MULTI-CLASS DISCRIMINANT FUNCTION BASED ON CANONICAL CORRELATION IN HIGH DIMENSION LOW SAMPLE SIZE

Tamatani, Mitsuru
Graduate school of science and engineering, Shimane University

Naito, Kanta
Graduate school of science and engineering, Shimane University

Koch, Inge
School of Mathematical Sciences, The University of Adelaide

https://doi.org/10.5109/1563533

出版情報：Bulletin of informatics and cybernetics. 45, pp.67-101, 2013-12. 統計科学研究会
バージョン：
権利関係：

MULTI-CLASS DISCRIMINANT FUNCTION BASED ON CANONICAL CORRELATION IN HIGH DIMENSION LOW SAMPLE SIZE

by

Mitsuru Tamatani, **Kanta Naito** and **Inge Koch**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.45

FUKUOKA, JAPAN
2013

Bulletin of Informatics and Cybernetics, Vol. 45, 2013

MULTI-CLASS DISCRIMINANT FUNCTION BASED ON CANONICAL CORRELATION IN HIGH DIMENSION LOW SAMPLE SIZE

By

Mitsuru Tamatani[*], **Kanta Naito**[†] and **Inge Koch**[‡]

Abstract

In multi-class discriminant analysis for High Dimension Low Sample Size settings it is not possible to define Fisher's discriminant function, since the sample covariance matrix is singular. For the special case of two-class problems, the naive Bayes rule has been studied, and combined with feature selection, this approach yields good practical results. We show how to extend the naive Bayes rule based on the naive canonical correlation matrix to a general setting for $K \geq 2$ classes, and we propose variable ranking and feature selection methods which integrate information from all $K-1$ eigenvectors. Provided the dimension does not grow too fast, we show that the $K-1$ sample eigenvectors are consistent estimators of the corresponding population parameters as both the dimension and sample size grow, and we give upper bounds for the misclassification rate. For real and simulated data we illustrate the performance of the new method which results in lower errors and typically smaller numbers of selected variables than existing methods.

Key Words and Phrases: High Dimension Low Sample Size, Canonical Correlations, Consistency, Naive Bayes rule, Misclassification, Multi-class Linear Discriminant Analysis.

1. Introduction

Fisher's well-known linear discriminant function for K-class problems is the solution to his paradigm: maximize the between-class variance while minimising the within-class variances. See Fisher (1936), and Rao (1948) for the general multi-class setting. Fisher's rule applies in the classical domain where the sample size n is much larger than the dimension, and the within-class covariance matrix is invertible, but breaks down for high dimension low sample size (HDLSS) problems, since the sample covariance matrix becomes singular when the dimension d exceeds the sample size n.

Several discriminant rules have been proposed for the HDLSS context which overcome the problem of the singular covariance matrix in different ways: Dudoit et al. (2002) proposed diagonal linear discriminant analysis (DLDA) which only uses the diagonal elements of the sample covariance matrix, Srivastava and Kubokawa (2007) proposed a discriminant function based on the Moore-Penrose inverse, and Aoshima and Yata (2011) considered a discriminant rule based on second moments in conjunction with geometric representations of high-dimensional data.

[*] Graduate school of science and engineering, Shimane University, Japan.
[†] Graduate school of science and engineering, Shimane University, Japan.
[‡] School of Mathematical Sciences, The University of Adelaide, Australia.

In this paper we focus on the 'diagonal' approach of Dudoit et al. (2002), which has special appeal since it is conceptually simpler than the competitors, admits a theoretical and asymptotic foundation, and can be implemented efficiently for large and complex data sets. We refer to Fisher's rule based on the diagonal of the covariance matrix as the *naive Bayes rule*.

For two-class problems in a HDLSS setting, Bickel and Levina (2004) and Fan and Fan (2008) investigated the asymptotic behavior of the naive Bayes rule, and calculated bounds for its misclassification rate. Tamatani, Koch and Naito (2012) defined the naive canonical correlation matrix, a modification of the canonical correlation matrix that is suitable for classification problems, and studied the asymptotic behavior of the eigenvector and the discriminant direction of the naive canonical correlation matrix in the context of two-class problems.

The purpose of this paper is to generalise works by Fan and Fan (2008) and Tamatani, Koch and Naito (2012) in two-class setting to multi-class. We propose a naive Bayes rule in a general multi-class setting and investigate its asymptotic properties for high-dimensional data when both the dimension d and the sample size n grow. Throughout this paper we focus on HDLSS data from K classes, that is, we assume that the dimension d of the data is much bigger than the sample size n. In this framework, we consider the problem of classifying d-dimensional random vectors $\boldsymbol{X}$ into one of K classes. Our discriminant approach is based on canonical correlations, and in particular on a modification of the canonical correlation matrix suitable for vector-valued class labels from K classes. In this framework we replace the covariance matrix by its diagonal counterpart as discussed in Dudoit et al. (2002), Bickel and Levina (2004) and Fan and Fan (2008). We call such a matrix a *naive canonical correlation matrix*, and observe that this matrix plays an important role in the present theory. The $K-1$ eigenvectors belonging to $K-1$ non-zero eigenvalues of the estimated naive canonical correlation matrix yield discriminant directions which inform our choice of a discriminant function for K classes. For this setting we study the asymptotic behavior of the eigenvectors and associate discriminant directions. Related research on eigenvalues and eigenvectors for high-dimensional data includes Johnstone (2001), Ahn et al. (2007), Jung and Marron (2009), Fujikoshi et al. (2010) and Tamatani, Koch and Naito (2012).

For HDLSS data from K multivariate normal classes, we derive an upper bound for the misclassification rate of the proposed multi-class discriminant function. Our asymptotic results for the misclassification rate split into two disjoint types depending on the precise growth rates of d and n. Depending on the two distinct growth rates, we also develop HDLSS asymptotic results for estimators of the eigenvectors and discriminant directions.

In high-dimensional settings, it is necessary to select a subset of relevant features (or variables) for discrimination. There are several discussions on feature selection methods for multi-class discriminant analysis, see Saeys et al. (2007). It appears that many of these proposals are variable selection methods which are not specific for or connected with a particular discriminant function. We believe that in discriminant analysis, a comprehensive feature selection method should include the discriminant function in the variable ranking and in the choice of the number of selected features, and we propose new feature selection methods specifically for our multi-class discriminant function.

This paper is organized as follows. In Section 2 we review a relationship between Fisher's rule and canonical correlations for K classes. To avoid the singularity of the

sample covariance matrix in a HDLSS setting, we derive a multi-class version of the naive Bayes rule. Section 3 details the asymptotic behavior of the eigenvectors of the estimated naive canonical correlation matrix and the associated discriminant directions in a HDLSS setting under general distributional assumptions. We derive an upper bound for the asymptotic misclassification rate of the proposed multi-class discriminant function under assumptions of normality, and we show that the upper bound for the asymptotic misclassification rate is indeed a multi-class extension of that obtained in Fan and Fan (2008) and Tamatani, Koch and Naito (2012). Section 5 includes new feature selection methods for multi-class discriminant analysis, which naturally follow from our analysis of the naive canonical correlation matrix. In Section 6, we apply proposed methods to real and simulated data sets, and compare the performance of our methods with other ranking methods. These comparisons demonstrate that our feature selection method works well and yields a parsimonious set of features which lead to good classification results. The conclusions are summarized in Section 7, and the Appendix contains proofs of our theoretical results.

2. Discriminant Function based on Canonical Correlations

In this section we derive direction vectors for discrimination using canonical correlations. For two classes, the relationship between canonical correlation vectors and Fisher's rule is well known. Tamatani, Koch and Naito (2012) discussed two-class discriminant analysis in a HDLSS context based on a naive canonical correlation vector. We extend their theory to the general multi-class HDLSS problem.

2.1. Canonical Correlations

Let $\mathcal{C}_\ell$ $(\ell = 1, \ldots, K)$ be d-dimensional populations with different means $\boldsymbol{\mu}_\ell$ and common covariance matrix Σ. For a random vector $\boldsymbol{X}$ from one of the K classes, let π_ℓ be the probability that $\boldsymbol{X}$ belongs to $\mathcal{C}_\ell$. Let $\boldsymbol{Y}$ be the K-dimensional vector-valued class label with ℓth entry 1 if $\boldsymbol{X}$ belongs to $\mathcal{C}_\ell$ and 0 otherwise, so $P(\boldsymbol{Y} = \boldsymbol{e}_\ell) = \pi_\ell$, and $\sum_{\ell=1}^{K} \pi_\ell = 1$, where $\boldsymbol{e}_\ell$ is the vector with ℓth entry 1 and 0 otherwise.

In a canonical correlation analysis of two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, with $\boldsymbol{Y}$ the vector-valued labels and $\boldsymbol{\mu} = \sum_{\ell=1}^{K} \pi_\ell \boldsymbol{\mu}_\ell$, the matrix

$$\widetilde{C} = \Sigma^{-1/2} E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] \left\{E[\boldsymbol{Y}\boldsymbol{Y}^T]\right\}^{-1/2}$$

plays an important role. From the definition of $(\boldsymbol{X}, \boldsymbol{Y})$ it follows that

$$E[\boldsymbol{Y}\boldsymbol{Y}^T] = \Pi \quad \text{and} \quad E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] = M_0\Pi, \tag{1}$$

where $\Pi = \mathrm{diag}(\pi_1, \ldots \pi_K)$ and $M_0 = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{\mu}_K - \boldsymbol{\mu}]$. Using the between-class covariance matrix $M = M_0 \Pi M_0^T$, we have

$$\widetilde{C}\widetilde{C}^T = \Sigma^{-1/2} M \Sigma^{-1/2}.$$

If we put $\widetilde{\boldsymbol{b}} = \Sigma^{-1/2}\widetilde{\boldsymbol{p}}$, where $\widetilde{\boldsymbol{p}}$ is a solution to the eigenvalue problem $\widetilde{C}\widetilde{C}^T\widetilde{\boldsymbol{p}} = \tilde{\lambda}\widetilde{\boldsymbol{p}}$ with $\tilde{\lambda} > 0$, then the vector $\widetilde{\boldsymbol{b}}$ is the maximizer of the criterion

$$\widetilde{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T \Sigma \boldsymbol{b}} \tag{2}$$

over vectors $\boldsymbol{b}$. Note that (2) is nothing other than the criterion which yields Fisher's rule for the multi-class setting. In particular, the rank of $\widetilde{C}\widetilde{C}^T$ is $K-1$, so the $K-1$ eigenvectors $[\widetilde{\boldsymbol{p}}_1 \cdots \widetilde{\boldsymbol{p}}_{K-1}]$ belonging to $K-1$ non-zero eigenvalues should be used for constructing the discriminant directions

$$\widetilde{B} \equiv [\widetilde{\boldsymbol{b}}_1 \cdots \widetilde{\boldsymbol{b}}_{K-1}] = \Sigma^{-1/2}[\widetilde{\boldsymbol{p}}_1 \cdots \widetilde{\boldsymbol{p}}_{K-1}].$$

Using the discriminant directions $\widetilde{B}$, we want to define a discriminant function $\widetilde{g}$ for classifying new observations $\boldsymbol{X}$ whose class is unknown. For $1 \leq \alpha \leq K-1$, put $Z_\alpha(\boldsymbol{X}) = \widetilde{\boldsymbol{b}}_\alpha^T \boldsymbol{X}$, and define the vector $Z(\boldsymbol{X}) = [Z_1(\boldsymbol{X}), \ldots, Z_{K-1}(\boldsymbol{X})]^T = \widetilde{B}^T \boldsymbol{X}$. For $Z(\boldsymbol{X})$ and $\mathcal{C}_\ell$, define Mahalanobis distances Δ_ℓ by

$$\Delta_\ell(Z(\boldsymbol{X})) = \sqrt{(Z(\boldsymbol{X}) - \boldsymbol{\nu}_\ell)^T \Sigma_\ell^{-1} (Z(\boldsymbol{X}) - \boldsymbol{\nu}_\ell)},$$

where $\boldsymbol{\nu}_\ell = E\left[Z(\boldsymbol{X}) | \boldsymbol{Y} = \boldsymbol{e}_\ell\right] = \widetilde{B}^T \boldsymbol{\mu}_\ell$ and $\Sigma_\ell = V\left[Z(\boldsymbol{X}) | \boldsymbol{Y} = \boldsymbol{e}_\ell\right] = \widetilde{B}^T \Sigma \widetilde{B}$. Hence $\Delta_\ell^2(Z(\boldsymbol{X}))$ can be rewritten as

$$\Delta_\ell^2(Z(\boldsymbol{X})) \quad = \quad (\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T \, \widetilde{B}(\widetilde{B}^T \Sigma \widetilde{B})^{-1} \widetilde{B}^T \, (\boldsymbol{X} - \boldsymbol{\mu}_\ell) \, . \tag{3}$$

Using (3), we now derive the multi-class discriminant function $\widetilde{g}$ as the minimizer of the Mahalanobis distance, and let

$$\widetilde{g}(\boldsymbol{X}) = \underset{\ell \in \{1,\ldots,K\}}{\arg\min} \, \Delta_\ell^2(Z(\boldsymbol{X})).$$

Note that both $\tilde{B}$ and $\Delta_\ell^2(Z(\boldsymbol{X}))$ depend on Σ.

2.2. Naive Canonical Correlations

In the HDLSS two-class discrimination settings of Fan and Fan (2008), the within-class sample covariance matrix $\widehat{\Sigma}$ is singular. For a population framework, it therefore does not make sense to define a discriminant function based on Σ. To be able to establish a consistent theory, we first require a suitable framework for the population. We define the *naive canonical correlation matrix* C and vectors $\boldsymbol{b}_\alpha$ and $\boldsymbol{p}_\alpha$ by

$$C = D^{-1/2} E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] \left\{E[\boldsymbol{Y}\boldsymbol{Y}^T]\right\}^{-1/2} \quad \text{and} \quad \boldsymbol{b}_\alpha = D^{-1/2}\boldsymbol{p}_\alpha,$$

where $D = \mathrm{diag}\Sigma$, and $\boldsymbol{p}_\alpha$ is eigenvector of the matrix CC^T corresponding to the αth largest eigenvalue λ_α^*. Put $P = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_{K-1}]$. The discriminant directions

$$B \equiv [\boldsymbol{b}_1 \cdots \boldsymbol{b}_{K-1}] = [D^{-1/2}\boldsymbol{p}_1 \cdots D^{-1/2}\boldsymbol{p}_{K-1}] = D^{-1/2}P$$

can now be seen to maximise the analogous *naive* criterion

$$J(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T D \boldsymbol{b}}. \tag{4}$$

Note that Σ in (2) has been replaced by the diagonal matrix D in (4). The corresponding discriminant function g is therefore

$$g(\boldsymbol{X}) = \underset{\ell \in \{1,\ldots,K\}}{\arg\min} \, (\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T \, B(B^T D B)^{-1} B^T \, (\boldsymbol{X} - \boldsymbol{\mu}_\ell) \, . \tag{5}$$

Note that Σ has been replaced by D both in (2) and (3) to yield (4) and (5) respectively.

It is worth noting that for $K = 2$, the discriminant function g reduces to the naive Bayes discriminant function of Bickel and Levina (2004).

3. Asymptotic Behavior of the Eigenvector of the Empirical Naive Canonical Correlation Matrix

In this section we investigate the asymptotic behavior of suitable estimators of P and B.

3.1. The Empirical Setting

Consider data $(\boldsymbol{X}_{\ell i}, \boldsymbol{Y}_{\ell i})$ $(\ell = 1, \ldots, K, i = 1, \ldots, n_\ell)$, where the independently distributed $\boldsymbol{X}_{\ell i}$ are from K disjoint classes, and the $\boldsymbol{Y}_{\ell i}$ are independent realizations of vector labels

$$\boldsymbol{Y}_{\ell i} = (Y_{\ell i1}, \ldots, Y_{\ell iK})^T \quad \text{with} \quad Y_{\ell ij} = \begin{cases} 1 & \ell\text{-th component,} \\ 0 & \text{otherwise.} \end{cases}$$

Let X and Y be matrices defined by $X = [\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{Kn_K}]$ and $Y = [\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{Kn_K}]$. Then X is of size $d \times n$, and Y is of size $K \times n$, where $n = \sum_{\ell=1}^{K} n_\ell$.

Next we derive an empirical version of C and its left eigenvectors $\boldsymbol{p}_\alpha$. Define estimators $\widehat{\boldsymbol{\mu}}_\ell$ and $\widehat{\Sigma}$ of $\boldsymbol{\mu}_\ell$ and Σ by

$$\widehat{\boldsymbol{\mu}}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \boldsymbol{X}_{\ell i} \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{K} \sum_{\ell=1}^{K} \widehat{S}_\ell,$$

where

$$\widehat{S}_\ell = \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} (\boldsymbol{X}_{\ell i} - \widehat{\boldsymbol{\mu}}_\ell)(\boldsymbol{X}_{\ell i} - \widehat{\boldsymbol{\mu}}_\ell)^T.$$

Using the centering matrix, an estimator for C is

$$\widehat{C} = \widehat{D}^{-1/2} \left\{ \frac{1}{n} \left(X \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \right) Y^T \right\} \left(\frac{1}{n} Y Y^T \right)^{-1/2} = \widehat{D}^{-1/2} \widehat{M}_0 N^{1/2}, \tag{6}$$

where $\widehat{D} = \mathrm{diag}\widehat{\Sigma}$, I_n is the $n \times n$ identity matrix, $\mathbf{1}_n$ is the n-dimensional vector of ones, $\widehat{M}_0 = [\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}, \ldots, \widehat{\boldsymbol{\mu}}_K - \widehat{\boldsymbol{\mu}}]$ and $N = \mathrm{diag}(n_1/n, \ldots, n_K/n)$. Hence we obtain the expression

$$\widehat{C}\widehat{C}^T \quad = \quad \widehat{D}^{-1/2} \widehat{M} \widehat{D}^{-1/2},$$

where $\widehat{M} = \widehat{M}_0 N \widehat{M}_0^T$.

Since the rank of $\widehat{C}\widehat{C}^T$ is $K - 1$, we use the $K - 1$ eigenvectors $\widehat{\boldsymbol{p}}_\alpha$ of $\widehat{C}\widehat{C}^T$ corresponding to the $K - 1$ non-zero eigenvalues in the definition of the discrimination directions $\widehat{\boldsymbol{b}}_\alpha$, and put $\widehat{\boldsymbol{b}}_\alpha = \widehat{D}^{-1/2} \widehat{\boldsymbol{p}}_\alpha$ $(\alpha = 1, ..., K - 1)$. We note that these $\widehat{\boldsymbol{b}}_\alpha$ can be obtained as the maximizers of the function

$$\widehat{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T \widehat{M} \boldsymbol{b}}{\boldsymbol{b}^T \widehat{D} \boldsymbol{b}}.$$

Put

$$\widehat{B} = [\widehat{\boldsymbol{b}}_1 \cdots \widehat{\boldsymbol{b}}_{K-1}] = [\widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_1 \cdots \widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_{K-1}] \equiv \widehat{D}^{-1/2}\widehat{P}, \tag{7}$$

then $\widehat{g}$, defined by

$$\widehat{g}(\boldsymbol{X}) = \underset{\ell\in\{1,\ldots,K\}}{\arg\min}\ (\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_i)^T\, \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T\,(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_i)\,, \tag{8}$$

is a natural estimator of g in (5), and this $\widehat{g}$ is our proposed discriminant function in the HDLSS multi-class setting. To elucidate the asymptotic behavior of $\widehat{g}$, it is necessary to develop first asymptotics for $\widehat{B}$ as well as $\widehat{P}$ in a HDLSS setting.

3.2. Asymptotic Behavior of $\widehat{P}$ and $\widehat{B}$

In this section, we evaluate the asymptotic behavior of the eigenvectors $\widehat{\boldsymbol{p}}_\alpha$ and discriminant directions $\widehat{\boldsymbol{b}}_\alpha$ in a HDLSS setting under general assumptions about the underlying distributions. Throughout this paper, we make the assumption that the sample size of each of the K classes satisfies $c \le n_\ell/n$ for some positive constant c and $\ell = 1, \ldots, K$.

In what follows we use the asymptotic notation:

1. $a_{n,d} = O(b_{n,d})$ to mean that $a_{n,d}/b_{n,d} \to M \in (0,\infty)$ as $n, d \to \infty$.
2. $a_{n,d} = o(b_{n,d})$ to mean that $a_{n,d}/b_{n,d} \to 0$ as $n, d \to \infty$.

The definition of o is usually included in that of big O, however we distinguish these cases in this paper.

DEFINITION 3.1. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a non-stochastic unit vector, and let $\widehat{\boldsymbol{x}}$, a vector of length one, denote an estimate of $\boldsymbol{x}$ based on the sample of size n . If

$$\widehat{\boldsymbol{x}}^T\boldsymbol{x} \xrightarrow{P} 1 \quad \text{as} \quad n, d \to \infty,$$

where $\xrightarrow{P}$ refers to convergence in probability, then $\widehat{\boldsymbol{x}}$ is HDLSS consistent with $\boldsymbol{x}$.

In order to evaluate the asymptotic behavior of $\widehat{\boldsymbol{p}}_\alpha$, we need the following conditions:

CONDITION A. Let $X = [\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{Kn_K}]$ be a data matrix from K classes. Each column of X can be written as

$$\boldsymbol{X}_{\ell i} = \boldsymbol{\mu}_\ell + \boldsymbol{\varepsilon}_{\ell i}, \tag{9}$$

where $\boldsymbol{\varepsilon}_{\ell i}$ $(\ell = 1, \ldots, K;\ i = 1, \ldots, n_\ell)$ are i.i.d. copies of an underlying random vector $\boldsymbol{\varepsilon}$ with mean 0 and covariance matrix Σ.

CONDITION B. (CRAMÉR'S CONDITION) There exist constants ν_1, ν_2, M_1 and M_2 such that each component of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^T$ satisfies

$$E[|\varepsilon_j|^m] \le m!M_1^{m-2}\nu_1/2 \quad \text{and} \quad E[|\varepsilon_j^2 - \sigma_{jj}|^m] \le m!M_2^{m-2}\nu_2/2 \quad \text{for all } m \in \mathbb{N}.$$

CONDITION C. Let $n_\ell \to \infty (\ell = 1, ..., K)$, $d \to \infty$, $\log d = o(n)$, $n = o(d)$. There exists a positive sequence C_d depending only on the dimension d such that $d/(nC_d) \to \xi$, where $\xi \geq 0$.

CONDITION D. All eigenvalues λ_α^* of $C^T C$ are simple (so $\lambda_1^* > \cdots > \lambda_K^*$) and satisfy

$$\lambda_\alpha^* = O(C_d) \quad \text{and} \quad \frac{\lambda_\alpha^* - \lambda_{\alpha+1}^*}{C_d} > \xi \quad \text{for} \quad \alpha = 1, \ldots, K-1,$$

and C_d as in condition C.

CONDITION E. As $d \to \infty$, $\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_\ell = O(C_d)$, and there exists $\delta \in (0, 1)$ such that $\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_k = O(C_d^\delta)$ for all $k, \ell \in \{1, \ldots, K\}$, and C_d as in condition C.

CONDITION F. For $k, \ell \in \{1, \ldots, K\}$, and $k \neq \ell$,

$$\lim_{d\to\infty} \sqrt{\pi_\ell} \frac{\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_\ell}{C_d} \neq \lim_{d\to\infty} \sqrt{\pi_k} \frac{\boldsymbol{\mu}_k^T D^{-1} \boldsymbol{\mu}_k}{C_d},$$

and C_d as in condition C.

We note that the positive sequence C_d will play an important role in the subsequent discussions since it controls the gap between d and n. To establish the asymptotic behavior of the estimators $\widehat{\boldsymbol{p}}_\alpha$ and $\widehat{\boldsymbol{b}}_\alpha$, we start with an asymptotic expansion of the matrix $\widehat{C}^T \widehat{C}/C_d$. In what follows, Θ denotes the parameter space for our multi-class setting:

$$\Theta = \left\{ (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma) \;\middle|\; \begin{array}{l} \min_{k\neq\ell} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)^T D^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \geq C_d, \\ \lambda_{\max}(R) \leq b_0, \quad \min_{1\leq j\leq d} \sigma_{jj} > 0 \end{array} \right\}, \tag{10}$$

where R is the correlation matrix $R = D^{-1/2} \Sigma D^{-1/2}$, $\lambda_{\max}(R)$ is the largest eigenvalue of R and σ_{jj} is jth diagonal entry of Σ.

LEMMA 3.2. *Suppose that conditions A – E hold. Then for all parameters $\theta \in \Theta$, $\widehat{C}^T \widehat{C}/C_d$ can be expanded as*

$$\frac{\widehat{C}^T \widehat{C}}{C_d} = \frac{C^T C}{C_d} + \xi (I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2}) + \mathbf{1}_K \mathbf{1}_K^T o_P(1),$$

where Π is the diagonal matrix given in (1).

If $\xi = 0$, which means $d = o(nC_d)$ by condition C, then the above expansion becomes very simple.

To proceed with the theoretical considerations, we need to show that the $K-1$ eigenvalues of $C^T C/C_d + \xi(I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2})$ are simple in the case of $d = O(nC_d)$, that is, $\xi > 0$.

LEMMA 3.3. *Let λ_α/C_d be αth largest eigenvalue of $C^T C/C_d + \xi(I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2})$. Suppose that the λ_α/C_d satisfy condition D, and that $d = O(nC_d)$. Then for all parameters $\theta \in \Theta$,*

$$\lambda_\alpha = O(C_d) \quad \textit{and} \quad \frac{\lambda_\alpha}{C_d} > \frac{\lambda_{\alpha+1}}{C_d} \quad \textit{for} \quad \alpha = 1, \ldots, K-1.$$

In this paper, eigenvectors have unit length. In addition, we assume that the first entry of each eigenvector is positive. This assumption avoids any ambiguity about the direction of the eigenvector. Using Lemmas 3.2 and 3.3, we can now describe the asymptotic behavior of $\widehat{\boldsymbol{p}}_\alpha$ as follows.

THEOREM 3.4. *Suppose that conditions A – D hold. Then, for all parameters* $\theta \in \Theta$,

$$\widehat{\boldsymbol{p}}_\alpha^T \frac{\widehat{C}\boldsymbol{\gamma}_\alpha}{||\widehat{C}\boldsymbol{\gamma}_\alpha||} = 1 + o_P(1) \quad \textit{for} \quad \alpha = 1, \ldots, K-1,$$

where $\boldsymbol{\gamma}_\alpha$ *is eigenvector of* $C^T C/C_d + \xi(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$ *belonging to the non-zero eigenvalue* λ_α.

Proofs of Lemmas 3.2, 3.3 and Theorem 3.4 are given in the Appendix. To gain further insight in the behavior of the $\widehat{\boldsymbol{p}}_\alpha$, we define the vectors

$$\boldsymbol{p}_\alpha = \frac{C\boldsymbol{\gamma}_\alpha}{||C\boldsymbol{\gamma}_\alpha||} \quad \text{for} \quad \alpha = 1, \ldots, K-1 \tag{11}$$

by referring to Theorem 3.4. Under the assumption that $d = O(nC_d)$, one can show that the vectors $\widehat{\boldsymbol{p}}_\alpha$ are consistent estimators for the $\boldsymbol{p}_\alpha$. We have the following theorem and corollary:

THEOREM 3.5. *Suppose that conditions A – E hold, and that* $d = O(nC_d)$. *Moreover, assume that* $\lambda_\alpha/C_d \to \kappa_\alpha$ *and* $\boldsymbol{\gamma}_\alpha^T\Pi^{1/2}\mathbf{1}_K \to \eta_\alpha$. *Then for all parameters* $\theta \in \Theta$,

$$\widehat{P}^T P \xrightarrow{P} \left(\frac{\kappa_\beta\delta_{\alpha\beta} - \xi(\delta_{\alpha\beta} - \eta_\alpha\eta_\beta)}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta - \xi(1-\eta_\beta^2)}} \right)_{1 \le \alpha,\beta \le K-1},$$

where $\delta_{\alpha\beta}$ *is the Kronecker delta-function.*

COROLLARY 3.6. *Suppose that conditions A – E hold, and that* $d = o(nC_d)$. *Then for all parameters* $\theta \in \Theta$,

$$\widehat{P}^T P \xrightarrow{P} I_{K-1}.$$

Theorem 3.5 and Corollary 3.6 are extension of Theorems 1 and 2 in Tamatani, Koch and Naito (2012). Corollary 3.6 states that if $d = o(nC_d)$ is satisfied, then $\widehat{\boldsymbol{p}}_\alpha$ is HDLSS consistent with $\boldsymbol{p}_\alpha$. Furthermore, $\widehat{\boldsymbol{p}}_\alpha$ is asymptotically orthogonal to $\boldsymbol{p}_\beta$ for $\alpha \neq \beta$. On the other hand, if $d = O(nC_d)$ is satisfied, then the angle between $\widehat{\boldsymbol{p}}_\alpha$ and $\boldsymbol{p}_\beta$ converges to a particular non-zero angle for all α and β.

Next we turn to the asymptotic behavior of the vectors $\widehat{\boldsymbol{b}}_\alpha$. We define normalized versions of direction vectors for discrimination by

$$\widehat{\boldsymbol{b}}_\alpha^* = \frac{\widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_\alpha}{\sqrt{\widehat{\boldsymbol{p}}_\alpha^T\widehat{D}^{-1}\widehat{\boldsymbol{p}}_\alpha}} \quad \text{and} \quad \boldsymbol{b}_\alpha^* = \frac{D^{-1/2}\boldsymbol{p}_\alpha}{\sqrt{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\alpha}}. \tag{12}$$

Then we have the following theorem and corollary:

THEOREM 3.7. *Suppose that conditions A – E hold, and that* $d = O(nC_d)$. *Put* $\sigma_{\max} = \max_{1\le j\le d}\sigma_{jj}$ *and* $\sigma_{\min} = \min_{1\le j\le d}\sigma_{jj}$. *Then for all parameters* $\theta \in \Theta$,

$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^* \le \frac{\boldsymbol{b}_\alpha^{*T}\boldsymbol{b}_\beta^*\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\min}/\sigma_{\max})}}(1+o_P(1)),$$
$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^* \ge \frac{\boldsymbol{b}_\alpha^{*T}\boldsymbol{b}_\beta^*\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\max}/\sigma_{\min})}}(1+o_P(1)).$$

COROLLARY 3.8. *Let* $\widehat{B}^* = [\widehat{\boldsymbol{b}}_1^*, \ldots, \widehat{\boldsymbol{b}}_{K-1}^*]$ *and* $B^* = [\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{K-1}^*]$. *Suppose that conditions A – E hold, and that* $d = o(nC_d)$. *Then for all parameters* $\theta \in \Theta$,

$$\widehat{B}^{*T}B^* - B^{*T}B^* \quad \xrightarrow{P} \quad O.$$

Theorem 3.7 states that the upper and lower bounds of $\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^*$ are determined by the ratio of $\sigma_{\max}$ and $\sigma_{\min}$. For example, if all diagonal elements of Σ are equal, then

$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\alpha^* - \boldsymbol{b}_\alpha^{*T}\boldsymbol{b}_\alpha^*\sqrt{1-\frac{\xi}{\kappa_\alpha}(1-\eta_\alpha^2)} \quad \xrightarrow{P} \quad 0. \tag{13}$$

If $d = o(nC_d)$ is satisfied, then the angle between $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\alpha^*$ converges to 0 in probability, which shows that the $\widehat{\boldsymbol{b}}_\alpha^*$ are HDLSS consistent with the corresponding $\boldsymbol{b}_\alpha^*$. However, $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\beta^*$ may not necessarily be orthogonal for $\alpha \ne \beta$, since

$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^* - \frac{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\beta}{\sqrt{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\alpha}\sqrt{\boldsymbol{p}_\beta^T D^{-1}\boldsymbol{p}_\beta}} = o_P(1).$$

4. Asymptotic Misclassification Rate

In this section, we study the misclassification rate of our method in a multi-class setting. The misclassification rate for two classes has been investigated in Fan and Fan (2008) who derived an upper bound for the misclassification rate in a HDLSS setting. A related discussion can be found in Tamatani, Koch and Naito (2012).

Specifically, for the results in this section we will assume that condition A1 holds.

CONDITION A1. Let $X = [\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{Kn_K}]$ be a data matrix from K classes. Each column of X can be written as

$$\boldsymbol{X}_{\ell i} = \boldsymbol{\mu}_\ell + \boldsymbol{\varepsilon}_{\ell i},$$

where $\boldsymbol{\varepsilon}_{\ell i}$ are independently and identically distributed as $N(\boldsymbol{0}, \Sigma)$.

In addition to condition A, condition A1 makes statements about the distribution of X.

Suppose that $\boldsymbol{X}$ belongs to class $\mathcal{C}_k$. The misclassification rate of $\widehat{g}$, an estimate of the discriminant function g in (5), for class $\mathcal{C}_k$ is defined as

$$\begin{aligned} W_k(\widehat{g},\theta) &= P\left(\widehat{g}(\boldsymbol{X}) \ne k \,|\, \boldsymbol{X}_{\ell i}, \ell = 1,\ldots,K, i = 1,\ldots,n_\ell\right) \\ &= 1 - \int_{\widehat{\mathscr{D}}_k} \frac{1}{\sqrt{|2\pi\widehat{\Sigma}_k|}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\widehat{\Sigma}_k^{-1}\boldsymbol{z}\right) d\boldsymbol{z} \\ &\equiv 1 - \Phi_{K-1}\left(\widehat{\mathscr{D}}_k; \boldsymbol{0}, \widehat{\Sigma}_k\right), \end{aligned}$$

where $\widehat{\Sigma}_k$ is the transformed covariance matrix of size $(K-1)\times(K-1)$ which is defined in (26) in the Appendix, and $\widehat{\mathscr{D}}_k$ is the $(K-1)$-dimensional region given by

$$\widehat{\mathscr{D}}_k = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \,\middle|\, z_\alpha < \widehat{\underline{d}}_{k\alpha},\ \alpha = 1,\dots,K-1 \right\},$$

where

$$\begin{aligned}
\widehat{\underline{d}}_{k\alpha} &= I(\alpha<k)\widehat{d}_{k\alpha} + I(\alpha \geq k)\widehat{d}_{k(\alpha+1)} \quad \text{and} \\
\widehat{d}_{k\alpha} &= \frac{(\boldsymbol{\mu}_k - (\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)/2)^T\, \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}{\sqrt{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T\widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T\Sigma\widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}}.
\end{aligned}$$

We note that the region $\widehat{\mathscr{D}}_1$ results in the interval obtained in Theorem 1 in Fan and Fan (2008) for their special case of $K=2$.

Let Θ_k be the parameter space associated with the misclassification rate of $\widehat{g}$ for class $\mathcal{C}_k$:

$$\Theta_k = \left\{ (\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_K,\Sigma) \;\middle|\; \begin{array}{l} \min_{\ell\neq k}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k) \geq C_d, \\ \lambda_{\max}(R) \leq b_0, \min_{1\leq j\leq d}\sigma_{jj} > 0 \end{array} \right\}.$$

In addition to the region $\widehat{\mathscr{D}}_k$ we also require the following region and quantities in Theorem 4.1:

$$\mathscr{D}_{k,O} = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \mid z_\alpha < \underline{d}_{k\alpha}(1+o_P(1)),\ \alpha = 1,\dots,K-1 \right\},$$

where $\underline{d}_{k\alpha} = I(\alpha<k)d_{k\alpha} + I(\alpha\geq k)d_{k(\alpha+1)}$,

$$d_{k\alpha} = \frac{S_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T}{\sqrt{\lambda_{\max}(R)}\sqrt{Q_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T}},$$

$\Gamma = [\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_{K-1}]$, $S_{k\alpha} = M_{k\alpha}/2 + (d/n)\boldsymbol{s}_{k\alpha}\Pi^{-1/2}$, $Q_{k\alpha} = M_{k\alpha} + (d/n)\boldsymbol{q}_{k\alpha}\Pi^{-1/2}$,

$$\begin{aligned}
M_{k\alpha} &= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1/2}C, \\
\boldsymbol{s}_{k\alpha} &= [s_1,\dots,s_K], \quad s_\ell = \pi_\ell - \frac{1}{2}\left\{I(\ell=k) + I(\ell=\alpha)\right\}, \\
\boldsymbol{q}_{k\alpha} &= [q_1,\dots,q_K], \quad q_\ell = I(\ell=k) - I(\ell=\alpha).
\end{aligned}$$

Furthermore, we need the following region in Corollary 4.2:

$$\mathscr{D}_{k,o} = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \mid z_\alpha < \underline{d}^*_{k\alpha}(1+o_P(1)),\ \alpha = 1,\dots,K-1 \right\},$$

where $\underline{d}^*_{k\alpha} = I(\alpha<k)d^*_{k\alpha} + I(\alpha\geq k)d^*_{k(\alpha+1)}$ and

$$d^*_{k\alpha} = \frac{\sqrt{M_{k\alpha}\Gamma\left(\Gamma^TC^TC\Gamma\right)^{-1}\Gamma^TM_{k\alpha}^T}}{2\sqrt{\lambda_{\max}(R)}}.$$

We have the following theorem and corollary using Theorem 3.4.

THEOREM 4.1. *Suppose that conditions A1 and B – F hold, and that* $d = O(nC_d)$. *Then, for all parameters* $\theta \in \Theta_k$,

$$W_k(\widehat{g}, \theta) \leq 1 - \Phi_{K-1}\left(\mathscr{D}_{k,O}; \mathbf{0}, \widehat{\Sigma}_k\right).$$

COROLLARY 4.2. *Suppose that conditions A1 and B – F hold, and that* $d = o(nC_d)$. *Then, for all parameters* $\theta \in \Theta_k$,

$$W_k(\widehat{g}, \theta) \leq 1 - \Phi_{K-1}\left(\mathscr{D}_{k,o}; \mathbf{0}, \widehat{\Sigma}_k\right).$$

Note that Theorem 4.1 and Corollary 4.2 extend Theorem 1 in Fan and Fan (2008) to the general multi-class setting considered in this paper.

To appreciate the usefulness of the upper bounds which we derived in Theorem 4.1 and Corollary 4.2, we give a numerical example for the case $K = 3$ with parameters $d = 1200$, $\boldsymbol{\mu}_\ell = \ell(\mathbf{1}_d - \boldsymbol{e}_\ell)/n^s$ $(\ell = 1, 2, 3)$ and $\Sigma = \left(0.6^{|i-j|}\right)_{1 \leq i,j \leq d}$. We see that $d = O(nC_d)$ if $s = 1/2$, and $d = o(nC_d)$ if $s < 1/2$ (see condition E). The empirical regions $\widehat{\mathscr{D}}_{1,O}$ or $\widehat{\mathscr{D}}_{1,o}$ are obtained from the sample mean over 100 iterations.

Figure 1 shows estimates of the misclassification rate $W_1(\widehat{g}, \theta)$ and of the upper bounds given in Theorem 4.1 and Corollary 4.2. Both figure panels show that the upper bounds are actually upper bounds of the error rates even for the moderate sample sizes used in the example. The left panel refers to the case $d = o(nC_d)$, and the right panel covers the case $d = O(nC_d)$. Note that the difference between the upper bound and $W_1(\widehat{g}, \theta)$ is smaller for the regime $d = O(nC_d)$ than the case $d = o(nC_d)$ as sample size increases. In both cases, the estimated misclassification rate increases with n, whereas the upper bounds tend to flatten as n approaches 500.

5. Feature Selection for the Multi-class Setting

In this section, we propose a method for feature selection in a HDLSS multi-class setting which accompanies our discriminant function. For two-class problems a number of algorithms exist for extracting and ranking salient features, including the features annealed independence rules (FAIR) of Fan and Fan (2008), which are based on two-sample t-statistics. Tamatani, Koch and Naito (2012) showed that FAIR is essentially equivalent to variable ranking based on the absolute value of the components of $\widehat{\boldsymbol{p}}_1$. Furthermore, Tamatani, Koch and Naito (2012) proposed the naive canonical correlation (NACC) approach for feature selection, which exploits the first canonical correlation vector $\widehat{\boldsymbol{b}}_1$. Feature selection algorithms for more than two classes have been discussed in the machine learning and bioinformatics literature; see, for example, the comprehensive survey by Saeys et al. (2007).

Feature selection algorithms generally consist of two steps:

Step A Variable Ranking: Using some reference vectors including information about features(variables), we make a ranking vector $\widehat{\boldsymbol{c}}$, of which components satisfy

$$\hat{c}_{j_1} \geq \cdots \geq \hat{c}_{j_m} \geq \cdots \geq \hat{c}_{j_d} \geq 0. \tag{14}$$

A variable ranking scheme for a vector $\boldsymbol{X}$ (or each column of the data matrix X) is a permutation of the variables of $\boldsymbol{X}$ according to the order inherited from $\widehat{\boldsymbol{c}}$.

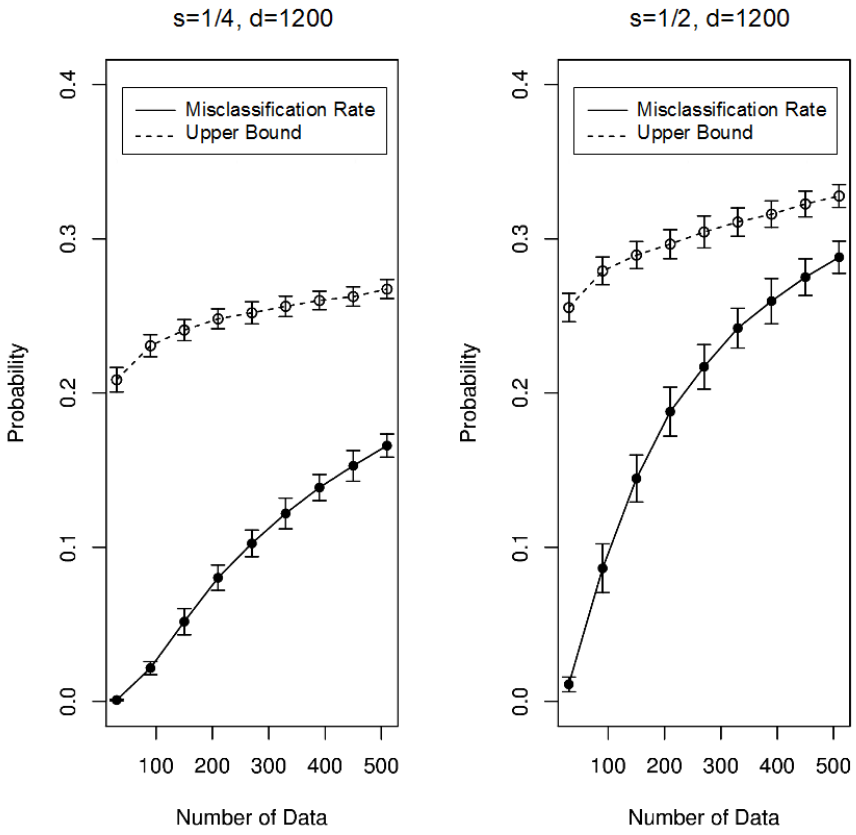

Figure 1: Mean upper bound for the misclassification rates over 100 simulations shown together with intervals of one standard deviation. Left panel: $d = o(nC_d)$; right panel: $d = O(nC_d)$.

Step B Number of Discriminant Features: Once the entries of $\boldsymbol{X}$ (or each column of the data matrix X) have been ranked, we determine the number of effective features, $\hat{m}$ and then only use the features $X_{j_1}, ..., X_{j_{\hat{m}}}$.

FAIR uses $\widehat{\boldsymbol{p}}_1$ as the reference vector. NACC works in a similar way, but it is based on $\widehat{\boldsymbol{b}}_1$ instead of $\widehat{\boldsymbol{p}}_1$. Both rules minimize an upper bound for the misclassification rate as the stopping criterion in Step B. Note that $\widehat{\boldsymbol{p}}_1$, and respectively $\widehat{\boldsymbol{b}}_1$, is the eigenvector belonging to the unique non-zero eigenvalue for $K = 2$.

We aim to extend FAIR and NACC to the multi-class setting. A natural extension is the use of all $K-1$ vectors $\widehat{\boldsymbol{p}}_1, ..., \widehat{\boldsymbol{p}}_{K-1}$ or $\widehat{\boldsymbol{b}}_1, ..., \widehat{\boldsymbol{b}}_{K-1}$ as reference vectors. Hence we need a rule for combining the $K-1$ vectors. We integrate the $K-1$ reference vectors into a component-wise 'best' ranking vector whose entries are chosen as described in Table 1. The ranking of the variables is then inherited from the ranking vector obtained by combining these reference vectors.

Table 1: The ranking vector $\widehat{\boldsymbol{c}}$ associated with NACC and FAIR for K classes.

Name	Component of $\widehat{\boldsymbol{c}}$
M-NACC	$\hat{c}_j = \max\limits_{1\le\alpha\le K-1} \lvert\hat{b}_{\alpha j}\rvert,\ j = 1, ..., d$
M-FAIR	$\hat{c}_j = \max\limits_{1\le\alpha\le K-1} \lvert\hat{p}_{\alpha j}\rvert,\ j = 1, ..., d$

The criterion for selecting the number of discriminant features is similar to that in Fan and Fan (2008): We choose the number $\hat{m}$ which minimizes the upper bound of the misclassification rate given in Theorem 4.1. Suppose that the rows of the data matrix X are sorted according to some $\widehat{\boldsymbol{c}}$, and then X is truncated into the upper $\hat{m} \times n$ matrix. The discrimination rule (8) with feature selection becomes

$$\widehat{g}_{\hat{m}}(\boldsymbol{X}) = \underset{\ell\in\{1,...,K\}}{\arg\min} \left(\boldsymbol{X}_{\hat{m}} - \widehat{\boldsymbol{\mu}}_{\ell,\hat{m}}\right)^T \widehat{B}_{\hat{m}} (\widehat{B}_{\hat{m}} \widehat{D}_{\hat{m}} \widehat{B}_{\hat{m}}^T)^{-1} \widehat{B}_{\hat{m}}^T \left(\boldsymbol{X}_{\hat{m}} - \widehat{\boldsymbol{\mu}}_{\ell,\hat{m}}\right), \tag{15}$$

where $\boldsymbol{X}_{\hat{m}}$, $\widehat{\boldsymbol{\mu}}_{\ell,\hat{m}}$ and each column of $\widehat{B}_{\hat{m}} = \left[\widehat{\boldsymbol{b}}_{1,\hat{m}}, \ldots, \widehat{\boldsymbol{b}}_{K-1,\hat{m}}\right]$ are the corresponding first $\hat{m}$-dimensional subvectors and

$$\widehat{D}_{\hat{m}} = \mathrm{diag}(\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{\hat{m}\hat{m}})$$

is the $\hat{m} \times \hat{m}$ left-upper submatrix of $\widehat{D}$. We summarize the new classification method based on feature selection in the following steps:

Feature Selection Algorithm M-NACC (M-FAIR)

Step 1 Calculate $\widehat{P}$ and $\widehat{B}$.

Step 2 Sort the components of $\widehat{\boldsymbol{c}}$ in descending order of their absolute values as

$$\hat{c}_{j_1} \ge \hat{c}_{j_2} \ge \cdots \ge \hat{c}_{j_m} \ge \cdots \ge \hat{c}_{j_d} \ge 0,$$

where $\hat{c}_j$ is based on $\widehat{\boldsymbol{b}}_\ell (\ell = 1, ..., K-1)$ for M-NACC, and on $\widehat{\boldsymbol{p}}_\ell (\ell = 1, ..., K-1)$ for M-FAIR (see Table 1).

Step 3 Apply the permutation $\tau : \{1, 2, \ldots, d\} \to \{j_1, j_2, \ldots, j_d\}$ to the rows of the data X, and put $\boldsymbol{X}_i \leftarrow \tau(\boldsymbol{X}_i)$.

Step 4 Let $[\,\cdot\,]_m$ be the vector or matrix calculated by using the first m rows of the matrix X made in Step 3. Find the best truncation $\hat{m}$ based on Theorem 4.1:

$$\hat{m} = \underset{K-1 \le m \le d}{\arg\max} \sum_{k \neq \alpha} \frac{n_k n_\alpha}{n^2} \hat{d}_{k\alpha}(m), \tag{16}$$

where $\hat{d}_{k\alpha}(m)$ is calculated by

$$\begin{aligned}
\hat{d}_{k\alpha}(m) &= \frac{\widehat{S}_{k\alpha,m} \widehat{\Gamma} \left(\widehat{\Gamma}^T \widehat{U}_m \widehat{\Gamma} \right)^{-1} \widehat{\Gamma}^T \widehat{Q}_{k\alpha,m}^T}{\sqrt{\lambda_{\max}(\widehat{R}_m)} \sqrt{\widehat{Q}_{k\alpha,m} \widehat{\Gamma} \left(\widehat{\Gamma}^T \widehat{U}_m \widehat{\Gamma} \right)^{-1} \widehat{\Gamma}^T \widehat{Q}_{k\alpha,m}^T}}, \\
\widehat{U}_m &= \left[\widehat{C}^T \widehat{C} \right]_m + \frac{m}{n} (I_K - N^{1/2} \mathbf{1}_K \mathbf{1}_K^T N^{1/2}), \\
\widehat{Q}_{k\alpha,m} &= \left[\widehat{M}_{k\alpha} \right]_m + \frac{m}{n} \boldsymbol{q}_{k\alpha} N^{-1/2}, \\
\widehat{S}_{k\alpha,m} &= \frac{1}{2} \left[\widehat{M}_{k\alpha} \right]_m + \frac{m}{n} \widehat{\boldsymbol{s}}_{k\alpha} N^{-1/2}, \\
\widehat{\boldsymbol{s}}_{k\alpha} &= (\hat{s}_1, \ldots, \hat{s}_K), \qquad \hat{s}_\ell = \frac{n_\ell}{n} - \frac{1}{2} \{ I(\ell = k) + I(\ell = \alpha) \}
\end{aligned}$$

and $\widehat{R}_m = (\hat{\rho}_{ij})_{1 \le i,j \le m}$ is the appropriate $m \times m$ correlation submatrix of $\widehat{R}$.

Step 5 Apply the same permutation τ as in Step 3 to a new datum $\boldsymbol{X} \leftarrow \tau(\boldsymbol{X})$, use the first $\hat{m}$ entries of $\boldsymbol{X}$ and apply rule (15) to assign $\boldsymbol{X}$ to one of the K classes.

Note that if $\hat{m} = d$, then (15) is nothing other than (8) without feature selection. We will refer to this method Multi-class Diagonal Linear Discriminant Analysis (M-DLDA). We investigate the performance of M-NACC, M-FAIR and M-DLDA for real and simulated data in the next section.

6. Numerical Studies

In this section, we apply the proposed discriminant function to real and simulated data, and compare its performance on simulated data with the theoretical results developed in the previous sections. Further, we compare the performance of our feature selection with the Minimum Redundancy – Maximum Relevance (MRMR) proposed by Ding and Peng (2005), which maximizes an F-statistic or mutual information while minimizing a redundancy criterion which is based on the correlation coefficient. It is worth noting that MRMR is a variable ranking method which does not include any determination of the number of ranked features as in Step B of Section 5.. The ranking approach MRMR does not exclude the integration of criteria for determining the number of features. Indeed, in the simulations below we show how one can implement the criterion of (16) in MRMR, as well as in our new method.

6.1. Simulation I

We start with an investigation of the angle between $\widehat{\boldsymbol{b}}^*_\alpha$ and $\boldsymbol{b}^*_\beta$ for simulated data from the Gaussian distribution. In Simulation I, we generate n_ℓ d-dimensional observations $\boldsymbol{X}_{\ell i} \overset{\text{i.i.d.}}{\sim} N_d(\boldsymbol{\mu}_\ell, \Sigma)$ for $d = 600$ and 1200 from $K = 3$ classes. For each d, we choose $n = \sum_{\ell=1}^{K} n_\ell$ such that $n \leq d$. For the simulated data we calculate $\widehat{\boldsymbol{b}}^*_1, \ldots, \widehat{\boldsymbol{b}}^*_{K-1}$, and the angles $\angle(\widehat{\boldsymbol{b}}^*_\alpha, \boldsymbol{b}^*_\beta) = (180/\pi)\arccos(\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\beta)$ (for $\alpha, \beta = 1, \ldots, K-1$), and display the distributions of these angles based on 1000 iterations.

We take $\Sigma = (\sigma_{ij})$ to be the covariance of an AR structure, with $\sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq d$. For this example we take $\rho = 0.6$. In order to avoid CC^T being singular, we take

$$[\boldsymbol{\mu}_1 \;\; \boldsymbol{\mu}_2 \;\; \boldsymbol{\mu}_3] = \frac{4}{\sqrt{n}} \begin{bmatrix} \mathbf{1}_{d/2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{d/4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{d/4} \end{bmatrix}.$$

In this setting, it is easily confirmed that

$$\min_{i \neq j} \frac{n(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T D^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{d} \longrightarrow 2.$$

Thus, if we take $C_d = d^\gamma$ for the sequence in condition C such that $d = nd^\gamma$ for $\gamma \in (0, 1)$, then the condition $d = O(nC_d)$ in Theorem 3.7 is satisfied. We obtain the following angles in degrees $\angle(\widehat{\boldsymbol{b}}^*_1, \boldsymbol{b}^*_1) \approx 33.855$, $\angle(\widehat{\boldsymbol{b}}^*_2, \boldsymbol{b}^*_2) \approx 40.893$ and $\angle(\widehat{\boldsymbol{b}}^*_1, \boldsymbol{b}^*_2)$, $\angle(\widehat{\boldsymbol{b}}^*_2, \boldsymbol{b}^*_1) \approx 90$ from (13) with $\sigma_{\max} = \sigma_{\min} = 1$.

Table 2 summarizes the results for Simulation I. The table shows that each angle approaches the theoretical value shown in Theorem 3.7. Figure 2 depicts the behavior of $\angle(\widehat{\boldsymbol{b}}^*_1, \boldsymbol{b}^*_1)$ and $\angle(\widehat{\boldsymbol{b}}^*_2, \boldsymbol{b}^*_2)$ using kernel density estimates. The top panel of Figure 2 shows the results for $\angle(\widehat{\boldsymbol{b}}^*_1, \boldsymbol{b}^*_1)$, and the bottom panel shows the results for $\angle(\widehat{\boldsymbol{b}}^*_2, \boldsymbol{b}^*_2)$ for $n = 15,\ 30,\ 150,\ 600$ and $d = 1200$. These figures show that each angle converges as the sample size increases.

6.2. Simulation II

Simulation II focuses on the performance of the proposed discriminant function with feature selection. In Simulation II, we consider data $\boldsymbol{X}_{\ell i} \overset{\text{i.i.d.}}{\sim} N_d(\boldsymbol{\mu}_\ell, \Sigma_0)$ from $K = 10$ classes with:

$$\boldsymbol{\mu}_\ell = (\mu_{\ell 1}, \ldots, \mu_{\ell d})^T \in \mathbb{R}^d,\ \mu_{\ell j} = \begin{cases} 11 - \ell & , \quad j = \ell, \\ 0 & , \quad \text{otherwise}, \end{cases} \quad \text{and } \Sigma_0 = A^{1/2} \Sigma A^{1/2},$$

where $A = \text{diag}(a_{jj})$ with $a_{jj} = j$ and Σ has an AR structure with $\rho = -0.6$. The mean parameters $\boldsymbol{\mu}_\ell$s show that only the first 10 features differ, and we therefore, expect to select these features in the simulated data. Furthermore, we note that the diagonal elements of the covariance matrix Σ_0 are monotonically increasing.

We compare M-NACC with MRMR, and use the same $\hat{m}$ for both methods, which we calculate for M-NACC as described in (16). A similar comparison is given for M-FAIR and MRMR. For $d = 1000$ and $n_\ell = 50$ (with $\ell = 1, ..., 10$), we obtained estimates for the misclassification rate based on 100 iterations.

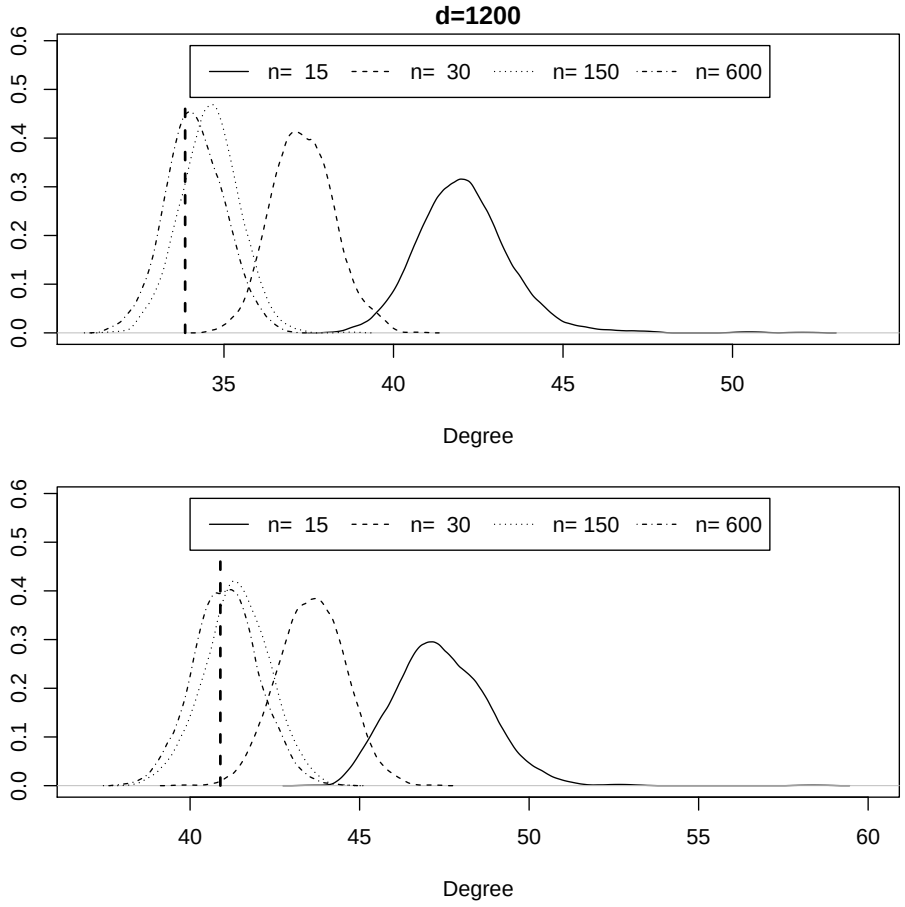

Figure 2: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\alpha^*$. The top panel and bottom panel are for $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$ and $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$, respectively. The vertical dashed line is the asymptotic value derived in Theorem 3.7.

Table 2: Average of 1000 simulated angles between $\widehat{\boldsymbol{b}}_i^*$ and $\boldsymbol{b}_j^*$ in degrees.

	$d = 600$			
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$
15	42.237	47.407	90.327	89.976
30	37.392	43.652	90.243	90.127
90	35.159	41.842	90.265	90.026
150	34.746	41.359	90.288	90.074
300	34.388	41.182	90.277	89.959
450	34.257	41.181	89.990	89.881
	$d = 1200$			
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$
15	42.084	47.428	90.273	89.995
30	37.328	43.590	90.302	89.992
90	35.042	41.671	90.046	90.071
150	34.592	41.385	90.073	90.061
300	34.261	41.146	90.196	90.058
450	34.203	41.075	90.164	90.059
600	34.147	41.062	90.120	89.946
900	34.078	41.068	90.021	90.041

In this simulation study we introduce another measure for comparing feature selection methods. We can see from definition of the $\boldsymbol{\mu}_\ell$ that the set of features $\{1, ..., 10\}$ should be picked with a suitable feature selection. We now want to determine how successful the features selection (16) is in choosing the correct features. Let $\{j_1^{(t)}, ..., j_{\hat{m}(t)}^{(t)}\}$ be a set of selected features based on the tth simulated data set, where $\hat{m}(t)$ is the number of selected features determined by the tth data set. For $i \in \{1, ..., 10\}$, let CSR(i) be the *correct selection rate* for feature i and put

$$\mathrm{CSR}(i) = \frac{1}{100} \sum_{t=1}^{100} I\left(i \in \{j_1^{(t)}, ..., j_{\hat{m}(t)}^{(t)}\}\right),$$

where $I(A)$ is the indicator function of a set A. The CSR counts how often the set of selected features contains the features which should be selected. A CSR(i) close to 1 means that feature i is frequently selected by the feature selection method which is good. We calculated CSR(i) for M-NACC, M-FAIR and MRMR in this study. For the MRMR-based discriminant function with feature selection we did not calculate estimates of the misclassification rate, since MRMR is just a competitor for variable ranking.

The results of the simulations are given in Table 3, which shows the correct selection rate as well as the estimated misclassification rate for each method. The estimate for the error based on (15), and called training error here, was obtained as the sample mean over 100 iterations, while the estimate for the test error was calculated as the average of the leave-one out cross-validation over 100 iterations. Table 3 shows that the CSR(i)-values of M-NACC are typically higher than those of MRMR, and a similar tendency can be

verified for M-FAIR. M-NACC recorded CSR(i) < 1 for $i = 9, 10$, while M-FAIR resulted in CSR(i) < 1 for $i = 8, 9, 10$. Further CSR(9) and CSR(10) of M-FAIR are smaller than those of M-NACC. As far as the mean number of selected features $\hat{m}$ is concerned, we note that M-FAIR has a tendency to select more (and indeed by far too many) features than M-NACC, and M-NACC selects the correct features more frequently than M-FAIR.

In view of training error, M-FAIR is slightly superior to M-NACC, which can be understood by its number of selected features. However it is worth noting that M-NACC recorded a smaller test error than M-FAIR and used fewer features.

Table 3: Results for Simulation II. CSR(i) and estimates for the misclassification rate of each method. [1]: MRMR with $\hat{m}$ determined by M-FAIR. [2]: MRMR with $\hat{m}$ determined by M-NACC.

Feature(i)	1	2	3	4	5	6	7	8	9	10
M-FAIR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.47	0.03
MRMR[1]	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.68	0.20	0.02
M-NACC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.14
MRMR[2]	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.42	0.08	0.00

	M-FAIR	M-NACC
No. of selected features	24.31	9.22
Training error (%)	27.46	30.94
Test error (%)	31.84	28.21

6.3. Real Data

6.3.1. SRBCT data

We apply our method to the small round blue-cell tumors (SRBCT) data that were analyzed in Khan et al. (2001), and are available at `http://statweb.stanford.edu/~tibs/ElemStatLearn/` (see Hastie et al. (2001)). The data have four classes: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). There are 2308 genes and a total of 83 samples: 63 training samples (12 NB, 20 RMS, 8 NHL and 23 EWS) and 20 test samples (6 NB, 5 RMS, 3 NHL and 6 EWS).

We applied M-NACC, M-FAIR and MRMR with feature selection to the SRBCT data, and also applied M-DLDA (without feature selection) to the same data. The results for approaches with feature selection, including selected gene numbers, training error and test error are summarized in Table 4. We observe from Table 4 that M-FAIR selected 13 genes which resulted in a training error of 0 and a test error of 2, while M-NACC selected 12 genes which yielded a training error of 0 and a slightly higher test error of 3. 20 genes were selected by MRMR, which yielded a 0 training error and a test error of 2. Although M-FAIR and MRMR resulted in the same test error, it should be noted that M-FAIR achieved this accuracy with a smaller number of features. M-DLDA recorded 1 training error and 5 test errors, both of these are worse than the error rates obtained with M-NACC, M-FAIR and MRMR. This reveals that feature selection works well for this data set, and is superior to similar methods without feature selection.

6.3.2. Isolet data

The last example deals with the Isolet data studied in Weinberger et al. (2006), which are available at `http://archive.ics.uci.edu/ml/datasets/ISOLET`. The data have 26 classes corresponding to the letters of the alphabet. There are 617 genes and and a total of 7797 samples: 6238 training samples (238 samples from class 6, and 240 samples from each of the other classes), and 1559 test samples (59 samples from class 13, and 60 samples from each of the other classes).

As our setting is that of HDLSS, we randomly picked 20 samples form each class of the training data. We applied each of the four methods considered in Section 6.3.1. to the 520 samples of the new training data, and evaluated their performances on the full test data consisting of 1559 samples. We repeated the above procedure 100 times. Boxplots of the test errors and the number of selected features are shown in Figure 3. The left panel of Figure 3 shows the number of selected features. Figure 3 exhibits that the number of selected features of M-FAIR is smaller than that of the other approaches. However, there seems to be an unstable trend in the misclassification rate arising from the M-FAIR calculations, as can be observed in the right panel. This could be a consequence of the small number of selected features. On the other hand, the number of selected features of M-NACC and MRMR result in about the same number, however, the test error of M-NACC is smaller than that for MRMR and M-FAIR. The detailed values of the boxplots are summarized in Table 5. From these values we can see that the average test error of M-NACC is almost equal to that of M-DLDA, but M-NACC obtains the same accuracy with only one-third of the number of features.

Table 4: Results for the SRBCT data. Numbers in bold show common selected genes in each approach.

selected genes	M-FAIR	**1955**	2050	1954	**1194**	1158	174	**1003**
		1389	246	107	1645	951	1980	
	M-NACC	**1955**	481	1158	1954	**1194**	1888	951
		879	**1003**	174	**1389**	246		
		509	107	867	879	1708	**1955**	2050
	MRMR	**1194**	246	742	**1003**	**1389**	819	851
		338	368	1706	1319	2	545	

	M-FAIR	M-NACC	MRMR	M-DLDA
No. of selected genes	13	12	20	2308
Training error	0/63	0/63	0/63	1/63
Test error	2/20	3/20	2/20	5/20

7. Conclusion

In this paper, we proposed a linear discriminant function for a general multi-class setting with K classes in a HDLSS context, and derived asymptotic properties of the sample estimators. Our approach extends linear discrimination based on a suitably adjusted naive canonical correlation matrix from two classes to K classes.

In the asymptotic theory, both the dimension d and the sample size n grow, and provided d does not grow too fast, we show that all $K-1$ eigenvectors and canonical correlation vectors are HDLSS consistent. Under suitable conditions, we are able to

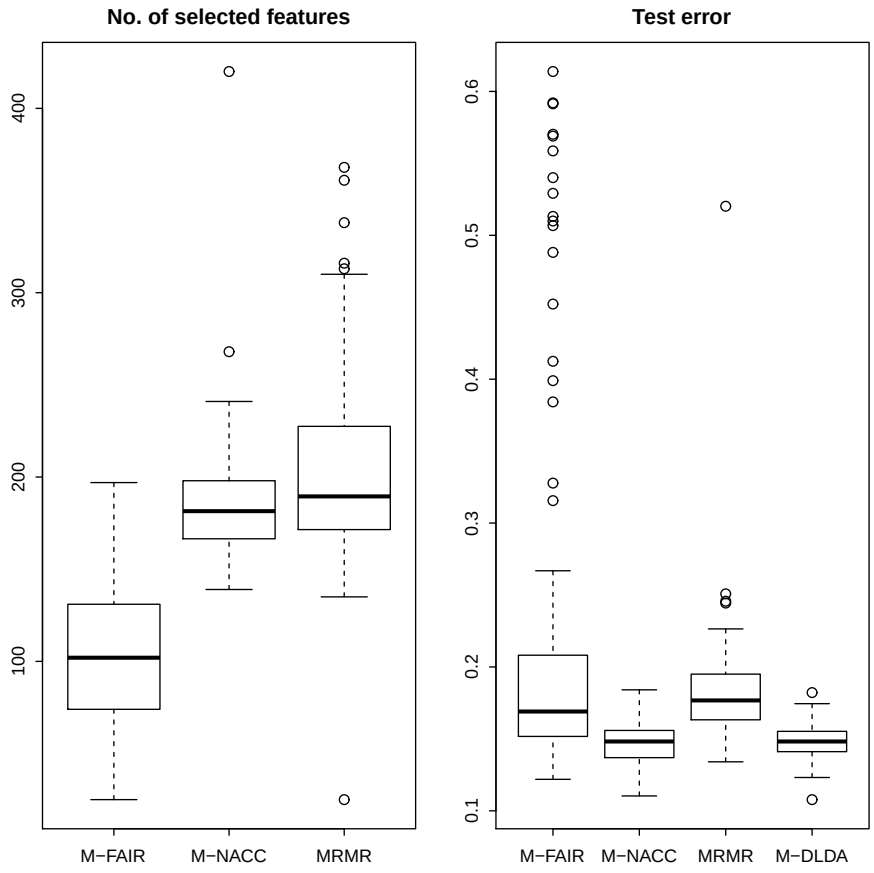

Figure 3: Isolet data. The left panel shows boxplots of the number of selected features for M-NACC, M-FAIR and MRMR over 100 iterations. M-FAIR is smallest, and the number of its outliers (open circles) is zero. The number of outliers for M-NACC is two, and that of MRMR is six. The right panel shows boxplots of test errors of M-NACC, M-FAIR, MRMR and M-DLDA using the simulated data of the left panel. M-FAIR is not as stable as the other approaches because there are many outliers and large errors. M-NACC and M-DLDA are superior to other approaches.

Table 5: Results for the Isolet data. Descriptive statistics of the number of selected features and test error for each method over 100 iterations.

	No. of selected features			
	M-FAIR	M-NACC	MRMR	M-DLDA
SD	43.69	33.38	52.31	0.00
min	25.00	139.00	25.00	617.00
1st quartile	74.00	166.80	171.80	617.00
median	102.00	181.50	189.50	617.00
average	99.33	185.60	205.10	617.00
3rd quartile	130.50	198.00	227.20	617.00
max	197.00	420.00	368.00	617.00
	Test error			
	M-FAIR	M-NACC	MRMR	M-DLDA
SD	0.13361	0.01483	0.04152	0.01242
min	0.12190	0.11030	0.13410	0.10780
1st quartile	0.15190	0.13710	0.16340	0.14110
median	0.16900	0.14820	0.17670	0.14820
average	0.22500	0.14770	0.18200	0.14850
3rd quartile	0.20510	0.15590	0.19500	0.15520
max	0.61390	0.18410	0.52020	0.18220

derive an upper bound for the worst case misclassification rate in the K class setting.

Feature selection has been shown to be essential in correctly classifying HDLSS data arising from two classes. For the general multi-class setting, we propose and discuss two methods for feature selection, M-NACC and M-FAIR, which extend their respective two-class analogues. If the variance is large relative to the difference between the means, we illustrate in Simulation II that M-NACC performs better than M-FAIR. Applications to real data sets demonstrate that M-NACC performs well.

Our approach exploits the naive Bayes rule and replaces $\widehat{\Sigma}^{-1}$ by the diagonal matrix $\widehat{D}^{-1}$. On the other hand, replacing $\widehat{D}^{-1}$ by a certain type of band matrices could also yield efficient linear discriminant functions in a HDLSS setting. Such discriminant functions are of interest in practice, especially when relevant correlation information between the observations is lost in the replacement of $\widehat{\Sigma}^{-1}$ by the diagonal matrix $\widehat{D}^{-1}$. Theoretical research of band matrices has been considered in Bickel and Levina (2008), and their results are expected to apply to linear discriminant function in HDLSS settings.

Appendix

Proof of Lemma 3.2 .

By condition A, for fixed $\ell \leq K$ we have

$$\widehat{\boldsymbol{\mu}}_\ell - \widehat{\boldsymbol{\mu}} \;=\; (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) + (\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}) + \sum_{\ell=1}^{K} \Big(\frac{n_\ell}{n} - \pi_\ell \Big) \boldsymbol{\mu}_\ell,$$

where $\overline{\boldsymbol{\varepsilon}}_\ell = (1/n_\ell)\sum_{i=1}^{n_\ell}\boldsymbol{\varepsilon}_{\ell i}$ and $\overline{\boldsymbol{\varepsilon}} = (1/n)\sum_{\ell=1}^{K}\sum_{i=1}^{n_\ell}\boldsymbol{\varepsilon}_{\ell i}$. Thus,

$$\widehat{M}_0 N^{1/2} - M_0\Pi^{1/2} \;=\; M_0(N^{1/2}-\Pi^{1/2}) + \widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^T N^{1/2} + E_0 N^{1/2},$$

where $\widetilde{M} = [\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K]$ and $E_0 = [\overline{\boldsymbol{\varepsilon}}_1 - \overline{\boldsymbol{\varepsilon}},\ldots,\overline{\boldsymbol{\varepsilon}}_K - \overline{\boldsymbol{\varepsilon}}]$. Therefore, $\widehat{C}^T\widehat{C}$ can be written as

$$\begin{aligned}
\widehat{C}^T\widehat{C} = {} & \left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right) \\
& + \left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}\widetilde{M}\,(N-\Pi)\,\mathbf{1}_K\mathbf{1}_K^T N^{1/2} \\
& + \left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}E_0 N^{1/2} + \left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}M\Pi^{1/2} \\
& + N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right) \\
& + N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^T N^{1/2} \\
& + N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}E_0 N^{1/2} + N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}M_0\Pi^{1/2} \\
& + N^{1/2}E_0^T\widehat{D}^{-1}M_0\left(N^{1/2}\Pi^{1/2}\right) + N^{1/2}E_0^T\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^T N^{1/2} \\
& + N^{1/2}E_0^T\widehat{D}^{-1}E_0 N^{1/2} + N^{1/2}E_0^T\widehat{D}^{-1}M_0\Pi^{1/2} \\
& + \Pi^{1/2}M_0^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right) + \Pi^{1/2}M_0\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^T N^{1/2} \\
& + \Pi^{1/2}M_0^T\widehat{D}^{-1}E_0 N^{1/2} + \Pi^{1/2}M_0\widehat{D}^{-1}M_0\Pi^{1/2}. \qquad (17)
\end{aligned}$$

From condition B, it follows that $\widehat{D} = D(1+o_P(1))$ (see Fan and Fan (2008)), and this leads to the following expressions

$$\begin{aligned}
M_0^T\widehat{D}^{-1}M_0 &= \left((\boldsymbol{\mu}_k-\boldsymbol{\mu})^T D^{-1}(\boldsymbol{\mu}_\ell-\boldsymbol{\mu})\right)_{1\le k,\ell\le K}(1+o_P(1)) = \mathbf{1}_K\mathbf{1}_K^T O(C_d), \\
M_0^T\widehat{D}^{-1}\widetilde{M} &= \left((\boldsymbol{\mu}_k-\boldsymbol{\mu})^T D^{-1}\boldsymbol{\mu}_\ell\right)_{1\le k,\ell\le K}(1+o_P(1)) = \mathbf{1}_K\mathbf{1}_K^T O(C_d), \\
\widetilde{M}^T\widehat{D}^{-1}\widetilde{M} &= \left(\boldsymbol{\mu}_k^T D^{-1}\boldsymbol{\mu}_\ell\right)_{1\le k,\ell\le K}(1+o_P(1)) = \mathbf{1}_K\mathbf{1}_K^T O(C_d^\delta) + I_K O(C_d)
\end{aligned}$$

by condition E. From the evaluation of the term I_3 on p.2626 of Fan and Fan (2008), we have

$$M_0^T\widehat{D}^{-1}E_0 = \mathbf{1}_K\mathbf{1}_K^T o_P(C_d), \quad \widetilde{M}^T\widehat{D}^{-1}E_0 = \mathbf{1}_K\mathbf{1}_K^T o_P(C_d).$$

Consider the matrix $E_0^T\widehat{D}^{-1}E_0$ of $\widehat{C}^T\widehat{C}$. We have

$$E_0^T\widehat{D}^{-1}E_0 \;=\; \left((\overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right)_{1\le k,\ell\le K}(1+o_P(1)).$$

In particular, we need to evaluate the variance term $V\left[(\overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right]$.

If $k=\ell$, this variance can be obtained as

$$\begin{aligned}
& V\left[(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right] \\
& = \mathrm{tr}\left\{(D^{-1}\otimes D^{-1})E\left[(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^T\otimes(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^T\right]\right\} - \left\{\mathrm{tr}(D^{-1}\Sigma^*)\right\}^2 \qquad (18)
\end{aligned}$$

by using Theorem 9.18 of Schott (1996), where $\otimes$ is Kronecker product and $\Sigma^* = V\left[\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}}\right] = (1/n_\ell - 1/n)\,\Sigma$. Thus, we have $\left\{\mathrm{tr}(D^{-1}\Sigma^*)\right\}^2 = d^2(1/n_\ell-1/n)^2$. Since

$D^{-1} \otimes D^{-1}$ is a diagonal matrix, (18) can be written as

$$V\left[(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})\right]$$
$$= \operatorname{tr}\left\{(D^{-1}\otimes D^{-1})E\left[\operatorname{diag}\left\{(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T \otimes (\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T\right\}\right]\right\} - d^2\left(\frac{1}{n_\ell} - \frac{1}{n}\right)^2,$$

by the property of the trace of the relevant matrix. The diagonal elements can be written as

$$D^{-1}\otimes D^{-1}E\left[\operatorname{diag}\left\{(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T \otimes (\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T\right\}\right] = \operatorname{diag}(v_1, \ldots, v_{d^2}), \quad (19)$$

where

$$v_j = \begin{cases} \dfrac{E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^4\right]}{\sigma_{ss}^2}, & \text{for } j = (s-1)d + s, \quad s \in \{1, \ldots, d\}, \\ \dfrac{E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^2 (\overline{\varepsilon}_{\ell t} - \overline{\varepsilon}_t)^2\right]}{\sigma_{ss}\sigma_{tt}}, & \text{for all other values of } j < d^2, \text{ and } s \neq t, \end{cases}$$

and $\overline{\varepsilon}_{\ell s}$ and $\overline{\varepsilon}_s$ are sth element of $\overline{\boldsymbol{\varepsilon}}_\ell$ and $\overline{\boldsymbol{\varepsilon}}$ respectively.

Next, we expand $\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s$. This difference can be written as

$$\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s = \left(\frac{1}{n_\ell} - \frac{1}{n}\right)\sum_{i=1}^{n_\ell} \varepsilon_{\ell i s} - \frac{1}{n}\sum_{k\neq\ell}\sum_{i=1}^{n_k} \varepsilon_{kis}.$$

Using the properties $E[\overline{\varepsilon}_{\ell s}\overline{\varepsilon}_{ks}] = E[\overline{\varepsilon}_{\ell s}]E[\overline{\varepsilon}_{ks}]$ and $E[\overline{\varepsilon}_{\ell s}] = 0$, we have

$$E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^4\right] = \left(\frac{1}{n_\ell} - \frac{1}{n}\right)^4 E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell i s}\right)^4\right] + \frac{1}{n^4}E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^4\right]$$
$$+\frac{6}{n^2}\left(\frac{1}{n_\ell} - \frac{1}{n}\right)^2 E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell i s}\right)^2\right] E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\right].$$

In particular, we find that

$$E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell i s}\right)^2\right] = n_\ell\sigma_{ss}, \quad E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^4\right] = (n-n_\ell)\xi_{ss} + 3\sum_{k\neq\ell} n_k(n_k-1)\sigma_{ss}^2,$$
$$E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\right] = (n-n_\ell)\sigma_{ss}, \quad E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell i s}\right)^4\right] = n_\ell\xi_{ss} + 3n_\ell(n_\ell-1)\sigma_{ss}^2,$$

where $\xi_{st} = E\left[\varepsilon_{11s}^2\varepsilon_{11t}^2\right]$. Therefore, the $((s-1)d+s)$th diagonal element of (19) becomes

$$E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^4\right] = \frac{(n-n_\ell)(3n_\ell^2 - 3nn_\ell + n^2)}{n^3 n_\ell^3}\xi_{ss} + \frac{3}{n^4}\sum_{m\neq\ell} n_m(n_m-1)\sigma_{ss}^2$$
$$+3\left(\frac{1}{n_\ell} - \frac{1}{n}\right)^2 \frac{(n-n_\ell)(n_\ell(n_\ell+1) + n(n_\ell-1))}{n_\ell n^2}\sigma_{ss}^2.$$

From tedious but direct calculations we have

$$E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)\right] = n_\ell\sigma_{st},$$

$$E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)^2\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)^2\right] = n_\ell(n-n_\ell)\sigma_{ss}\sigma_{tt},$$

$$E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)^2\right] = n_\ell(n-n_\ell)\sigma_{ss}\sigma_{tt},$$

$$E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)\right] = n_\ell(n-n_\ell)\sigma_{st}^2,$$

$$E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)^2\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)^2\right] = n_\ell\xi_{st} + n_\ell(n_\ell-1)\sigma_{ss}\sigma_{tt} + \tau_\ell\sigma_{st}^2,$$

$$E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)^2\right] = (n-n_\ell)\xi_{st} + (n-n_\ell)(n-n_\ell-1)\sigma_{ss}\sigma_{tt} + \tau_{-\ell}\sigma_{st}^2,$$

where τ_ℓ and $\tau_{-\ell}$ are the numbers of combinations that arose throughout the calculations, and whose orders are $O(n^2)$. The above expressions lead to

$$\begin{aligned} E\left[\left(\overline{\varepsilon}_{\ell s}-\overline{\varepsilon}_s\right)^2\left(\overline{\varepsilon}_{\ell t}-\overline{\varepsilon}_t\right)^2\right] &= \frac{(n-n_\ell)(3n_\ell^2-3n_\ell n+n^2)}{n^3n_\ell^3}\xi_{st} \\ &\quad -\frac{(n-n_\ell)(n_\ell^2n+3n_\ell^2-n_\ell n^2-3n_\ell n+n^2)}{n_\ell^3n}\sigma_{ss}\sigma_{tt}+\tau\sigma_{st}^2, \end{aligned}$$

where $\tau = (1/n_\ell - 1/n)^4\tau_\ell + (1/n)^4\tau_{-\ell} + 4\left(1/n_\ell-1/n\right)^2 n_\ell(n-n_\ell)/n^2$. Combining the above calculations results in

$$V\left[(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^TD^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right] = O\left(\frac{d^2}{n^3}\right)+O\left(\frac{1}{n^2}\right)\sum_{s,t}\rho_{st}^2,$$

where ρ_{st} is the (s,t) component of the correlation matrix R. The sum is evaluated as

$$\sum_{s,t}\rho_{st}^2 = \mathbf{1}_d^T(R\odot R)\mathbf{1}_d \le \lambda_{\max}(R)\left\{\max_{1\le s\le d}\rho_{ss}\right\}\mathbf{1}_d^T\mathbf{1}_d \le b_0 d$$

by the definition of the parameter space Θ in (10), where $\odot$ is the Hadamard product. Therefore, (18) can be evaluated as

$$V\left[(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^TD^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right] = O\left(\frac{d^2}{n^3}\right).$$

Using Chebyshev's inequality, for any $\varepsilon>0$, we have

$$P\left(\left|\frac{(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^TD^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})-E\left[(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})^TD^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell-\overline{\boldsymbol{\varepsilon}})\right]}{C_d}\right|>\varepsilon\right) \le O\left(\frac{d^2}{n^3C_d^2}\right) = o(1).$$

Hence, $(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})$ can be evaluated as

$$(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}) = \left(\frac{1}{n_\ell} - \frac{1}{n}\right) d + o_P(C_d).$$

Next, we evaluate $V\left[(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})\right]$ for $\ell \neq k$. Using Theorems 7.7 and 7.14–7.16 of Schott (1996), we get

$$\begin{aligned} & V\left[(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})\right] \\ & = \operatorname{tr}\left\{(D^{-1} \otimes D^{-1}) E\left[\operatorname{diag}\{(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T \otimes (\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T\}\right]\right\} \\ & \quad - \left\{\operatorname{tr}(D^{-1} E\left[\operatorname{diag}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T\right]\right\}^2. \end{aligned}$$

We first calculate the jth diagonal element of $(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T$. By noting that

$$\overline{\varepsilon}_{\ell j} - \overline{\varepsilon}_j = \left(\frac{1}{n_\ell} - \frac{1}{n}\right) \sum_{i=1}^{n_\ell} \varepsilon_{\ell ij} - \frac{1}{n} \sum_{i=1}^{n_k} \varepsilon_{kij} - \frac{1}{n} \sum_{h \neq \ell, k} \sum_{i=1}^{n_h} \varepsilon_{hij},$$

we have $E\left[(\overline{\varepsilon}_{\ell j} - \overline{\varepsilon}_j)(\overline{\varepsilon}_{kj} - \overline{\varepsilon}_j)\right] = -\sigma_{jj}/n$. Consequently, we obtain

$$\left\{\operatorname{tr}(D^{-1} E\left[\operatorname{diag}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T\right]\right\}^2 \quad = \quad \frac{d^2}{n^2}.$$

Next, we consider the diagonal matrix

$$(D^{-1} \otimes D^{-1}) E\left[\operatorname{diag}\{(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T \otimes (\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})^T\}\right] = \operatorname{diag}\left(u_1, \ldots, u_{d^2}\right),$$

where

$$u_j = \begin{cases} \dfrac{E\left[\left(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s\right)^2 \left(\overline{\varepsilon}_{ks} - \overline{\varepsilon}_s\right)^2\right]}{\sigma_{ss}^2}, & j = (s-1)d + s, \quad s \in \{1, \ldots, d\}, \\ \dfrac{E\left[\left(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s\right)\left(\overline{\varepsilon}_{ks} - \overline{\varepsilon}_s\right)\left(\overline{\varepsilon}_{\ell t} - \overline{\varepsilon}_t\right)\left(\overline{\varepsilon}_{kt} - \overline{\varepsilon}_t\right)\right]}{\sigma_{ss}\sigma_{tt}}, & \text{for all other values of } j < d^2, \text{ and } s \neq t. \end{cases}$$

If $j = (s-1)d + s$, then we have

$$E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^2 (\overline{\varepsilon}_{ks} - \overline{\varepsilon}_s)^2\right] \quad = \quad \frac{n(n_\ell + n_k) - 3n_\ell n_k}{n_\ell n_k n^3} \xi_{ss} + \kappa(\ell, k)\sigma_{ss}^2,$$

where $\kappa(\ell, k)$ is the coefficient of σ_{ss}^2. Note that the order of $\kappa(\ell, k)$ is $O(1/n^2)$ which we state here without giving a detailed proof. On the other hand, we have

$$\begin{aligned} & E\left[\left(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s\right)\left(\overline{\varepsilon}_{ks} - \overline{\varepsilon}_s\right)\left(\overline{\varepsilon}_{\ell t} - \overline{\varepsilon}_t\right)\left(\overline{\varepsilon}_{kt} - \overline{\varepsilon}_t\right)\right] \\ & = \frac{n(n_\ell + n_k) - 3n_\ell n_k}{n_\ell n_k n^3} \xi_{st} + \left\{\frac{3}{n^3} - \frac{1}{n^2}\left(\frac{1}{n_\ell} + \frac{1}{n_k}\right) + \frac{1}{n^2}\right\} \sigma_{ss}\sigma_{tt} + \tau\sigma_{st}^2 \end{aligned}$$

when $\ell = k$, where $\tau = O(1/n^2)$. From the above calculations, we have

$$V\left[(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})\right] = O\left(\frac{d^2}{n^3}\right) + O\left(\frac{1}{n^2}\right) \sum_{s,t} \rho_{st}^2 = O\left(\frac{d^2}{n^3}\right).$$

Chebyshev's inequality now implies that $(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}) = -d/n + o_P(C_d)$, and consequently,

$$\begin{aligned}
&N^{1/2} E_0^T \widehat{D}^{-1} E_0 N^{1/2} \\
&= N^{1/2} \left((\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}})^T D^{-1} (\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}) \right)_{1 \le \ell, k \le K} N^{1/2} (1 + o_P(1)) \\
&= N^{1/2} \left(d \left(\frac{1}{n_\ell} - \frac{1}{n} \right) \delta_{\ell,k} - \frac{d}{n}(1 - \delta_{\ell,k}) + o_P(C_d) \right)_{1 \le \ell, k \le K} N^{1/2} (1 + o_P(1)) \\
&= \frac{d}{n} (I_K - N^{1/2} \mathbf{1}_K \mathbf{1}_K^T N^{1/2}) + \mathbf{1}_K \mathbf{1}_K^T o_P(C_d).
\end{aligned}$$

The previous calculations can now be summarized and lead to the desired expansion of $\widehat{C}^T \widehat{C}/C_d$, namely

$$\begin{aligned}
\frac{\widehat{C}^T \widehat{C}}{C_d} &= \frac{C^T C}{C_d} + \frac{d}{n C_d} (I_K - N^{1/2} \mathbf{1}_K \mathbf{1}_K^T N^{1/2}) + \mathbf{1}_K \mathbf{1}_K^T o_P(1) \\
&= \frac{C^T C}{C_d} + \xi (I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2}) + \mathbf{1}_K \mathbf{1}_K^T o_P(1).
\end{aligned}$$

□

PROOF OF LEMMA 3.3 . From Weyl's inequality (see e.g. Bhatia (1997)), λ_α can be evaluated as

$$\max \left\{ \frac{\lambda^*_{\alpha+1}}{C_d} + \xi, \frac{\lambda^*_\alpha}{C_d} \right\} \le \frac{\lambda_\alpha}{C_d} \le \frac{\lambda^*_\alpha}{C_d} + \xi, \tag{20}$$

for $\alpha = 1, \ldots, K-1$ and $0 \le \lambda_K / C_d \le \lambda^*_K / C_d + \xi = \xi$. In particular, it follows from (20) that

$$\frac{\lambda^*_{\alpha+1}}{C_d} + \xi < \frac{\lambda_\alpha}{C_d} \le \frac{\lambda^*_\alpha}{C_d} + \xi$$

by condition D. Therefore, λ_α / C_d should be simple. □

PROOF OF THEOREM 3.4 .

Put $\Gamma_K = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$, where $\boldsymbol{\gamma}_\ell$ is eigenvector of $C^T C/C_d + \xi(I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2})$ belonging to the ℓth largest eigenvalue. By Lemma 3.2, we obtain

$$\Gamma_K^T \frac{\widehat{C}^T \widehat{C}}{C_d} \Gamma_K = \operatorname{diag} \left(\frac{\lambda_1}{C_d}, \ldots, \frac{\lambda_K}{C_d} \right) (1 + o_P(1)).$$

Let $\widehat{H} = \left[\widehat{\boldsymbol{h}}_1, \ldots, \widehat{\boldsymbol{h}}_K \right]$, where $\widehat{\boldsymbol{h}}_\ell$ is eigenvector of $\Gamma_K^T \left(\widehat{C}^T \widehat{C}/C_d \right) \Gamma_K$ belonging to the ℓth largest eigenvalue. Since all eigenvalues λ_α (for $\alpha = 1, \ldots, K-1$) are simple by Lemma 3.3, it follows that $\widehat{H} \xrightarrow{P} I_K$. From the equation $\Gamma_K^T \left(\widehat{C}^T \widehat{C}/C_d \right) \Gamma_K \widehat{\boldsymbol{h}}_\ell = (\widehat{\lambda}_\ell / C_d) \widehat{\boldsymbol{h}}_\ell$ we can see that

$$\begin{aligned}
&\Gamma_K^T \frac{\widehat{C}^T \widehat{C}}{C_d} \Gamma_K \widehat{\boldsymbol{h}}_\ell = \frac{\widehat{\lambda}_\ell}{C_d} \widehat{\boldsymbol{h}}_\ell \\
&\Longrightarrow \frac{\widehat{C}^T \widehat{C}}{C_d} \left(\Gamma_K \widehat{\boldsymbol{h}}_\ell \right) = \frac{\widehat{\lambda}_\ell}{C_d} \left(\Gamma_K \widehat{\boldsymbol{h}}_\ell \right) \\
&\Longrightarrow \frac{\widehat{C} \widehat{C}^T}{C_d} \left\{ \frac{\widehat{C} \boldsymbol{\gamma}_\ell}{||\widehat{C} \boldsymbol{\gamma}_\ell||} (1 + o_P(1)) \right\} = \frac{\widehat{\lambda}_\ell}{C_d} \left\{ \frac{\widehat{C} \boldsymbol{\gamma}_\ell}{||\widehat{C} \boldsymbol{\gamma}_\ell||} (1 + o_P(1)) \right\}.
\end{aligned} \tag{21}$$

On the other hand,

$$\frac{\widehat{C}\widehat{C}^T}{C_d}\widehat{\boldsymbol{p}}_\ell = \frac{\widehat{\lambda}_\ell}{C_d}\widehat{\boldsymbol{p}}_\ell \tag{22}$$

follows from the definition in Section 3. Now, from (21), (22) and Lemma 3.3, we conclude that the linear span of the $\widehat{\boldsymbol{p}}_\alpha$ is asymptotically equal to that of the $\widehat{C}\boldsymbol{\gamma}_\alpha/||\widehat{C}\boldsymbol{\gamma}_\alpha||$. Since eigenvectors have unit length, $||\boldsymbol{p}_\alpha|| = 1$ and $\mathrm{sgn}\,(\hat{p}_{\alpha 1}) = \mathrm{sgn}\left(\left(\widehat{C}\boldsymbol{\gamma}_\alpha/||\widehat{C}\boldsymbol{\gamma}_\alpha||\right)_1\right)$, where $(\cdot)_1$ denotes the first component of the vector. Therefore, we have

$$\widehat{\boldsymbol{p}}_\alpha = \frac{\widehat{C}\boldsymbol{\gamma}_\alpha}{||\widehat{C}\boldsymbol{\gamma}_\alpha||}\,(1+o_P(1)) \implies \widehat{\boldsymbol{p}}_\alpha^T\frac{\widehat{C}\boldsymbol{\gamma}_\alpha}{||\widehat{C}\boldsymbol{\gamma}_\alpha||} = 1+o_P(1).$$

□

Proof of Theorem 3.5.

From Theorem 3.4 and (11), the inner product of $\widehat{\boldsymbol{p}}_\alpha$ and $\boldsymbol{p}_\beta$ is given by

$$\widehat{\boldsymbol{p}}_\alpha^T\boldsymbol{p}_\beta = \frac{\boldsymbol{\gamma}_\alpha^T\Pi^{1/2}\widehat{M}_0^T D^{-1}M_0\Pi^{1/2}\boldsymbol{\gamma}_\beta(1+o_P(1))}{\sqrt{\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{C}\boldsymbol{\gamma}_\alpha}\sqrt{\boldsymbol{\gamma}_\beta^T C^T C\boldsymbol{\gamma}_\beta}}. \tag{23}$$

The numerator of (23) can be evaluated as

$$\begin{aligned}\boldsymbol{\gamma}_\alpha^T\Pi^{1/2}\widehat{M}_0^T D^{-1}M_0\Pi^{1/2}\boldsymbol{\gamma}_\beta &= \boldsymbol{\gamma}_\alpha^T\Pi^{1/2}M_0^T D^{-1}M_0\Pi^{1/2}\boldsymbol{\gamma}_\beta(1+o_P(1))\\ &= \boldsymbol{\gamma}_\alpha^T C^T C\boldsymbol{\gamma}_\beta(1+o_P(1))\end{aligned}$$

by Chebyshev's inequality. By Lemma 3.2, $\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{C}\boldsymbol{\gamma}_\alpha$ of (23) becomes $\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{C}\boldsymbol{\gamma}_\alpha = \lambda_\alpha(1+o_P(1))$. Notice that $\boldsymbol{\gamma}_\beta^T C^T C\boldsymbol{\gamma}_\beta$ of the denominator of (23) can be written as

$$\begin{aligned}&\boldsymbol{\gamma}_\beta^T C^T C\boldsymbol{\gamma}_\beta\\ &= \boldsymbol{\gamma}_\beta^T\left\{(C^T C + C_d\xi(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})) - C_d\xi(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})\right\}\boldsymbol{\gamma}_\beta\\ &= \lambda_\beta - C_d\xi(1-\boldsymbol{\gamma}_\beta^T\Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2}\boldsymbol{\gamma}_\beta).\end{aligned}$$

Therefore, we obtain

$$\widehat{\boldsymbol{p}}_\alpha^T\boldsymbol{p}_\beta = \frac{\kappa_j\delta_{\alpha\beta} - \xi(\delta_{\alpha\beta}-\eta_\alpha\eta_\beta)}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta - \xi(1-\eta_\beta^2)}}(1+o_P(1)).$$

□

Proof of Corollary 3.6. Follows, since $\xi \to 0$. □

Proof of Theorem 3.7.

The inner product of $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\beta^*$ becomes

$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^* = \frac{\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{D}^{-1/2}D^{-1/2}C\boldsymbol{\gamma}_\beta}{\sqrt{\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{D}^{-1}\widehat{C}\boldsymbol{\gamma}_\alpha}\sqrt{\boldsymbol{\gamma}_\beta^T C^T D^{-1}C\boldsymbol{\gamma}_\beta}}(1+o_P(1)). \tag{24}$$

using Theorem 3.4, (11) and (12). The numerator of (24) can be evaluated as

$$\boldsymbol{\gamma}_\alpha^T \widehat{C}^T \widehat{D}^{-1/2} D^{-1/2} C \boldsymbol{\gamma}_\beta = \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\beta (1 + o_P(1)).$$

Using (17), $\widehat{C}^T \widehat{D}^{-1} \widehat{C}$ of (24) is given by

$$\widehat{C}^T \widehat{D}^{-1} \widehat{C} = C^T D^{-1} C + N^{1/2} E^T \widehat{D}^{-2} E N^{1/2} + \mathbf{1}_K \mathbf{1}_K^T o(C_d).$$

Therefore, we have

$$\begin{aligned}
&\boldsymbol{\gamma}_\alpha^T \widehat{C}^T \widehat{D}^{-1} \widehat{C} \boldsymbol{\gamma}_\alpha \\
&\leq \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha + \frac{1}{\sigma_{\min}} \boldsymbol{\gamma}_\alpha^T N^{1/2} E^T \widehat{D}^{-1} E N^{1/2} \boldsymbol{\gamma}_\alpha (1 + o_P(1)) + o(C_d) \\
&= \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha \\
&\quad \times \left(1 + C_d \xi \frac{1}{\sigma_{\min}} \frac{1 - \boldsymbol{\gamma}_\alpha^T \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2} \boldsymbol{\gamma}_\alpha}{\boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \gamma_\alpha} (1 + o_P(1)) + o\left(\frac{C_d}{\boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha} \right) \right) \\
&\leq \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha \\
&\quad \times \left(1 + C_d \xi \frac{\sigma_{\max}}{\sigma_{\min}} \frac{1 - \boldsymbol{\gamma}_\alpha^T \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2} \boldsymbol{\gamma}_\alpha}{\boldsymbol{\gamma}_\alpha^T C^T C \boldsymbol{\gamma}_\alpha} (1 + o_P(1)) + o\left(\frac{C_d}{\boldsymbol{\gamma}_\alpha^T C^T C \boldsymbol{\gamma}_\alpha} \right) \right) \\
&= \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha \\
&\quad \times \left(1 + C_d \xi \frac{\sigma_{\max}}{\sigma_{\min}} \frac{1 - \eta_\alpha^2}{\lambda_\alpha - C_d \xi (1 - \eta_\alpha^2)} (1 + o_P(1)) + o\left(\frac{C_d}{\lambda_\alpha - C_d \xi (1 - \eta_\alpha^2)} \right) \right) \\
&= \boldsymbol{\gamma}_\alpha^T C^T D^{-1} C \boldsymbol{\gamma}_\alpha \frac{\kappa_\alpha - \xi (1 - \eta_\alpha^2) (1 - \sigma_{\max}/\sigma_{\min})}{\kappa_\alpha - \xi (1 - \eta_\alpha^2)} (1 + o_P(1)),
\end{aligned}$$

where $\sigma_{\max} = \max_{1 \leq j \leq d} \sigma_{jj}$ and $\sigma_{\min} = \min_{1 \leq j \leq d} \sigma_{jj}$. Hence it follows that

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \geq \boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \frac{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2) (1 - \sigma_{\max}/\sigma_{\min})}} (1 + o_P(1)).$$

Similarly, we obtain

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \leq \boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \frac{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2) (1 - \sigma_{\min}/\sigma_{\max})}} (1 + o_P(1)).$$

□

PROOF OF COROLLARY 3.8. Follows, since $\xi \to 0$. □

EVALUATION OF THE MISCLASSIFICATION RATE $W(\widehat{g}, \theta)$.

Suppose that the random vector $\boldsymbol{X}$ belongs to $\mathcal{C}_k$. The correct classification rate of $\widehat{g}$ for class $\mathcal{C}_k$ is defined as

$$\begin{aligned}
\overline{W}_k(\widehat{g}, \theta) &= P\left(\widehat{g}(\boldsymbol{X}) = k \,|\, \boldsymbol{X}_{\ell i}, \ell = 1, \ldots, K, i = 1, \ldots, n_\ell \right) \\
&= P\left(\widehat{g}(\boldsymbol{X}) = k \,|\, X \right).
\end{aligned}$$

We have

$$\begin{aligned}\overline{W}_k(\widehat{g},\theta) &= P\left(\bigcap_{\alpha\neq k}\left\{\omega\in\Omega\,\middle|\,\left(\boldsymbol{X}(\omega)-\frac{1}{2}(\widehat{\boldsymbol{\mu}}_k+\widehat{\boldsymbol{\mu}}_\alpha)\right)^T\widehat{\boldsymbol{w}}_{k\alpha}>0\right\}\middle|\,X\right)\\ &= P\left(\bigcap_{\alpha\neq k}\left\{\omega\in\Omega\,\middle|\,\widehat{\delta}_{k\alpha}(\boldsymbol{X}(\omega))>0\right\}\middle|\,X\right),\end{aligned}$$

where $\widehat{\boldsymbol{w}}_{k\alpha}=\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_\alpha)$. We can easily see that

$$\widehat{\delta}_{k\alpha}(\boldsymbol{X}) \sim N\left(\left(\boldsymbol{\mu}_k-\frac{1}{2}(\widehat{\boldsymbol{\mu}}_k+\widehat{\boldsymbol{\mu}}_\alpha)\right)^T\widehat{\boldsymbol{w}}_{k\alpha},\ \widehat{\boldsymbol{w}}_{k\alpha}^T\Sigma\widehat{\boldsymbol{w}}_{k\alpha}\right),\quad \alpha\neq k.$$

Therefore, $\overline{W}_k(\widehat{g},\theta)$ can be written as

$$\overline{W}_k(\widehat{g},\theta) = P\left(\bigcap_{\alpha\neq k}\left\{\omega\in\Omega\,\middle|\,\widehat{Z}_{k\alpha}(\omega)>-\widehat{d}_{k\alpha}\right\}\middle|\,X\right),$$

where $\widehat{Z}_{k\alpha}=\left(\widehat{\delta}_{k\alpha}(\boldsymbol{X})-E\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]\right)/\sqrt{V\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]}\sim N(0,1)$ and

$$\begin{aligned}\widehat{d}_{k\alpha} &= \frac{E\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]}{\sqrt{V\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]}}\\ &= \frac{(\boldsymbol{\mu}_k-(\widehat{\boldsymbol{\mu}}_k+\widehat{\boldsymbol{\mu}}_\alpha)/2)^T\,\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_\alpha)}{\sqrt{(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_\alpha)^T\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}\Sigma\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_\alpha)}}.\end{aligned}\tag{25}$$

Next, we evaluate the (i,j)th element of the covariance matrix of $(\widehat{Z}_{k1},\ldots,\widehat{Z}_{kK})^T$, where $i,j\in\{1,\ldots,K\}-\{k\}$ and $i\neq j$. From $\widehat{\delta}_{k\alpha}(\boldsymbol{X})-E\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]=(\boldsymbol{X}-\boldsymbol{\mu}_k)^T\widehat{\boldsymbol{w}}_{k\alpha}$, $Cov(\widehat{Z}_{ki},\widehat{Z}_{kj})$ can be written as

$$Cov(\widehat{Z}_{ki},\widehat{Z}_{kj}) = \frac{\widehat{\boldsymbol{w}}_{ki}^T\Sigma\widehat{\boldsymbol{w}}_{kj}}{\sqrt{\widehat{\boldsymbol{w}}_{ki}^T\Sigma\widehat{\boldsymbol{w}}_{ki}}\sqrt{\widehat{\boldsymbol{w}}_{kj}^T\Sigma\widehat{\boldsymbol{w}}_{kj}}}.$$

Therefore, the covariance matrix of $\widehat{\boldsymbol{Z}}_k=(\underline{\widehat{Z}}_{k1},\ldots,\underline{\widehat{Z}}_{k(K-1)})^T$ is

$$\widehat{\Sigma}_k=\widehat{W}_k^T\Sigma\widehat{W}_k,\tag{26}$$

where $\underline{\widehat{Z}}_{k\alpha}=I(\alpha<k)\widehat{Z}_{k\alpha}+I(\alpha\geq k)\widehat{Z}_{k(\alpha+1)}$,

$$\widehat{W}_k=\left[\frac{\underline{\widehat{\boldsymbol{w}}}_{k1}}{\sqrt{\underline{\widehat{\boldsymbol{w}}}_{k1}^T\Sigma\underline{\widehat{\boldsymbol{w}}}_{k1}}},\ldots,\frac{\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}}{\sqrt{\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}^T\Sigma\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}}}\right]$$

and $\underline{\widehat{\boldsymbol{w}}}_{k\alpha} = I(\alpha < k)\widehat{\boldsymbol{w}}_{k\alpha} + I(\alpha \geq k)\widehat{\boldsymbol{w}}_{k(\alpha+1)}$. Now consider the region

$$\widehat{\mathscr{D}}_k = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \,\middle|\, z_j < \underline{\widehat{d}}_{k\alpha}, \alpha \in \{1, \ldots, K-1\} \right\},$$

where $\underline{\widehat{d}}_{k\alpha} = I(\alpha < k)\widehat{d}_{k\alpha} + I(\alpha \geq k)\widehat{d}_{k(\alpha+1)}$. Since $-\boldsymbol{Z}_k$ is also distributed as $N_{K-1}(\boldsymbol{0}, \widehat{\Sigma}_k)$, the correct probability can be obtained as

$$\begin{aligned}
\overline{W}_k(\widehat{g}, \theta) &= P\left(\bigcap_{\alpha=1}^{K-1} \left\{ \omega \in \Omega \,\middle|\, -\widehat{Z}_{k\alpha}(\omega) < \underline{\widehat{d}}_{k\alpha} \right\} \,\middle|\, X \right) \\
&= \int_{\widehat{\mathscr{D}}_k} \frac{1}{\sqrt{|2\pi\widehat{\Sigma}_k|}} \exp\left(-\frac{1}{2} \boldsymbol{z}^T \widehat{\Sigma}_k^{-1} \boldsymbol{z} \right) d\boldsymbol{z} \\
&= \Phi_{K-1}\left(\widehat{\mathscr{D}}_k; \boldsymbol{0}, \widehat{\Sigma}_k \right).
\end{aligned}$$

Therefore, the misclassification rate of $\widehat{g}$ for class $\mathcal{C}_k$ becomes

$$W_k(\widehat{g}, \theta) = 1 - \overline{W}_k(\widehat{g}, \theta) = 1 - \Phi_{K-1}\left(\widehat{\mathscr{D}}_k; \boldsymbol{0}, \widehat{\Sigma}_k \right).$$

□

PROOF OF THEOREM 4.1.

By Theorem 3.4, $\widehat{B}$ is given by

$$\widehat{B} = \widehat{D}^{-1/2}\widehat{P} = \widehat{D}^{-1}\widehat{M}_0 N^{1/2} \Gamma \widehat{L}^{-1}(1 + o_P(1)),$$

where $\widehat{L} = \mathrm{diag}\left(||\widehat{C}\boldsymbol{\gamma}_1||, \ldots, ||\widehat{C}\boldsymbol{\gamma}_{K-1}|| \right)$. Using $\widehat{D} = D(1+o_P(1))$, (25) can be evaluated as

$$\begin{aligned}
\widehat{d}_{k\alpha} &= \frac{\left(\boldsymbol{\mu}_k - (\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)/2\right)^T \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}{\sqrt{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T \Sigma \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}} \\
&\geq \frac{1}{\sqrt{\lambda_{\max}(R)}} \frac{I_1 N^{1/2}\Gamma(\Gamma^T N^{1/2} I_2 N^{1/2}\Gamma)^{-1}\Gamma^T N^{1/2} I_3^T}{\sqrt{I_3 N^{1/2}\Gamma(\Gamma^T N^{1/2} I_2 N^{1/2}\Gamma)^{-1}\Gamma^T N^{1/2} I_3^T}}(1 + o_P(1)),
\end{aligned}$$

where $I_1 = \left(\boldsymbol{\mu}_k - (\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)/2\right)^T D^{-1}\widehat{M}_0$, $I_2 = \widehat{M}_0^T D^{-1}\widehat{M}_0$ and $I_3 = (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M}_0$. We first calculate I_3. Note that I_3 can be decomposed as

$$\begin{aligned}
I_3 &= (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M}_0 \\
&= \left[(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}), \ldots, (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_K - \widehat{\boldsymbol{\mu}}) \right]. \qquad (27)
\end{aligned}$$

From condition A, a typical component of (27) can be expressed as

$$\begin{aligned}
&(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_\ell - \widehat{\boldsymbol{\mu}}) \\
&= \sum_{h=1}^{K} \frac{n_h}{n} \left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h) + (\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_j)^T \widehat{D}^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h) \right. \\
&\quad \left. + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) + (\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) \right].
\end{aligned}$$

Then we have

$$(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h) = o_P\left((\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h)\right)$$
$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) = o_P\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)\right)$$

by p.2625 of Fan and Fan (2008). Next we examine $\sum_{h=1}^{K}(n_h/n)(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h)$, which can be written as

$$\sum_{h=1}^{K}\frac{n_h}{n}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) = \overline{\boldsymbol{\varepsilon}}_k^T D^{-1}\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_k^T D^{-1}\overline{\boldsymbol{\varepsilon}} - \overline{\boldsymbol{\varepsilon}}_\alpha^T D^{-1}\overline{\boldsymbol{\varepsilon}}_\ell + \overline{\boldsymbol{\varepsilon}}_\alpha^T D^{-1}\overline{\boldsymbol{\varepsilon}}.$$

By an argument similar to that given on p.2627 of Fan and Fan (2008), we obtain

$$\sum_{h=1}^{K}\frac{n_h}{n}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) = \begin{cases} \dfrac{d}{n_k} + o_P\left(\sqrt{\dfrac{d}{n}}\right) & \text{if } \ell = k, \\ -\dfrac{d}{n_\alpha} + o_P\left(\sqrt{\dfrac{d}{n}}\right) & \text{if } \ell = \alpha, \\ o_P\left(\sqrt{\dfrac{d}{n}}\right) & \text{otherwise.} \end{cases}$$

We also need to evaluate the asymptotic order of $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0$, which can be written as

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0 = \mathbf{1}_K^T \Pi F,$$

where $F = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K] = ((\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j))_{1\le i,j\le K}$. Using conditions E and F, ℓth component of $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0$ has the following form

$$\begin{aligned} &\mathbf{1}_K^T \Pi \boldsymbol{f}_\ell \\ &= \begin{cases} C_d\left(-\sum_{h\neq k}\sqrt{\pi_h}\dfrac{\boldsymbol{\mu}_k^T D^{-1}\boldsymbol{\mu}_k}{C_d} - \sqrt{\pi_\alpha}\dfrac{\boldsymbol{\mu}_\alpha D^{-1}\boldsymbol{\mu}_\alpha}{C_d} + \sum_{\beta\neq k}\dfrac{c_{\beta k}}{C_d}\right) & \text{if } \ell = k, \\ C_d\left(\sqrt{\pi_k}\dfrac{\boldsymbol{\mu}_k D^{-1}\boldsymbol{\mu}_k}{C_d} + \sum_{h\neq \alpha}\sqrt{\pi_h}\dfrac{\boldsymbol{\mu}_\alpha^T D^{-1}\boldsymbol{\mu}_\alpha}{C_d} + \sum_{\beta\neq\alpha}\dfrac{c_{\beta\alpha}}{C_d}\right) & \text{if } \ell = \alpha, \\ C_d\left(\sqrt{\pi_k}\dfrac{\boldsymbol{\mu}_k D^{-1}\boldsymbol{\mu}_k}{C_d} - \sqrt{\pi_\alpha}\dfrac{\boldsymbol{\mu}_\alpha D^{-1}\boldsymbol{\mu}_\alpha}{C_d} + \sum_{\beta\neq\ell}\dfrac{c_{\beta\ell}}{C_d}\right) & \text{otherwise.} \end{cases} \\ &= O(C_d), \end{aligned}$$

where $c_{\beta\ell} = O(C_d^{\zeta_{\beta\ell}})$ and $\zeta_{\beta\ell} \in (0,1)$ for all β, ℓ. Therefore, we have

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0 = \mathbf{1}_K^T \Pi F = O(C_d)\mathbf{1}_K^T.$$

Using the above calculations, we have

$$
\begin{aligned}
&(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_\ell - \widehat{\boldsymbol{\mu}}) \\
&= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(1 + o_P(1)) + \sum_{h=1}^{K} \frac{n_h}{n}(\overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_\alpha)^T D^{-1}(\overline{\boldsymbol{\varepsilon}}_\ell - \overline{\boldsymbol{\varepsilon}}_h) \\
&\quad + o_P\left(\max_{h\in\{1,\dots K\}} \left\{(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h), (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)\right\}\right) \\
&= \begin{cases}
\left\{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) + \dfrac{d}{n_k}\right\}(1 + o_P(1)) & \text{if } \ell = k, \\
\left\{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) - \dfrac{d}{n_\alpha}\right\}(1 + o_P(1)) & \text{if } \ell = \alpha, \\
(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(1 + o_P(1)) & \text{otherwise,}
\end{cases}
\end{aligned}
$$

by condition D. Thus, it follows that

$$
I_3 = \left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)D^{-1}M_0 + \boldsymbol{\beta}_{k\alpha}\right)(1 + o_P(1)),
$$

where

$$
\boldsymbol{\beta}_{k\alpha} = \left(0, \dots, 0, \frac{d}{n_k}, 0, \dots, 0, -\frac{d}{n_\alpha}, 0, \dots, 0\right).
$$

Next, we consider I_1. We find that

$$
\begin{aligned}
I_1 &= \left(\boldsymbol{\mu}_k - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)\right)^T D^{-1}\widehat{M}_0 \\
&= -\overline{\boldsymbol{\varepsilon}}_k^T D^{-1}\widehat{M}_0 + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M}_0. \qquad (28)
\end{aligned}
$$

Similarly, (28) becomes

$$
\begin{aligned}
&-\overline{\boldsymbol{\varepsilon}}_k^T D^{-1}\widehat{M}_0 + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M}_0 \\
&= \begin{cases}
\dfrac{d}{n}\left(1 - \dfrac{n}{n_k}\right) + \dfrac{1}{2}\left\{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) + \dfrac{d}{n_k}\right\}(1 + o_P(1)) & \text{if } \ell = k, \\
\dfrac{d}{n} + \dfrac{1}{2}\left\{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) - \dfrac{d}{n_\alpha}\right\}(1 + o_P(1)) & \text{if } \ell = \alpha, \\
\dfrac{d}{n} + \dfrac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(1 + o_P(1)) & \text{otherwise,}
\end{cases} \\
&= \begin{cases}
\left[\dfrac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) + \dfrac{d}{n}\left(1 - \dfrac{n}{2n_\ell}\right)\right](1 + o_P(1)) & \text{if } \ell = k, \alpha, \\
\left[\dfrac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}) + \dfrac{d}{n}\right](1 + o_P(1)) & \text{otherwise.}
\end{cases}
\end{aligned}
$$

Therefore, we have

$$
I_1 = \left[\frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}M_0 + \frac{d}{n}\boldsymbol{\alpha}_{k\alpha}\right](1 + o_P(1)),
$$

where

$$
\boldsymbol{\alpha}_{k\alpha} = \left(1, \dots, 1, 1 - \frac{n}{2n_k}, 1, \dots, 1, , 1 - \frac{n}{2n_\alpha}, 1, \dots, 1\right).
$$

Finally, we consider I_2. It can be written as

$$I_2 = \widehat{M}_0 D^{-1} \widehat{M}_0 = \left((\widehat{\boldsymbol{\mu}}_\alpha - \widehat{\boldsymbol{\mu}})^T D^{-1} (\widehat{\boldsymbol{\mu}}_\beta - \widehat{\boldsymbol{\mu}}) \right)_{1 \le \alpha, \beta \le K}. \tag{29}$$

Each component of (29) can be decomposed as

$$(\widehat{\boldsymbol{\mu}}_\alpha - \widehat{\boldsymbol{\mu}})^T D^{-1} (\widehat{\boldsymbol{\mu}}_\beta - \widehat{\boldsymbol{\mu}}) = \left\{ (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu})^T D^{-1} (\boldsymbol{\mu}_\beta - \boldsymbol{\mu}) + J_1 + J_2 \right\} (1 + o_P(1)) + J_3, \tag{30}$$

where $J_1 = (\overline{\boldsymbol{\varepsilon}}_\alpha - \overline{\boldsymbol{\varepsilon}})^T D^{-1} (\boldsymbol{\mu}_\beta - \boldsymbol{\mu})$, $J_2 = (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu})^T D^{-1} (\overline{\boldsymbol{\varepsilon}}_\beta - \overline{\boldsymbol{\varepsilon}})$ and $J_3 = (\overline{\boldsymbol{\varepsilon}}_\alpha - \overline{\boldsymbol{\varepsilon}})^T D^{-1} (\overline{\boldsymbol{\varepsilon}}_\beta - \overline{\boldsymbol{\varepsilon}})$. From calculations similar to those carried out in the derivation of I_1 and I_3, we get

$$\begin{aligned}
J_1 &= o_P\left((\boldsymbol{\mu}_\beta - \boldsymbol{\mu}) D^{-1} (\boldsymbol{\mu}_\beta - \boldsymbol{\mu}) \right), \\
J_2 &= o_P\left((\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}) D^{-1} (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}) \right), \\
J_3 &= \begin{cases} \dfrac{d}{n}\left(\dfrac{n}{n_\alpha} - 1 \right) + o_P\left(\sqrt{\dfrac{d}{n}} \right) & \text{if } \alpha = \beta, \\ -\dfrac{d}{n} + o_P\left(\sqrt{\dfrac{d}{n}} \right) & \text{if } \alpha \neq \beta. \end{cases}
\end{aligned}$$

Consequently, (30) results in

$$\begin{aligned}
&(\widehat{\boldsymbol{\mu}}_\alpha - \widehat{\boldsymbol{\mu}})^T D^{-1} (\widehat{\boldsymbol{\mu}}_\beta - \widehat{\boldsymbol{\mu}}) \\
&= \begin{cases} \left\{ (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu})^T D^{-1} (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}) + \dfrac{d}{n}\left(\dfrac{n}{n_\alpha} - 1 \right) \right\} (1 + o_P(1)) & \text{if } \alpha = \beta, \\ \left\{ (\boldsymbol{\mu}_\alpha - \boldsymbol{\mu})^T D^{-1} (\boldsymbol{\mu}_\beta - \boldsymbol{\mu}) - \dfrac{d}{n} \right\} (1 + o_P(1)) & \text{if } \alpha \neq \beta. \end{cases}
\end{aligned}$$

Therefore, we have

$$I_2 = \left\{ M_0^T D^{-1} M_0 + \frac{d}{n} \left(N^{-1} - \mathbf{1}_K \mathbf{1}_K^T \right) \right\} (1 + o_P(1)).$$

In summary, the components of $\widehat{d}_{k\alpha}$ can be evaluated as

$$\begin{aligned}
I_1 N^{1/2} &= \left[\frac{1}{2} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0 + \frac{d}{n} \boldsymbol{\alpha}_{k\alpha} \right] N^{1/2} (1 + o_P(1)) \\
&= \left[\frac{1}{2} M_{k\alpha} + \frac{d}{n} \boldsymbol{s}_{k\alpha} \Pi^{-1/2} \right] (1 + o_P(1)) \\
&= S_{k\alpha} (1 + o_P(1)), \\
N^{1/2} I_2 N^{1/2} &= N^{1/2} \left[M_0^T D^{-1} M_0 + \frac{d}{n} \left(N^{-1} - \mathbf{1}_K \mathbf{1}_K^T \right) \right] N^{1/2} (1 + o_P(1)) \\
&= \left[C^T C + \frac{d}{n} \left(I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2} \right) \right] (1 + o_P(1)), \\
I_3 N^{1/2} &= \left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha) D^{-1} M_0 + \boldsymbol{\beta}_{k\alpha} \right] N^{1/2} (1 + o_P(1)) \\
&= \left[M_{k\alpha} + \frac{d}{n} \boldsymbol{q}_{k\alpha} \Pi^{-1/2} \right] (1 + o_P(1)) \\
&= Q_{k\alpha} (1 + o_P(1))
\end{aligned}$$

since $N = \Pi(1 + o_P(1))$. Therefore, we have

$$\widehat{d}_{k\alpha} \geq \frac{S_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)\left(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2}\right)\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T(1 + o_P(1))}{\sqrt{\lambda_{\max}(R)}\sqrt{Q_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)\left(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2}\right)\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T}}. \tag{31}$$

This completes the proof of Theorem 4.1. □

PROOF OF COROLLARY 4.2. Using $M_{k\alpha} = \mathbf{1}_K^TO(C_d)$ and $C^TC = \mathbf{1}_K\mathbf{1}_K^TO(C_d)$, $S_{k\alpha}$ of (31) becomes

$$S_{k\alpha} = \frac{M_{k\alpha}}{2} + C_d\left(\frac{d}{nC_d}\right)\boldsymbol{s}_{k\alpha}\Pi^{-1/2} = \frac{M_{k\alpha}}{2} + \mathbf{1}_K^To(C_d) = \frac{M_{k\alpha}}{2}(1 + o_P(1)).$$

Similarly, we can obtain $Q_{k\alpha} = M_{k\alpha}(1 + o_P(1))$ and $U = C^TC(1 + o_P(1))$. □

Acknowledgement

The research of the second author was partially supported by KAKENHI 23500350.

References

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*. **94** 760–766.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential analysis*. **30** 356–399.

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York.

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.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*. **36** 199–227.

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*. **3** 185–205.

Dudoit, S., Fridlyand, 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**(457) 77–87.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*. **36** 2605–2637.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7** 179–188.

Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* John Wiley & Sons.

Hastie, T., Tibshirani, R. and Friedman J. (2001). *The Elements of Statistical Learning*. Springer.

Johnstone, I. M. (2001). On the distribution of the largest principal component. *The Annals of Statistics*. **29** 295–327.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*. **37** 4104–4130.

Khan, J., Wei, J., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001), Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. **7** 673–679.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*. **10**(2) 159–203.

Saeys, Y. and Inza, I. and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*. **23**(19) 2507–2517.

Schott, J. (1996). *Matrix Analysis for Statistics*. New York: Wiley.

Srivastava, M. S. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *Journal of the Japan statistical Society*. **1** 123–134.

Tamatani, M., Koch, I. and Naito, K. (2012). Pattern recognition based on canonical correlations in a high dimension low sample size context. *Journal of Multivariate Analysis*. **111** 350–367.

Weinberger, K., Blitzer, J. and Saul, L. (2006). Distance Metric Learning for Large Margin Nearest Neighbor Classification *Advances in Neural Information Processing Systems 18*. 1473–1480.

Received July 30, 2013
Revised November 6, 2013