat{\mu}_0) = c_D D^{-1/2} \hat{p},
\]

(17)

where $c_D = ||D^{-1/2}(\hat{\mu}_1 - \hat{\mu}_0)||$, and the sample based naive Bayes discriminant function using this direction $\hat{b}$, which is suitable for a HDLSS setting.

In this section, we investigate the asymptotic behavior of $\tilde{b}$, or more precisely, its normalized version. Put

\[
\tilde{b}_{NB} = \frac{D^{-1/2} \tilde{p}}{\sqrt{\tilde{p}^T D^{-1} \tilde{p}}}
\]

for the population, and

\[
\hat{b}_{NB} = \frac{D^{-1/2} \hat{p}}{\sqrt{\hat{p}^T D^{-1} \hat{p}}}
\]

for the sample.

We consider the behavior of $\hat{b}_{NB}$ on the parameter space

\[
\Gamma^* = \{(\mu_1, \mu_0, \Sigma)(\mu_1 - \mu_0)^T D^{-2}(\mu_1 - \mu_0) \geq C_d, \lambda_{\max}(R) \leq b_0, \min_{1 \leq j \leq d} \sigma_{jj} > 0\}.
\]

(18)

We have the following theorem:

Theorem 4. Suppose that the data $X$ satisfy Conditions A–C, and that $d = O(n_C)$ and $n_0/n_1 = O(1)$. Then, for all parameters $\theta \in \Gamma^*$, we have

\[
L(\tilde{b}_{NB}, \hat{b}_{NB}) \overset{p}{\to} 0, \quad \text{as } d \to \infty.
\]

To describe the behavior of $\hat{b}_{NB}$ for the faster rate of growth $d = O(n_C)$, which we considered in Theorem 2, we require the parameter space

\[
\Gamma^{**} = \{(\mu_1, \mu_0, \Sigma)|C_d \geq (\mu_1 - \mu_0)^T D^{-2}(\mu_1 - \mu_0) \geq C_d, \lambda_{\max}(R) \leq b_0, \min_{1 \leq j \leq d} \sigma_{jj} > 0\},
\]

where $C_d$ is a positive sequence that depends on $d$ only.

Theorem 5. Suppose that the data $X$ satisfy Conditions A–C, and that $C_d^*/C_d = O(1)$. Moreover, assume that $d/(n_0 C_d^*) \to k, d/(n_0 C_d) \to k^*, n_1/n \to 1/\xi, \min_{1 \leq j \leq d} \sigma_{jj} > 1/\sigma_0$ and $\max_{1 \leq j \leq d} \sigma_{jj} < 1/\sigma_0^*$ for $k, k^* > 0, \xi > 1$, and $\sigma_0, \sigma_0^* > 0$. Then, for all parameters $\theta \in \Gamma^{**}$,

\[
\arccos\left(\frac{1}{\sqrt{1 + \xi \sigma_0^* k^*}}\right) (1 + o_P(1)) < \angle(\tilde{b}_{NB}, \hat{b}_{NB}) < \arccos\left(\frac{1}{\sqrt{1 + \xi \sigma_0 k}}\right) (1 + o_P(1)).
\]

We note that

\[
0 = \arccos(1) < \arccos\left(\frac{1}{\sqrt{1 + \xi \sigma_0 k}}\right).
\]
It follows that \( \widehat{b}_{NB} \) of Theorem 5 is not consistent, and as in Theorems 1 and 2, we have the same two regimes for the growth rate of \( d \) which determine the HDLSS consistency, or the inconsistent behavior of the eigenvector and the naive Bayes direction vector. Now consider the parameter space

\[
\Gamma \cap \Gamma^{**} = \left\{ (\mu_1, \mu_0, \Sigma) \mid \left( \mu_1 - \mu_0 \right)^T D^{-1} \left( \mu_1 - \mu_0 \right) \geq C_d, \lambda_{\max}(R) \leq b_0, \min_{1 \leq j \leq d} \sigma_{\gamma} > 0 \right\}.
\]

If \( d = o(nC_d) \) is satisfied, then \( \widehat{b}_{NB} \) is HDLSS consistent for all \( \theta \in \Gamma \cap \Gamma^{**} \). By contrast, if \( d = O(nC_d) \) holds, then \( \widehat{b}_{NB} \) is not consistent for any \( \theta \in \Gamma \cap \Gamma^{**} \).

6. Relation to error probability

In this section we point out that, under the assumptions which lead to the HDLSS consistency of \( \widehat{b}_{NB} \), we can achieve a smaller upper bound for the error probability than has previously been established in [3].

The error probability of a discriminant function \( \delta \) for a parameter \( \theta \) is defined as

\[
W(\delta, \theta) = P(\delta(X) \leq 0 \mid X_{ki}, k = 1, 0, i = 1, \ldots, n_k),
\]

where the new observation \( X \) is assumed to be from class \( C_1 \). The worst case classification error for \( \delta \) is defined as

\[
W(\delta) = \max_{\theta \in \Gamma} W(\delta, \theta),
\]

where \( \Gamma \) is the parameter space in (15). Let

\[
\widehat{\delta}_{NB}(X) = \left( X - \frac{1}{2} (\mu_1 + \mu_0) \right)^T \widehat{D}^{-1} (\mu_1 - \mu_0)
\]

be the discriminant function of the naive Bayes rule. Our Theorem 6 quotes the upper bound for the classification error \( W(\widehat{\delta}_{NB}, \theta) \), which is derived in Theorem 1 of Fan and Fan [3].

**Theorem 6** (Fan and Fan [3]). Suppose that the data \( X \) satisfy Conditions A–C, and that \( d = O(nC_d) \). Then, for \( \theta \in \Gamma \), the classification error \( W(\widehat{\delta}_{NB}, \theta) \) is bounded above by

\[
W(\widehat{\delta}_{NB}, \theta) \leq 1 - \phi \left( \frac{\sqrt{n_1n_0/(dn)}a^1D^{-1} \alpha (1 + o_p(1)) + (n_1 - n_0) \sqrt{d/(nn_1n_0)}}{2\sqrt{\lambda_{\max}(R)} \sqrt{1 + n_1n_0a^1D^{-1} \alpha (1 + o_p(1))/dn}} \right),
\]

where \( \alpha = \mu_1 - \mu_0 \).

The worst case classification error for \( \widehat{\delta}_{NB} \) is also derived in [3], and is given under the assumption \( d = o(nC_d) \). Their result is

\[
W(\widehat{\delta}_{NB}) = 1 - \phi \left( \frac{1}{2} \left[ n_1n_0/(dnb_0) \right]^{1/2} C_d(1 + o_p(1)) \right).
\]

This result seems to be derived from the bound stated in Theorem 6—part (i) of their Theorem 1. Our calculations in Theorem 7 are based on the explicit assumption \( d = o(nC_d) \), which yields the tighter bound (19).

**Theorem 7.** Suppose that the data \( X \) satisfy Conditions A–C, and that \( d = o(nC_d) \). Then, for \( \theta \in \Gamma \), the classification error \( W(\widehat{\delta}_{NB}, \theta) \) is bounded above by

\[
W(\widehat{\delta}_{NB}, \theta) \leq 1 - \phi \left( \frac{\sqrt{\alpha^1D^{-1} \alpha}}{2\sqrt{\lambda_{\max}(R)}} (1 + o_p(1)) \right),
\]

with \( \alpha = \mu_1 - \mu_0 \). Moreover, for the worst case classification error, we have

\[
W(\widehat{\delta}_{NB}) = 1 - \phi \left( \frac{\sqrt{C_d}}{2\sqrt{b_0}} (1 + o_p(1)) \right).
\]

Note that Theorem 6 is the result for \( d = O(nC_d) \), while \( d = o(nC_d) \) is assumed in Theorem 7. We see that, again, the rate of growth of \( d \) relative to \( nC_d \) plays an important role. The upper bounds in Theorems 6 and 7 have the following relationship:
Corollary 8. Suppose that the data $X$ satisfy Conditions A–C and $n_0/n_1 = c + o(1)$ for $1 \leq c < \infty$. Then, as $d \to \infty$, for any $(\mu_1, \mu_0, \Sigma) \in \Gamma$ satisfying $d = o(nc_d)$ and, $\alpha^1D^{-1}\alpha/C_d = O(1)$, where $\alpha = \mu_1 - \mu_0$.

\begin{align}
1 - \Phi \left( \frac{\sqrt{n_1n_0/(dn)\alpha^1D^{-1}\alpha(1 + \alpha_\tau(1))} + (n_1 - n_0)\sqrt{d/(nn_1n_0)}}{2\sqrt{\lambda_{\max}(R)} / \sqrt{1 + n_1n_0\alpha^1D^{-1}\alpha(1 + \alpha_\tau(1))/(dn)}} \right) & > 1 - \Phi \left( \frac{\sqrt{\alpha^1D^{-1}\alpha(1 + \alpha_\tau(1))}}{2\sqrt{\lambda_{\max}(R)}} \right).
\end{align}

(20) (21)

Therefore, (21), the bound obtained in Theorem 7, is smaller than the bound (20) of Theorem 6. It is interesting to observe that the assumption $d = o(nc_d)$, which leads to the desirable HdllsS consistency of $\hat{b}_{NB}$ in Theorem 4, is also responsible for the smaller error bound (21).

7. Feature selection

Koch and Naito [8] propose feature selection in a regression context, which is based on two different ‘ranking vectors’: the eigenvector $\hat{p}_1$ of the matrix $\hat{C}C^T$ as in (4), and the first canonical correlation vector $\hat{b}_1 = \hat{\Sigma}_1^{-1/2}p_1$. Analogously, we consider a linear discriminant function $\hat{\delta}_q$ with $\hat{b}$ as in (17), which includes feature selection based on $\hat{b}$ and on $\hat{p}$ as in (13). For notational convenience we write $\hat{q}$ to denote either $\hat{p}$ or $\hat{b}$ as appropriate.

Let $X = (X_1, \ldots, X_q)^T$ be a random vector from one of the two classes $C_k$. Put

\[ \hat{\delta}_q(X) = \sum_{j=1}^d \left( X_j - \frac{1}{2} (\hat{\mu}_{ij} + \hat{\mu}_{0j}) \right) \hat{b}_j I(\hat{q}_j > \eta), \]

(22)

where $I$ is the indicator function and \( \eta > 0 \) is an appropriate threshold.

We interpret (22) in the following way. We first sort the features, that is, the variables of $X$ in decreasing order of the absolute value of the components $\hat{q}_j$ of $\hat{q}$, and then consider the first $m$ features to classify the data.

We write the sorted components of $\hat{q}$ as

\[ |\hat{q}_{i_1}| \geq |\hat{q}_{i_2}| \geq \cdots \geq |\hat{q}_{i_m}| \geq \cdots \geq |\hat{q}_{i_d}| \geq 0. \]

(23)

For the naive canonical correlation matrix $\hat{C}$ of (11), $\hat{b}$ of (17) is the naive version of the canonical correlation vector. In Section 5, $\hat{b}$ is the direction vector for the Hdlls naive Bayes rule; $\hat{b}$ therefore plays the dual role of ranking vector for variable selection, and direction of the naive Bayes discriminant function. We summarize our classification method based on feature selection with $\hat{b}$ in Steps 1–5 below.

Classification and variable ranking based on $\hat{b}$

Step 1. Calculate $\hat{b}$.
Step 2. Sort the components of $\hat{b}$ in descending order of their absolute values as in (23):

\[ |\hat{b}_{i_1}| \geq |\hat{b}_{i_2}| \geq \cdots \geq |\hat{b}_{i_m}| \geq \cdots \geq |\hat{b}_{i_d}| \geq 0. \]

Step 3. Apply the permutation $\tau : \{1, 2, \ldots, d\} \to \{i_1, i_2, \ldots, i_d\}$ to the rows of $X$, and to $\hat{b}$, and then put $\hat{b} \leftarrow \tau(\hat{b})$ and $X \leftarrow \tau(X)$.

Step 4. Find the best truncation $\hat{m}$ of (4.3) in [3]:

\[ \hat{m} = \arg \max_{1 \leq m \leq d} \frac{1}{\lambda_{\max}(R_m)} \left[ \frac{\sum_{j=1}^m (\hat{\mu}_{ij} - \hat{\mu}_{0j})^2 / \hat{\sigma}_{jj} + m(1/n_0 - 1/n_1)}{nm/(n_1n_0) + \sum_{j=1}^m (\hat{\mu}_{ij} - \hat{\mu}_{0j})^2 / \hat{\sigma}_{jj}} \right]^2, \]

where $R_m$ is the correlation matrix of the truncated observations.

Step 5. Classify a new datum $X$ by

1. putting $X \leftarrow \tau(X)$, and
2. assigning $X$ to class $C_1$ if

\[ \hat{\delta}_q(X) = \sum_{i=1}^{\hat{m}} \left( X_i - \frac{1}{2} (\hat{\mu}_{i1} + \hat{\mu}_{0i}) \right) \hat{b}_i > 0. \]

(24)
We refer to the classification of the five steps above as the Naive Canonical Correlation (NACC) approach, thus acknowledging the fact that Mardia et al. [10] call \( \hat{b} \) a canonical correlation vector. For the ranking vector \( \hat{q} = \hat{p} \) in (23), the rule (22) becomes

\[
\hat{\delta}_m(X) = \sum_{j=1}^{m} \left( X_{ij} - \frac{1}{2} (\hat{\mu}_{1j} + \hat{\mu}_{0j}) \right) \hat{b}_{ij}
\]

\[
= \sum_{j=1}^{m} \left( X_{ij} - \frac{1}{2} (\hat{\mu}_{1j} + \hat{\mu}_{0j}) \right) \frac{\hat{\sigma}_{ij}}{\hat{\sigma}_{ij}},
\]

where the \( X_{ij} \) are the sorted entries of \( X \), \( m = 1, \ldots, d \), \( \hat{\alpha} = \hat{\mu}_1 - \hat{\mu}_0 \), and \( \hat{\sigma}_{ij} \) is the \( j \)th diagonal element of \( \hat{D} \) given by

\[
\hat{\sigma}_{ij} = \frac{1}{n} \left\{ (n_0 - 1)S_{0j}^2 + (n_1 - 1)S_{1j}^2 \right\},
\]

and

\[
S_{kj}^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{ikj} - \hat{\mu}_{kj})^2, \quad k = 0, 1; \ j = 1, \ldots, d.
\]

A comparison of the feature selection induced by (23) with the Feature Annealed Independence Rules (FAIR), which Fan and Fan [3] propose, shows that their selection is induced by the two sample \( t \)-statistics, namely, for \( j \)th variable

\[
T_j = \frac{\hat{\delta}_j}{\sqrt{\hat{\sigma}_{jj} \left( \frac{1}{n_0} + \frac{1}{n_1} \right)}}.
\]  \quad (25)

A comparison of (13) and (25) yields that \( T_j = C_p \hat{p}_j \) for all \( j \), where the constant \( C_p \) depends on the sample size. Hence, feature selection or variable ranking based on (23) is essentially equivalent to FAIR, and the eigenvector \( \hat{p} \) of the naive canonical correlation matrix therefore offers a natural explanation for the variable selection in FAIR.

The classifications NACC and FAIR differ in that the initial ranking is based on different vectors; NACC uses \( \hat{b} \), while variable selection in FAIR is based on \( \hat{p} \). As a consequence the order of the variables and the ‘optimal’ number of variables will differ in the two approaches.

We investigate the behavior of \( \hat{\delta}_0 \) in (24) for real and simulated data in the next section.

8. Numerical studies

In this section, we illustrate the theoretical results of the previous sections via numerical experiments, and investigate the performance of discrimination with feature selection for real data.

8.1. Simulation I

In Simulation I, our interests focus on the error probability \( W(\hat{\delta}_{\text{NB}}, \theta) \), and the angle between \( \hat{b}_{\text{NB}} \) and \( \hat{b}_{\text{NB}} \).

We generate \( n_k \) \( d \)-dimensional observations \( X_{kij} \sim \text{i.i.d.} \ N_d(\mu_k, \Sigma) \) for \( d = 200 \) and \( d = 1000 \). For each value of \( d \), we choose \( n = n_0 + n_1 \) such that \( n \leq d \). The estimate of \( \angle(\hat{b}_{\text{NB}}, \hat{b}_{\text{NB}}) \) is obtained as the sample mean over 1000 iterations. Similarly, the estimate of \( W(\hat{\delta}_{\text{NB}}, \theta) \) is calculated as the average of the leave-one out CV (cross-validation) on the 1000 iterations.

In Simulation I, we take \( \mu_1 = 0, \mu_0 = t1 = (t, \ldots, t)^T \), for \( t > 0 \). The covariance matrix \( \Sigma = (\sigma_{ij}) \) has an AR structure:

\[
\sigma_{ij} = \begin{cases} 
1.0, & i = j, \\
(-0.6)^{|i-j|}, & i \neq j.
\end{cases}
\]

We can see that

\[
(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) = dt^2.
\]

Thus, for \( t = 1, (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) = d \). If we take \( C_d = d \) in (18), then the condition \( d = o(nC_d) \) in Theorem 4 is satisfied. For this choice of \( C_d \), parameters, which are elements of \( \Gamma^* \), are also in \( \Gamma^* \), and hence in \( \Gamma^* \). Therefore, the angle between \( \hat{b}_{\text{NB}} \) and \( \hat{b}_{\text{NB}} \) converges to 0.

On the other hand, if \( t = 2/\sqrt{n} \), then \( (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) = (4d)/n \). The condition \( C_d^*/C_d = O(1) \) of Theorem 5 is satisfied for \( C_d = C_d^* = d \). For \( 0 < t \ll 1 \), the parameters \( \sigma_0 = 1 + t \) and \( \sigma_0^* = 1 - t \) satisfy \( 1/\sigma_0 < 1 \) and \( 1/\sigma_0^* > 1 \).
are large, so we expect to select these features from the simulated data. Further, we note that the diagonal elements of $\Sigma$ is the same as in Simulation I. The mean parameters $\mu_0$ and $\mu_1$ show that only the 10th, 20th and 30th features are large, so we expect to select these features from the simulated data. Further, we note that the diagonal elements of

Furthermore, if we set $n_1/n_0 = 1$, we have $\kappa = \kappa^* = d/(n_0C_d) \approx (n/n_0)(d/(nC_d)) = 1/2$. Thus, the above parameters satisfy the conditions of Theorem 5, and the angle between $\hat{b}_{NB}$ and $\hat{b}_{NB}$ therefore does not converge to 0. We have

$$\arccos \left( \frac{1}{\sqrt{1+1-\varepsilon}} \right) (1 + o_p(1)) < \Sigma (\hat{b}_{NB}, \hat{b}_{NB}) < \arccos \left( \frac{1}{\sqrt{1+1+\varepsilon}} \right) (1 + o_p(1))$$

$$\implies \Sigma (\hat{b}_{NB}, \hat{b}_{NB}) \rightarrow \arccos \left( \frac{1}{\sqrt{1+1}} \right) = \frac{\pi}{4}, \quad n, d \to \infty, \varepsilon \to 0.$$  

In fact, the angle between $\hat{b}_{NB}$ and $\hat{b}_{NB}$ will converge to 45°, and $\hat{b}_{NB}$ is therefore strongly inconsistent in the sense of Jung and Marron [7].

Table 1 summarizes the results from Simulation I. The two columns pertaining to “consistent” relate the results under the assumptions of Theorem 4; the column ‘Error’ gives the estimated error probability $W(\hat{b}_{NB}, \theta)$, and the column ‘Degrees’ lists the estimated angle $\Sigma (\hat{b}_{NB}, \hat{b}_{NB})$ in degrees. The two columns “not consistent” show analogous results under the assumptions of Theorem 5.

We note that the angle in the “consistent” columns decreases to zero, as $n$ and $d$ increase. These results agree with Theorem 4. It is interesting to observe that the estimated angles for $n = 30, 50, 150, 200$ are very similar for both values of $d$. In contrast, for the “not consistent” columns, the angle clearly approaches 45°, which agrees with Theorem 5. Fig. 1 complements the results in Table 1; here we show kernel density estimates of the angles based on the 1000 iterations. The values of $n$ in the top row (for $d = 200$) are $n = 8, 20, 80$ and 180, and the bottom row shows $n = 40, 100, 400$ and 900 for $d = 1000$. The left panels focus on the “consistent” columns in the Table 1, and the right panels show density estimates for the “not consistent” columns. These figures clearly illustrate the behavior of the angle as the sample size increases.

Returning to the “Error” columns in Table 1, it is noticeable that the error probability is almost 0 in the “consistent” cases, and is largest when $n$ and $d$ are both large, that is, as $b_{NB}$ becomes strongly inconsistent. For these latter results, $t = 2/\sqrt{n}$, and this choice of $t$ makes the discrimination problem increasingly difficult as $n$ increases.

### 8.2. Simulation II

Simulation II focuses on the performance of $\hat{\delta}_C$ of (24). As mentioned at the end of Section 7, the FAIR approach of Fan and Fan [3] and the NAAC approach share the vector $\hat{b}$ for discrimination, but base their feature selection on different ranking vectors: FAIR essentially chooses features based on (13), while NAAC selects features based on $\bar{b}$.

For Simulation II the parameters $\mu_0$ and $\Sigma_0$ are

$$\mu_0 = 0 \in \mathbb{R}^d, \quad \mu_1 = (\mu_{11}, \ldots, \mu_{1d})^T \in \mathbb{R}^d, \quad \mu_{ij} = \begin{cases} j/4 & j \in \{10, 20, 30\}, \\ 0 & \text{otherwise} \end{cases}$$

$$\Sigma_0 = A^{1/2} \Sigma A^{1/2}, \quad A = \text{diag}(a_{jj}), \quad a_{jj} = j,$$

and $\Sigma$ is the same as in Simulation I. The mean parameters $\mu_0$ and $\mu_1$ show that only the 10th, 20th and 30th features are large, so we expect to select these features from the simulated data. Further, we note that the diagonal elements of
the covariance matrix $\Sigma_0$ are monotonically increasing. The observations about the large features and the behavior of $\Sigma_0$ allow us to compare the performance of the FAIR and NACC approaches under a non-homogeneous variance structure of the features.

For each pair $(d, n)$ and for both FAIR and NACC, we calculate estimates of the error probability as described in Simulation I. However, in this simulation, we use 100 iterations.

The estimates of the error probabilities are tabulated in Table 2, with standard deviation in parentheses. The column $d = 200$ of Table 2 shows that the error probabilities of NACC are smaller than the corresponding values for FAIR except for $n = 10$, where FAIR wins. For $d = 1000$, the superiority of NACC over FAIR is apparent, especially for $n = 50, 100, 150, 200$. The results show the merit of feature selection with $\hat{b}$ over that with the vector $\hat{p}$ of FAIR.

So far we have compared the error probabilities of the two approaches. We now look at the specific features that are selected in the simulations for $(d, n)$ pairs. Fig. 2 shows the frequency of the selected features over 100 simulations. The top panels show the results for $(d, n) = (200, 180)$, and the bottom panels show similar results for $(d, n) = (1000, 180)$. The left panels relate to NACC, and the right panels to FAIR. The figures show clearly that NACC correctly picks the large variables 10, 20 and 30 most of the time, while FAIR selects many other features, typically features with a large variance. The feature selection of NACC is based on $\hat{b}$ and thus on $(\hat{\mu}_1 - \hat{\mu}_0) / \hat{\sigma}_j$. Our results suggest that feature selection with $\hat{b}$ captures the data structure better than FAIR, in particular for larger values of $d$. 

**Fig. 1.** Kernel density estimates of angles in degrees between $\tilde{b}_{NB}$ and $\tilde{b}_{NB}$.
## Table 2
Estimated value of the error probability.

<table>
<thead>
<tr>
<th>d = 200</th>
<th>d = 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nacc</strong></td>
<td><strong>fair</strong></td>
</tr>
<tr>
<td><strong>n = 6</strong></td>
<td></td>
</tr>
<tr>
<td>0.53167</td>
<td>0.53500</td>
</tr>
<tr>
<td>(0.26025)</td>
<td>(0.25436)</td>
</tr>
<tr>
<td><strong>n = 10</strong></td>
<td></td>
</tr>
<tr>
<td>0.51100</td>
<td><strong>0.50300</strong></td>
</tr>
<tr>
<td>(0.20395)</td>
<td>(0.21670)</td>
</tr>
<tr>
<td><strong>n = 30</strong></td>
<td></td>
</tr>
<tr>
<td>0.33933</td>
<td>0.35700</td>
</tr>
<tr>
<td>(0.13848)</td>
<td>(0.15500)</td>
</tr>
<tr>
<td><strong>n = 50</strong></td>
<td></td>
</tr>
<tr>
<td>0.23360</td>
<td>0.25940</td>
</tr>
<tr>
<td>(0.08989)</td>
<td>(0.10830)</td>
</tr>
<tr>
<td><strong>n = 100</strong></td>
<td></td>
</tr>
<tr>
<td>0.17510</td>
<td>0.19280</td>
</tr>
<tr>
<td>(0.04237)</td>
<td>(0.04854)</td>
</tr>
<tr>
<td><strong>n = 150</strong></td>
<td></td>
</tr>
<tr>
<td>0.17993</td>
<td>0.18640</td>
</tr>
<tr>
<td>(0.03349)</td>
<td>(0.03718)</td>
</tr>
<tr>
<td><strong>n = 200</strong></td>
<td></td>
</tr>
<tr>
<td>0.17035</td>
<td>0.17295</td>
</tr>
<tr>
<td>(0.02616)</td>
<td>(0.02658)</td>
</tr>
</tbody>
</table>

## Table 3
Lung cancer data.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of selected genes</th>
<th>NACC</th>
<th>FAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Selected genes</td>
<td>1955</td>
<td>2039</td>
<td>2928</td>
</tr>
<tr>
<td></td>
<td>4345</td>
<td>8928</td>
<td>11238</td>
</tr>
<tr>
<td>Training error</td>
<td>0/32</td>
<td></td>
<td>0/32</td>
</tr>
<tr>
<td>Test error</td>
<td>8/149</td>
<td></td>
<td>8/149</td>
</tr>
</tbody>
</table>

The results of Simulation II illustrate the superiority of NACC over FAIR in two ways: the error probabilities are mostly smaller, and feature selection is more pertinent, especially for higher values of $d$.

### 8.3. Real data

In this section, we consider the lung cancer data that were analyzed in [5]. These data are available at [http://www.chestsurg.org/publications/2002-microarray.aspx](http://www.chestsurg.org/publications/2002-microarray.aspx). The data have two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). There are 12553 genes, the variables, and 181 samples (31 MPM and 150 ADCA). Gordon et al. [5] considered a training set of 16 MPM and 16 ADCA samples, and used the remaining 149 samples for testing. We use the same training and testing subsets.

For the lung cancer data, we compare FAIR with NACC. Table 3 shows the classification results of the two approaches. FAIR selected 14 genes, and resulted in 0 training errors and 8 test errors, while NACC selected only 7 genes which yielded 0 training errors and 8 test errors. Both approaches select features 2039 and 11368. These features may be of interest to medical experts, but we are not concerned with this aspect in the present analysis. The results show that both classification approaches have the same number of errors, so perform equally well, but NACC finds a more parsimonious set of features. Indeed, NACC requires only half the number of features in order to achieve the misclassification that FAIR achieved.

### 9. Discussion

In this paper we have exhibited the relationship between canonical correlation analysis and the naive Bayes discriminant rule for HDLSS data from two classes. We showed that the estimators of the first eigenvector and the canonical correlation vector of the naive canonical correlation matrix are HDLSS consistent, provided $d$ does not grow too fast. Under these growth conditions on $d$, the consistency results enable us to derive an upper bound for the worst case classification error which is smaller than the error previously given in [3].

Our approach, based on the naive canonical correlation matrix, naturally leads to two ranking vectors for feature selection. One of them, the eigenvector of the naive canonical correlation matrix, is equivalent to the vector Fan and Fan use in their FAIR. The second candidate for feature selection, the naive canonical correlation vector, plays the dual role of also being the natural discriminant direction for the naive Bayes rule in HDLSS data. We compare FAIR and our approach NACC, which uses the naive canonical correlation vector for feature selection, on simulated and real data. If the means of the two classes only differ for a small number of variables and the diagonal elements of the covariance matrix increase with the variable number (as in our Simulation II), NACC performs better than FAIR both in selecting the right features and in achieving a lower...
classification error. In addition to the features with a nonzero mean difference, FAIR typically selects features with a large variance, while NACC’s choice of features is not affected by the large variances.

For the HDLSS lung cancer data; FAIR and NACC resulted in the same number of misclassified observations, however, NACC obtained this result with only half the number of features, thus resulting in a more parsimonious model.

For HDLSS data, the inverse of the sample covariance matrix $\Sigma$ does not exist. To circumvent the problem caused by the singular sample covariance matrix, we replaced $\Sigma$ by its diagonal matrix which is invertible. Instead one could use a generalized inverse of $\Sigma$, and derive the ‘generalized versions’ of the vectors $\hat{p}$ and $\hat{b}$. Shin and Eubank [12] consider different Moore–Penrose generalized inverses and associated vectors $\hat{p}$: one based on the pooled sample covariance matrix, and one based on the sample covariance matrix of all observations. The latter inverse and associated vector $\hat{p}$ have some nice properties which are discussed in their paper. It would be of interest to investigate the asymptotic properties of this vector $\hat{p}$ for two and more classes. This will be the topic of future research.

Another possibility for overcoming the problem posed by the singular sample covariance matrix is to consider regularized canonical correlations as proposed in [9], who include smoothing parameters $\alpha$, similar to a ridge regression parameter, in the canonical correlation setting, and then find the appropriate ‘regularized’ vectors $\hat{p}_\alpha$. This solution path clearly applies to HDLSS settings, and the regularized sample solution $\hat{p}_\alpha$ could be used instead of our vector $\hat{p}$ for a suitable choice of the regularization parameters. We will not pursue this approach here.

**Fig. 2.** Frequency of obtained features.
Currently, our model consists of two classes. But it can be extended to general multi-class discriminant problems, which will be of interest in practice. We will pursue this in future work. Other possible research directions include extensions of our theoretical results to the “kernel method” in linear discrimination described in [11].

Acknowledgments

The authors would like to thank the reviewers for many helpful comments and suggestions which improved the paper. The research of the third author was supported in part by KAKENHI 23500350.

Appendix

Proof of Theorem 1. The inner product of \( \hat{p} \) and \( \hat{p} \) is

\[
\langle \hat{p}, \hat{p} \rangle = \frac{(\mu_1 - \mu_0)^T D^{-1/2} \hat{D}^{-1/2} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) \sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T D^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}}.
\]  

From Cramér’s condition, it follows that \( \hat{D} = D(1 + o_p(1)) \). Thus, (26) can be written as

\[
\langle \hat{p}, \hat{p} \rangle = \frac{(\mu_1 - \mu_0)^T \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) (1 + o_p(1))}{\sqrt{(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) \sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T D^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}}.
\]  

In particular, the denominator of (27) becomes

\[
\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T D^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} = \sqrt{(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) (1 + o_p(1)) + \frac{nd}{n_1 n_0}}
\]

by (A.4) in [3]. On the other hand, the numerator of (27) can be decomposed as

\[
(\mu_1 - \mu_0)^T \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) = (\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) + (\mu_1 - \mu_0)^T \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)
\]

where \( \hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} \xi_i}{n_1} \) and \( \hat{\mu}_0 = \frac{\sum_{i=n_1+1}^{n} \xi_i}{n_0} \). From the evaluation of the term \( I_3 \) on p.2626 of Fan and Fan [3], we have

\[
(\mu_1 - \mu_0)^T \hat{D}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) = (\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) (1 + o_p(1)).
\]

Thus, (27) becomes

\[
\langle \hat{p}, \hat{p} \rangle = \frac{1 + o_p(1)}{\sqrt{(1 + o_p(1)) + (n/n_1)(d/n_0(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0))}}
\]

\[
= \frac{1 + o_p(1)}{\sqrt{(1 + o_p(1)) + O(1) o(1)}}
\]

\[
= 1 + o_p(1).
\]

Since ‘arccos’ is a continuous function, the angle between \( \hat{p} \) and \( \hat{p} \) is

\[
\angle(\hat{p}, \hat{p}) = \text{arccos}(\hat{p}^T \hat{p}) = o_p(1),
\]

and this last equality completes the proof. □

Proof of Theorem 2. From (28) in the proof of Theorem 1, and the assumptions of Theorem 1, it follows that

\[
\langle \hat{p}, \hat{p} \rangle = \frac{1 + o_p(1)}{\sqrt{(1 + o_p(1)) + \xi k (1 + o(1))}}
\]

\[
= \frac{1}{\sqrt{1 + \xi k}} (1 + o_p(1)),
\]

and therefore,

\[
\angle(\hat{p}, \hat{p}) = \text{arccos}(\hat{p}^T \hat{p}) = \text{arccos}\left(\frac{1}{\sqrt{1 + \xi k}}\right) (1 + o_p(1))
\]

by an argument similar to that given in the proof of Theorem 1. □
Proof of Theorem 3. Using (A.4) of Fan and Fan [3],

$$(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) = \left\{ \begin{array}{ll}
(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) + \frac{nd}{n_1 n_0} (1 + o_p(1)), & d = o(nC_d), \\
(\mu_1 - \mu_0)^T D^{-1} (\mu_1 - \mu_0) (1 + o_p(1)), & d = O(nC_d).
\end{array} \right.$$

On the other hand, $n_1/n \overset{p}{\to} \pi$ and $n_0/n \overset{p}{\to} 1 - \pi$. Therefore, the ratio of $\lambda$ and $\hat{\lambda}$ satisfies (16). □

Proof of Theorem 4. The inner product $(\hat{b}_N, \hat{b}_N)$ can be expressed as

$$(\hat{b}_N, \hat{b}_N) = \frac{\hat{p}^T D^{-1/2} \hat{D}^{-1/2} \hat{p}}{\sqrt{\hat{p}^T D^{-1} \hat{p}} \sqrt{\hat{p}^T D^{-1} \hat{p}}} = \frac{(\mu_1 - \mu_0)^T D^{-1} \hat{D}^{-1} (\mu_1 - \mu_0)}{\sqrt{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) \sqrt{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)}}. \quad (29)$$

Since $\hat{D} = D(1 + o_p(1))$, we have

$$(\hat{b}_N, \hat{b}_N) = \frac{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)(1 + o_p(1))}{\sqrt{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) \sqrt{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)}}; \quad (29)$$

and we note that the numerator includes $\hat{D}$.

Consider the denominator of (29). We have

$$(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) = (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) + 2(\mu_1 - \mu_0)^T D^{-2} (\bar{\epsilon}_1 - \bar{\epsilon}_0) + (\bar{\epsilon}_1 - \bar{\epsilon}_0)^T D^{-2} (\bar{\epsilon}_1 - \bar{\epsilon}_0) \equiv (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) + 2E_1 + E_2. \quad (30)$$

Now define

$$\tilde{\epsilon} = \sqrt{\frac{n_1 n_0}{n}} V_R^{-1/2} Q_k^T D^{-1/2} (\bar{\epsilon}_1 - \bar{\epsilon}_0),$$

where $V_R$ and $Q_k$ are the matrices obtained from the spectral decomposition of

$$R = Q_k V_k Q_k^T. \quad (31)$$

Let $\lambda_{R,i}$ be the eigenvalues of $R$, so $V_R = \text{diag} [\lambda_{R,1}, \ldots, \lambda_{R,d}]$. We have $\{n/(n_1 n_0)\} G \Sigma G^T = I_d$, where $G = \sqrt{\{n_1 n_0\} / n} V_R^{-1/2} Q_k D^{-1/2}$. Thus, $\tilde{\epsilon} \sim N_d(0, I_d)$. On the other hand,

$$\tilde{\epsilon} = \sqrt{\frac{n_1 n_0}{n}} V_R^{-1/2} Q_k^T D^{-1/2} (\bar{\epsilon}_1 - \bar{\epsilon}_0) \iff D^{-1} (\bar{\epsilon}_1 - \bar{\epsilon}_0) = \sqrt{\frac{n}{n_1 n_0}} D^{-1/2} Q_k V_R^{1/2} \tilde{\epsilon},$$

and $E_2$ of (30) becomes

$$E_2 = (D^{-1} (\bar{\epsilon}_1 - \bar{\epsilon}_0))^T (D^{-1} (\bar{\epsilon}_1 - \bar{\epsilon}_0)) \leq \frac{n}{n_1 n_0} \tilde{\epsilon}^T V_R^{1/2} Q_k^T D^{-1/2} Q_k V_R^{1/2} \tilde{\epsilon} \leq \frac{n}{n_1 n_0} \Sigma_{\max} (D^{-1}) \tilde{\epsilon}^T V_R \tilde{\epsilon}.$$

In particular, $\Sigma_{\max} (D^{-1}) = 1 / \min_{1 \leq j \leq d} \sigma_{ij} < \infty$, since $D^{-1}$ is diagonal. Thus,

$$(\bar{\epsilon}_1 - \bar{\epsilon}_0)^T D^{-2} (\bar{\epsilon}_1 - \bar{\epsilon}_0) \leq \frac{n}{n_1 n_0} \min_{1 \leq j \leq d} \sigma_{ij} \tilde{\epsilon}^T V_R \tilde{\epsilon} \leq \frac{n}{n_1 n_0} \min_{1 \leq j \leq d} \sigma_{ij} (1 + o_p(1))$$

by the weak law of large numbers.
Next, consider $E_1$ of (30), which has the distribution

$$(\mu_1 - \mu_0)^T D^{-2} (\overline{\epsilon}_1 - \overline{\epsilon}_0) \sim N \left( 0, \frac{n}{n_1 n_0} (\mu_1 - \mu_0)^T D^{-2} \Sigma D^{-2} (\mu_1 - \mu_0) \right).$$

From the definition of $I^*$ in (18), the variance of $E_1$ is evaluated as follows:

$$V[E_1] = \frac{n}{n_1 n_0} (\mu_1 - \mu_0)^T D^{-3/2} RD^{-3/2} (\mu_1 - \mu_0)$$
$$\leq \frac{n}{n_1 n_0} \lambda_{\max}(R)(\mu_1 - \mu_0)^T D^{-3} (\mu_1 - \mu_0)$$
$$\leq \frac{n}{n_1 n_0} \lambda_{\max}(R) \lambda_{\max}(D^{-1})(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)$$
$$\leq \frac{n}{n_1 n_0} \frac{1}{\min \sigma_{\bar{\eta}}} (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0).$$

Therefore, using Chebyshev’s inequality,

$$P \left( \left| \frac{(\mu_1 - \mu_0)^T D^{-2} (\overline{\epsilon}_1 - \overline{\epsilon}_0)}{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)} \right| > \epsilon \right) \leq \frac{V[E_1]}{(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) \epsilon^2}$$
$$\leq \left( \frac{n^2}{n_1 n_0} \right) \left( \frac{b_0}{\min \sigma_{\bar{\eta}}} \right) \frac{1}{\epsilon^2 n C_d}$$
$$= o(1).$$

Consequently, $(\mu_1 - \mu_0)^T D^{-2} (\overline{\epsilon}_1 - \overline{\epsilon}_0) = (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) o_p(1)$.

The previous calculations lead to the following bound for (30):

$$(\bar{\mu}_1 - \bar{\mu}_0)^T D^{-2} (\bar{\mu}_1 - \bar{\mu}_0) \leq (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)$$
$$+ 2(\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) o_p(1) + \frac{1}{\min \sigma_{\bar{\eta}}} \frac{nd}{n_1 n_0} (1 + o_p(1))$$
$$\leq (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0)(1 + o_p(1)) + \frac{1}{\min \sigma_{\bar{\eta}}} \frac{nd}{n_1 n_0}. $$

Next, we consider the numerator of (29):

$$(\mu_1 - \mu_0)^T D^{-2} (\bar{\mu}_1 - \bar{\mu}_0) = (\mu_1 - \mu_0)^T D^{-2} (\mu_1 - \mu_0) + (\mu_1 - \mu_0)^T D^{-2} \overline{\epsilon}_1 - (\mu_1 - \mu_0)^T D^{-2} \overline{\epsilon}_0$$

follows by an argument similar to that given in the proof of Theorem 1 in [3].

Combining the calculations for the numerator and denominator of (29) leads to $(\tilde{b}_{\text{NB}}, \tilde{b}_{\text{NB}}) \leq 1$, and, in particular,

$$1 \geq (\tilde{b}_{\text{NB}}, \tilde{b}_{\text{NB}})$$
$$\geq \frac{1}{\sqrt{(1 + o_p(1)) + (n/n_1)(1/\min \sigma_{\bar{\eta}})(d/(n_0 C_d))}}$$
$$= (1 + o_p(1)).$$

Thus, the inner product of $\tilde{b}_{\text{NB}}$ and $\tilde{b}_{\text{NB}}$ converges to 1 in probability. Since arcoss is a continuous function, the angle of $\tilde{b}_{\text{NB}}$ and $b_{\text{NB}}$ satisfies

$$\angle(\tilde{b}_{\text{NB}}, b_{\text{NB}}) = \arccos(\tilde{b}_{\text{NB}}, \tilde{b}_{\text{NB}}) = o_p(1).$$

**Proof of Theorem 5.** Consider the inner product $(\tilde{b}_{\text{NB}}, \tilde{b}_{\text{NB}})$ in the form given in (29), and let $E_2$ be defined as in (30). If we define the spectral decomposition of $R$ as in (31) in the proof of Theorem 4, then

$$E_2 \geq \frac{n}{n_1 n_0} \min_{z \in \mathbb{R}^d - \{0\}} \frac{z^T D^{-1} z}{z^T z} \left( Q_R V_R^{1/2} \hat{\epsilon} \right)^T \left( Q_R V_R^{1/2} \hat{\epsilon} \right)$$
$$\geq \frac{nd}{n_1 n_0} \max_{1 \leq j \leq d} \sigma_j (1 + o_p(1)).$$
and the inner product \((\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}})\) of (29) is bounded by

\[
(\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}}) \leq \frac{(1 + o_p(1))}{\sqrt{(1 + o_p(1)) + \left(\frac{n/n_1}{1/n_1 + \min_{1 \leq j \leq d} \sigma_{jj}}\right)}}.
\]

Therefore, using the parameter space \(I^{**}\), (32) and (33), we have

\[
(\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}}) \geq \frac{(1 + o_p(1))}{\sqrt{(1 + o_p(1)) + \left(\frac{n/n_1}{1/n_1 + \max_{1 \leq j \leq d} \sigma_{jj}}\right)}}.
\]

and

\[
(\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}}) \leq \frac{(1 + o_p(1))}{\sqrt{(1 + o_p(1)) + \left(\frac{n/n_1}{1/n_1 + \max_{1 \leq j \leq d} \sigma_{jj}}\right)}}.
\]

Since \(\text{arccos}\) is a monotonically decreasing continuous function on \([0, 1]\), we derive upper and lower bounds of the angle between \(\widehat{b}_{\text{NB}}\) and \(\widehat{b}_{\text{NB}}\) using the following equivalent statements

\[
\frac{(1 + o_p(1))}{\sqrt{1 + \xi \sigma_0\kappa}} > (\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}}) > \frac{(1 + o_p(1))}{\sqrt{1 + \xi \sigma_0\kappa}}
\]

\[\iff\]

\[
\frac{(1 + o_p(1))}{\sqrt{1 + \xi \sigma_0\kappa}} < \arccos(\widehat{b}_{\text{NB}}, \widehat{b}_{\text{NB}}) < \frac{(1 + o_p(1))}{\sqrt{1 + \xi \sigma_0\kappa}}.
\]

This completes the proof. \(\square\)

**Proof of Corollary 8.** We derive (20) as follows:

\[
1 - \Phi \left( \frac{\sqrt{n_1 n_0/(dn) \alpha^T D^{-1} \alpha (1 + o_p(1)) + (n_1 - n_0) \sqrt{d/(nn_1 n_0)}}}{2 \sqrt{\lambda_{\max}(R) \sqrt{1 + n_1 n_0 \alpha^T D^{-1} \alpha (1 + o_p(1))/(dn)}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(nn_1 n_0 \alpha^T D^{-1} \alpha (n_1 - n_0))}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(dn)/(n_1 n_0 \alpha^T D^{-1} \alpha) + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(n_0 \alpha^T D^{-1} \alpha) (1 - n_0/n_1)}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(n_0 \alpha^T D^{-1} \alpha) (1 - (1 + o(1)))}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
\geq 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(n_0 \alpha^T D^{-1} \alpha) o(1)}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(n_0 \alpha^T D^{-1} \alpha) o(1)}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + d/(n_0 \alpha^T D^{-1} \alpha) o(1) + (1 + o_p(1))}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + O(1)] o(1) + (1 + o_p(1))}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]

\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha (1 + o_p(1)) + O(1)] o(1) + (1 + o_p(1))}}{2 \sqrt{\lambda_{\max}(R) \sqrt{(n/n_1) [d/(n_0 \alpha^T D^{-1} \alpha)] + (1 + o_p(1))}}} \right)
\]
\[
= 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha} (1 + o_P(1))}{2 \sqrt{\lambda_{\text{max}}(R)} \sqrt{1 + O(1)}} \right) \\
> 1 - \Phi \left( \frac{\sqrt{\alpha^T D^{-1} \alpha}}{2 \sqrt{\lambda_{\text{max}}(R)}} (1 + o_P(1)) \right). \quad \square
\]

References


