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

March, 2014

Mitsuru Tamatani

Graduate School of Science and Engineering, Shimane University

Abstract

This paper is concerned with pattern recognition for $K(\geq 2)$-class problems in a High Dimension Low Sample Size (HDLSS) context. The proposed method is based on canonical correlations between the predictors and response vector of class label. This paper proposes a modified version of the canonical correlation matrix which is suitable for discrimination of a new data in a HDLSS context. We call such a matrix the naive canonical correlation matrix which plays an important role in this paper. Provided the dimension does not grow too fast, we show that the $K-1$ sample eigenvectors of naive canonical correlation matrix 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. Furthermore, we propose variable ranking and feature selection methods which integrate information from all $K-1$ eigenvectors. 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.

Contents

1 Introduction

Pattern recognition is a statistical method by which a new data with unknown class label is classified into one of the previously known classes. Generally, the overall process can be divided into four main steps: preprocessing of data, feature extraction, feature selection, and discrimination. Following Devroye et al. (1996), we use the notion *Pattern Recognition* synonymously with *Discrimination*, and consider the problem of classifying d-dimensional random vectors $\boldsymbol{X}$ into one of K classes.

Fisher's well-known linear discriminant function for K-class problems is the solution to his paradigm: maximize the between-class variance while minimizing the within-class variances. See Fisher (1936), and Rao (1948) for the general multi-class setting. Especially, for the special case of 2-class problems, we assume that the two populations, $\mathcal{C}_1$ and $\mathcal{C}_2$, have multivariate normal distributions which differ in their means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, but share a common covariance matrix Σ. For $\boldsymbol{X}$ from one of the two classes, Fisher's linear discriminant function

$$g(\boldsymbol{X}) = \left(\boldsymbol{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)^T \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{1.1}$$

assigns $\boldsymbol{X}$ to $\mathcal{C}_1$ if $g(\boldsymbol{X}) > 0$, and to $\mathcal{C}_2$ otherwise. If the random vectors have equal probability of belonging to either of the two classes, then the misclassification rate of (1.1) is shown to be

$$\frac{1}{2}\left[\Phi\left(-\frac{\Delta^2}{2}\right) + \left\{1 - \Phi\left(\frac{\Delta^2}{2}\right)\right\}\right],$$

where Φ is the standard normal distribution function and Δ is the Mahalanobis distance between the two populations:

$$\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}.$$

A sample version of (1.1) is

$$\widehat{g}(\boldsymbol{X}) = \left(\boldsymbol{X} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2), \tag{1.2}$$

where $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_2$ and $\widehat{\Sigma}$ are appropriate estimates for the corresponding population parameters of the two classes. (1.2) can be applied to data sets in the general setting where the sample size n is much larger than the dimension d. However, for problems with a large dimension, it has been shown that Fisher's linear discriminant function performs poorly due to diverging spectra. Further for *High Dimension Low Sample Size* (HDLSS) problems, namely $n \ll d$, it is not possible to directly define Fisher's linear discriminant function since the sample covariance matrix $\widehat{\Sigma}$ in (1.2) 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 singularity of $\widehat{\Sigma}$ in different ways: Dudoit et al. (2002) proposed *diagonal linear discriminant analysis* 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, Ahn and Marron (2010) constructed discriminant function based on the Maximal Data Piling (MDP) direction vectors, and Aoshima and Yata (2011) considered a discriminant rule based on second moments in conjunction with geometric representations of high-dimensional data.

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. Further, this approach can be implemented efficiently for large and complex data sets. We refer to Fisher's linear discriminant function 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 et al. (2012) defined 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 modified canonical correlation matrix in the context of two-class problems. Also, for multi-class problems in a HDLSS context, Tamatani et al. (2013) proposed a naive Bayes rule in a general multi-class setting and investigated 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 $K(\geq 2)$ classes, that is, we assume that the dimension d of the data is much bigger than the sample size n. 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 the *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. In principal component analysis, Johnstone (2001) proposes the notion of *spiked* covariance matrices and Ahn et al. (2007) and Jung and Marron (2009) combine the spikiness of the covariance matrix with HDLSS consistency. HDLSS consistency is a measure of the closeness of two vectors in a HDLSS setting, and is given in terms of the angle between the vectors. Ahn et al. (2007) and Jung and Marron (2009) derive precise conditions under which the first eigenvector of the sample covariance matrix is HDLSS consistent. These conditions include the asymptotic rate of growth of the first eigenvalue which exceeds the dimension d. Like Ahn et al. (2007) and Jung and Marron (2009), we consider the asymptotic behavior of $K-1$ eigenvectors in a HDLSS classification context. In our context, however, their spiked covariance model is not appropriate since we are dealing with inverses of the covariance matrix. As a result, our conditions for consistency differ considerably from theirs.

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 are divided 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. We propose new feature selection methods specifically for our multi-class discriminant function.

The paper is organized as follows. In Section 2, we review the framework of pattern recognition and a relationship between Fisher's rule and canonical correlations for K classes. This relationship explicitly exhibits Fisher's ratio of the within-class and between-class variances, and its maximizer provides a discriminant rule. In Section 3, we focus on the gene expression microarray data, which is a typical example of HDLSS data. In general, HDLSS data are emerging in various areas of modern science such as genetic microarrays, medical imaging, text recognition, chemometrics, face recognition, and so on. Further, we present a similar discussion in Section 2 for HDLSS data; we replace Σ by $\mathrm{diag}\Sigma$, and obtain a corresponding criterion, which leads to a natural derivation of a multi-class version of the naive Bayes rule. Section 4 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 normal distribution, 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 et al. (2012). Section 5 includes new feature selection methods for K-class discriminant analysis, which naturally follow from our analysis of the naive canonical correlation matrix. Section 6 provides numerical studies to validate the theorems derived in Section 4. Our numerical results confirm the asymptotic behavior of discriminant directions of Subsection 4.3, and an upper bound for the misclassification rate of the multi-class discriminant function proposed in Subsection 4.4. Furthermore, we apply proposed methods in Section 5 to simulated data sets, and compare the performance of our methods with other ranking methods. Section 7 applies the proposed methods in Section 5 to microarray data example and English spoken data. These comparisons demonstrate that our feature selection methods work well and yield a parsimonious set of features which lead to good classification results. Section 8 contains proofs of our theoretical results. In Section 9, we summarize our results and survey related works and directions for future research.

2 Pattern Recognition

2.1 Problem of Pattern Recognition

Let K be the number of class labels. For observation vector $\boldsymbol{x} \in \mathbb{R}^d$ belonging to a class $k \in \{1, \ldots, K\}$, we define the function g, which is called *discriminant function*, as

$$g : \mathbb{R}^d \longrightarrow \{1, \ldots, K\}.$$

For evaluating the accuracy of the discriminant function g, it is important to focus on the probability

$$P(g(\boldsymbol{X}) \neq Y),$$

which is called misclassification rate of g, where $\boldsymbol{X}$ is a new data taking value in $\mathbb{R}^d$ and Y is the class label of $\boldsymbol{X}$ taking value in $\{1, \ldots, K\}$. Our purpose is to choose a function g^* such as misclassification rate is small, that is, for any discriminant function g,

$$P(g^*(\boldsymbol{X}) \neq Y) \leq P(g(\boldsymbol{X}) \neq Y). \tag{2.1}$$

(2.1) means that g^* is the optimal function, which is called the *Bayes rule* (see Devroye et al. (1996)).

In the following sections, we first consider two-class problem under normal distributions and then investigate the extension to general multi-class discriminant problems. The aim of these sections is to relate Fisher's linear discriminant function and canonical correlation analysis.

2.2 Fisher's Linear Discriminant Function

In this subsection, we consider two-class setting. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two d-dimensional populations with different means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the same covariance matrix Σ. For a random vector $\boldsymbol{X}$ from one of these populations/classes, let π_1 be the probability that $\boldsymbol{X}$ belongs to $\mathcal{C}_1$, and let π_2 be the probability that $\boldsymbol{X}$ belongs to $\mathcal{C}_2$, where $\pi_1 + \pi_2 = 1$.

Linear discriminant function is obtained by projecting observation vector $\boldsymbol{x}$ to one dimension using

$$z(\boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{x}, \tag{2.2}$$

where $\boldsymbol{a}$ is called a weight vector. For $\boldsymbol{x}$ from one of the two classes, (2.2) assigns $\boldsymbol{x}$ to $\mathcal{C}_1$ if $z(\boldsymbol{a}) \geq -a_0$, and otherwise class $\mathcal{C}_2$, where a_0 is a threshold. A characteristic of (2.2) is that it is easier to interpret than nonlinear functions. In addition, we can select a projection that maximizes the class separation by adjusting the components of weight vector. Here an important point is that we need to find an efficient weight vector which maximizes the separation between classes.

Fisher's idea is to obtain the vector $\boldsymbol{a}$ which maximizes the ratio of the *between-class variance* Σ_B and *within-class variance* Σ_W:

$$J^*(\boldsymbol{a}) = \frac{\boldsymbol{a}^T\Sigma_B\boldsymbol{a}}{\boldsymbol{a}^T\Sigma_W\boldsymbol{a}}, \tag{2.3}$$

where

$$\begin{aligned}\Sigma_B &= (E[\ \boldsymbol{X} \mid \mathcal{C}_1\ ] - E[\ \boldsymbol{X} \mid \mathcal{C}_2\ ])\,(E[\ \boldsymbol{X} \mid \mathcal{C}_1\ ] - E[\ \boldsymbol{X} \mid \mathcal{C}_2\ ])^T \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\end{aligned}$$

and

$$\begin{aligned}\Sigma_W &= \sum_{\ell=1}^{2} \pi_\ell V[\ \boldsymbol{X} \mid \mathcal{C}_\ell\ ] \\ &= \sum_{\ell=1}^{2} \pi_\ell E[\ (\boldsymbol{X} - \boldsymbol{\mu}_\ell)(\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T \mid \mathcal{C}_\ell\ ] \\ &= \Sigma.\end{aligned}$$

As is well known, a solution of $\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}\in\mathbb{R}^d-\{\boldsymbol{0}\}} J^*(\boldsymbol{a})$ is given by

$$\boldsymbol{a}^* = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{2.4}$$

see e.g. Bishop and Nasrabadi (2006). Here we select the value

$$\begin{aligned} a_0(\boldsymbol{a}^*) &= \frac{1}{2}\left(E[\ Z(\boldsymbol{a}^*) \mid \mathcal{C}_1\ ] + E[\ Z(\boldsymbol{a}^*) \mid \mathcal{C}_2\ ]\right) \\ &= \frac{1}{2}\boldsymbol{a}^{*T}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \end{aligned}$$

as the threshold using (2.4). Therefore, the linear discriminant function

$$g^*(\boldsymbol{x}) = z(\boldsymbol{a}^*) - a_0(\boldsymbol{a}^*)$$

becomes (1.1), and is called *Fisher's linear discriminant function* or *Fisher's rule.* It is also well known that the choice (2.4) yields the Bayes rule, see Devroye et al. (1996).

2.3 Discriminant Function Based on Canonical Correlations

In this subsection we review the relationship between canonical correlations and Fisher's linear discriminant function in pattern recognition.

In canonical correlation analysis of two subsets of variables $\boldsymbol{X}^{[1]}$ and $\boldsymbol{X}^{[2]}$ of a random vector

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{[1]} \\ \boldsymbol{X}^{[2]} \end{bmatrix}, \tag{2.5}$$

the canonical correlation matrix

$$C_* = \Sigma_1^{-1/2}\Sigma_{1,2}\Sigma_2^{-1/2}, \tag{2.6}$$

and the derived matrix $C_*C_*^T$ play important roles, see Mardia et al. (1979). Here Σ_ℓ is the covariance matrix of $\boldsymbol{X}^{[\ell]}$ ($\ell = 1, 2$), and $\Sigma_{1,2}$ is the between-covariance matrix of the two vectors:

$$\begin{aligned} \Sigma_\ell &= E\left[(\boldsymbol{X}^{[\ell]} - E[\boldsymbol{X}^{[\ell]}])(\boldsymbol{X}^{[\ell]} - E[\boldsymbol{X}^{[\ell]}])^T\right], \\ \Sigma_{1,2} &= E\left[(\boldsymbol{X}^{[1]} - E[\boldsymbol{X}^{[1]}])(\boldsymbol{X}^{[2]} - E[\boldsymbol{X}^{[2]}])^T\right]. \end{aligned}$$

For random vectors $\boldsymbol{X}$ belonging to one of the classes $\mathcal{C}_1$ and $\mathcal{C}_2$, we replace $\boldsymbol{X}^{[1]}$ in (2.5) by $\boldsymbol{X}$, and $\boldsymbol{X}^{[2]}$ by the vector of labels $\boldsymbol{Y}$ defined by

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \quad \text{and} \quad P\left(\boldsymbol{Y} = \boldsymbol{e}_\ell\right) = \pi_\ell, \tag{2.7}$$

where $\boldsymbol{e}_\ell$ is the vector with ℓth entry 1 and 0 otherwise. For $\boldsymbol{X}$ and $\boldsymbol{Y}$ we define the matrix

$$\widetilde{C} = \Sigma^{-1/2} E\left[(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{Y}^T\right]\left\{E\left[\boldsymbol{Y}\boldsymbol{Y}^T\right]\right\}^{-1/2}, \tag{2.8}$$

where $\boldsymbol{\mu} = \pi_1\boldsymbol{\mu}_1 + \pi_2\boldsymbol{\mu}_2$. Strictly speaking, the centered $\boldsymbol{Y}$ should be used for $\widetilde{C}$ in (2.8), however, for the vector of labels $\boldsymbol{Y}$, centering is not meaningful. From (2.7) and (2.8) it follows that

$$\begin{aligned}
E\left[\boldsymbol{Y}\boldsymbol{Y}^T\right] &= \begin{bmatrix} \pi_1 & 0 \\ 0 & \pi_2 \end{bmatrix}, \\
E\left[(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{Y}^T\right] &= E[\boldsymbol{X}\boldsymbol{Y}^T] - \boldsymbol{\mu}E[\boldsymbol{Y}^T] \\
&= [\ \pi_1\boldsymbol{\mu}_1 \quad \pi_2\boldsymbol{\mu}_2\ ] - \boldsymbol{\mu}[\pi_1,\ \pi_2] \\
&= [\ \pi_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}) \quad \pi_2(\boldsymbol{\mu}_2-\boldsymbol{\mu})\ ],
\end{aligned}$$

and hence

$$\widetilde{C}\widetilde{C}^T = \Sigma^{-1/2} M \Sigma^{-1/2}, \tag{2.9}$$

where

$$\begin{aligned}
M &= E\left[(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{Y}^T\right]\left\{E\left[\boldsymbol{Y}\boldsymbol{Y}^T\right]\right\}^{-1} E\left[(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{Y}^T\right]^T \\
&= \sum_{\ell=1}^{2} \pi_\ell(\boldsymbol{\mu}_\ell-\boldsymbol{\mu})(\boldsymbol{\mu}_\ell-\boldsymbol{\mu})^T.
\end{aligned}$$

Now, consider the eigenvalue problem $\widetilde{C}\widetilde{C}^T\widetilde{\boldsymbol{p}} = \tilde{\lambda}\widetilde{\boldsymbol{p}}$. Using the expression (2.9), we see that solving the eigenvalue problem is equivalent to maximizing $\widetilde{J}(\boldsymbol{b})$ defined by

$$\widetilde{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T \Sigma \boldsymbol{b}}. \tag{2.10}$$

From

$$\begin{aligned}
M &= \sum_{\ell=1}^{2} \pi_\ell(\boldsymbol{\mu}_\ell-\boldsymbol{\mu})(\boldsymbol{\mu}_\ell-\boldsymbol{\mu})^T \\
&= \pi_1\pi_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \\
&= \pi_1\pi_2\Sigma_B,
\end{aligned}$$

(2.10) is nothing other than the criterion which yields Fisher's rule (2.3).

2.4 Fisher's Rule for Multi-class

We consider the K-class discrimination problem. Let $\mathcal{C}_\ell$ ($\ell = 1, \dots, K$) 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 $P\left(\boldsymbol{Y} = \boldsymbol{e}_\ell\right) = \pi_\ell$, and $\sum_{\ell=1}^{K} \pi_\ell = 1$.

In 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 (2.8) plays an important role. From the definition of $(\boldsymbol{X}, \boldsymbol{Y})$ it follows that

$$E[\boldsymbol{Y}\boldsymbol{Y}^T] = \Pi \quad \text{and} \quad E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] = M_0 \Pi, \tag{2.11}$$

where $\Pi = \mathrm{diag}(\pi_1, \dots, \pi_K)$ and $M_0 = [\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}\ \cdots\ \boldsymbol{\mu}_K - \boldsymbol{\mu}\ ]$. Using

$$M \;=\; \sum_{\ell=1}^{K} \pi_\ell (\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})^T \;=\; M_0 \Pi M_0^T,$$

we have (2.9). 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

$$\widetilde{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T \Sigma \boldsymbol{b}} \tag{2.12}$$

over vectors $\boldsymbol{b}$. Note that (2.12) 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 $\alpha \in \{1, \dots, K-1\}$, put $Z_\alpha(\boldsymbol{X}) = \widetilde{\boldsymbol{b}}_\alpha^T \boldsymbol{X}$, and define the vector $Z(\boldsymbol{X}) = [Z_1(\boldsymbol{X}), \dots, Z_{K-1}(\boldsymbol{X})]^T = \widetilde{B}^T\boldsymbol{X}$. For $Z(\boldsymbol{X})$ and $\mathcal{C}_\ell$, define the Mahalanobis distance between $Z(\boldsymbol{X})$ and $\mathcal{C}_\ell$ by

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

where

$$\boldsymbol{\nu}_\ell = E[\ Z(\boldsymbol{X}) \mid \mathcal{C}_\ell\ ] = E[\ Z(\boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{e}_\ell\ ] = \widetilde{B}^T \boldsymbol{\mu}_\ell$$

and

$$\Sigma_\ell = V[\ Z(\boldsymbol{X}) \mid \mathcal{C}_\ell\ ] = V[\ Z(\boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{e}_\ell\ ] = \widetilde{B}^T \Sigma \widetilde{B}.$$

Hence $\Delta_\ell(Z(\boldsymbol{X}))^2$ can be rewritten as

$$\Delta_\ell(Z(\boldsymbol{X}))^2 \quad = \quad (\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T\, \widetilde{B}(\widetilde{B}^T \Sigma \widetilde{B})^{-1} \widetilde{B}^T\, (\boldsymbol{X} - \boldsymbol{\mu}_\ell)\,. \tag{2.13}$$

Using (2.13), we now derive the multi-class discriminant function $\widetilde{g}$ as the minimizer of the Mahalanobis distance, and let

$$\widetilde{g}(\boldsymbol{X}) = \underset{\ell \in \{1,\ldots,K\}}{\arg\min}\, \Delta_\ell(Z(\boldsymbol{X}))^2. \tag{2.14}$$

Note that $\widetilde{B}$ depends on Σ and that (2.14) reduces to the Fisher's linear discriminant function in the case of 2-class by considering the sign of $\Delta_1(Z(\boldsymbol{X}))^2 - \Delta_2(Z(\boldsymbol{X}))^2$.

2.5 Other Discriminant Functions

In previous subsections, we introduced Fisher's rule and discriminant function based on canonical correlations. Another discriminant functions have been discussed in literatures which include empirical risk minimization method, nearest neighbor rule, kernel method, neural network, and so on (see Devroye et al. (1996), Bishop and Nasrabadi (2006), Duda et al. (2012)).

3 HDLSS

3.1 HDLSS Setting

Gene expression microarray data is a form of high-throughput biological data providing relative measurements of mRNA levels for genes in a biological sample. One important application of gene expression microarray data is the classification of new samples into known classes. In the recent years, many laboratories have collected and analyzed microarray data, and such data have been appeared in public databases or on web sites. Indeed, the data collected from gene expression microarrays consist of thousands or tens of thousands of genes that constitute features: leukemia data (Golub et al. (1999), $K = 2,\ n = 73,\ d = 7129$), lung cancer data (Gordon et al. (2002), $K = 2,\ n = 184,\ d = 12433$), prostate cancer data (Singh et al. (2002), $K = 2,\ n = 134,\ d = 12600$), the small round blue cell tumors (SRBCT) data (Khan et al. (2001), $K = 4,\ n = 83,\ d = 2308$), and so on. These data, often called the High Dimension Low Sample Size (HDLSS) data, are characterized with large number of dimensions d and a relatively small number of sample size n, that is, we can write $n \ll d$.

To establish asymptotic theory in this paper, we consider an asymptotic situation in which the dimension d and the sample size n both approach infinity in such a way that d is much faster than n:

$$n = o(d) \quad \text{as} \quad n, d \ \longrightarrow\ \infty.$$

3.2 Difficulties in HDLSS

The HDLSS data involves serious problems. Let $\boldsymbol{X}_i$ $(i = 1, \ldots, n)$ be *independently and identically distributed* (i.i.d.) d-dimensional random vectors with the covariance matrix Σ. The usual estimates of Σ is defined as

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})^T, \tag{3.1}$$

where $\widehat{\boldsymbol{\mu}} = (1/n) \sum_{i=1}^{n} \boldsymbol{X}_i$ is the mean vector of the data $\{\boldsymbol{X}_i\}_{1 \le i \le n}$. When we assume that the data satisfies $n \ll d$, (3.1) is a singular matrix because the rank of (3.1) is at most $n - 1$. Hence, it means that the sample version of (2.8) is not directly applicable in the case of $n \ll d$.

The difficulty of high-dimensional classification in data analysis is intrinsically caused by the existence of many noise features that do not contribute to the decrease of misclassification rate. To avoid the above problem, the importance of feature selection has been stressed and several methods of feature selection for HDLSS data have been proposed in literatures, for example, *nearest shrunken centroids method* (Tibshirani et al. (2002)) , *feature annealed independence rules* (Fan and Fan (2008)), *minimum redundancy maximum relevance feature selection* (Ding and Peng (2005), Peng et al. (2005)), and so on.

3.3 Naive Canonical Correlations

In the HDLSS two-class discrimination settings, the sample covariance matrix $\widehat{\Sigma}$ is singular as was discussed in the previous subsection. For a population framework, it therefore does not make sense to define a discriminant function based on the within-class covariance matrix Σ. To overcome such a difficulty, 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

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

and $\boldsymbol{b}_\alpha = D^{-1/2}\boldsymbol{p}_\alpha$, 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

$$J(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T D \boldsymbol{b}}. \tag{3.2}$$

Note that Σ in (2.12) has been replaced by the diagonal matrix D in (3.2).

3.4 Discriminant Function in HDLSS

The corresponding discriminant function g is therefore

$$g(\boldsymbol{X}) = \underset{\ell\in\{1,\ldots,K\}}{\arg\min}\ (\boldsymbol{X}-\boldsymbol{\mu}_\ell)^T B(B^T D B)^{-1} B^T (\boldsymbol{X}-\boldsymbol{\mu}_\ell)\,. \tag{3.3}$$

Note that Σ has been replaced by D both in (2.12) and (2.13) to yield (3.2) and (3.3) respectively.

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

4 Asymptotic Theory

In this section, we evaluate the asymptotic behavior of the estimators of $\boldsymbol{p}_\alpha$ and $\boldsymbol{b}_\alpha$ in a HDLSS setting under general assumptions about the underlying distributions. In addition, we derive upper bounds for the misclassification rate of multi-class discriminant function under d-dimensional normal populations.

4.1 Preliminaries

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

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

Let X and Y be matrices defined by $X = [\ \boldsymbol{X}_{11}\ \cdots\ \boldsymbol{X}_{Kn_K}\ ]$ and $Y = [\ \boldsymbol{Y}_{11}\ \cdots\ \boldsymbol{Y}_{Kn_K}\ ]$. 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, a natural estimator for C is

$$\begin{aligned} \widehat{C} &= \widehat{D}^{-1/2} \left\{ \frac{1}{n} \left(X \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \right) Y^T \right\} \left(\frac{1}{n} Y Y^T \right)^{-1/2} \\ &= \widehat{D}^{-1/2} \widehat{M}_0 N^{1/2}, \qquad (4.1) \end{aligned}$$

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, $\widehat{M_0} = [\ \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}\ \cdots\ \widehat{\boldsymbol{\mu}}_K - \widehat{\boldsymbol{\mu}}\ ]$ and $N = \mathrm{diag}(n_1/n, \ldots, n_K/n)$. Hence we obtain the expression

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

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 $\widehat{\boldsymbol{b}}_\alpha = \widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_\alpha$ ($\alpha = 1, \ldots, K-1$). 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

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

then $\widehat{g}$, defined by

$$\widehat{g}(\boldsymbol{X}) = \underset{\ell \in \{1,\ldots,K\}}{\arg\min}\ (\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_\ell)^T\, \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T\, (\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_\ell)\,, \tag{4.2}$$

is a natural estimator of g in (3.3), 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.

Throughout this paper, we make the assumption that the sample size of each of the K classes satisfies $c \leq n_\ell/n$ 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.

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

DEFINITION 4.1 *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*

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

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

CONDITION A. Let $X = [\ \boldsymbol{X}_{11} \ \cdots \ \boldsymbol{X}_{Kn_K}\ ]$ 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 $\boldsymbol{\varepsilon}_{\ell i}$ $(\ell = 1, \ldots, K;\ i = 1, \ldots, n_\ell)$ 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

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

CONDITION C. Let $n_\ell \to \infty$ (for $\ell = 1, \ldots, K$), $d \to \infty$, $\log d = o(n)$, $n = o(d)$. 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 $k \neq \ell$,

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

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.

4.2 Asymptotic Results for Eigenvectors

To establish the asymptotic behaviour of the estimators $\widehat{\boldsymbol{p}}_\alpha$, 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:

$$\Theta = \left\{ (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma) \;\middle|\; \begin{array}{l} \min_{k\neq\ell} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)^T D^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \geq C_d, \\ \lambda_{\max}(R) \leq b_0, \quad \min_{1\leq j\leq d} \sigma_{jj} > 0 \end{array} \right\}, \tag{4.3}$$

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 4.1 *Suppose that* Condition A – Condition 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 (2.11).

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 $C^TC/C_d+\xi(I_K-\Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$ are simple in the case of $d=O(nC_d)$, that is, $\xi>0$.

LEMMA 4.2 *Let λ_α/C_d be αth largest eigenvalue of $C^TC/C_d+\xi(I_K-\Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$. 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 4.1 and 4.2, we can now describe the asymptotic behavior of $\widehat{\boldsymbol{p}}_\alpha$ as follows.

THEOREM 4.1 *Suppose that* CONDITION A – CONDITION 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 $C^TC/C_d+\xi(I_K-\Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$ belonging to the non-zero eigenvalue λ_α.

On the other hand, especially in the case of $K=2$, the eigenvector $\widehat{\boldsymbol{p}}_1$ of $\widehat{C}\widehat{C}^T$ corresponding to the non-zero eigenvalue can be easily calculated in a closed form. Using $\widehat{\boldsymbol{\mu}}=(n_1/n)\widehat{\boldsymbol{\mu}}_1+(n_2/n)\widehat{\boldsymbol{\mu}}_2$ and

$$\begin{aligned}\widehat{M_0} &= [\ \widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}} \quad \widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}\ ] \\ &= \left[\ \frac{n_2}{n}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2) \quad -\frac{n_1}{n}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)\ \right] \\ &= (\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)\left[\frac{n_2}{n},\ -\frac{n_1}{n}\right],\end{aligned}$$

(4.1) can be expressed as

$$\widehat{C} \;=\; \sqrt{\frac{n_1 n_2}{n^2}}\widehat{D}^{-1/2}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)\left[\sqrt{\frac{n_2}{n}},\ -\sqrt{\frac{n_1}{n}}\right] \tag{4.4}$$

from which we obtain the expression

$$\widehat{C}^T\widehat{C} \;=\; \hat{\lambda}^* \begin{bmatrix} \sqrt{\frac{n_2}{n}} \\ -\sqrt{\frac{n_1}{n}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{n_2}{n}} \\ -\sqrt{\frac{n_1}{n}} \end{bmatrix}^T , \tag{4.5}$$

where $\hat{\lambda}^* = \left\{(n_1 n_2)/n^2\right\}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)^T \widehat{D}^{-1}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)$. From (4.5), it follows that the rank of $\widehat{C}^T\widehat{C}$ is one. The non-zero eigenvalue of $\widehat{C}^T\widehat{C}$ is $\hat{\lambda}^*$ and its eigenvector is

$$\widehat{\boldsymbol{p}}_0 \equiv \begin{bmatrix} \sqrt{\frac{n_2}{n}} \\ -\sqrt{\frac{n_1}{n}} \end{bmatrix}.$$

Hence the normalized eigenvector $\widehat{\boldsymbol{p}}_1$ of $\widehat{C}\widehat{C}^T$ is given by

$$\widehat{\boldsymbol{p}}_1 = \frac{\widehat{C}\widehat{\boldsymbol{p}}_0}{||\widehat{C}\widehat{\boldsymbol{p}}_0||} = \frac{\widehat{D}^{-1/2}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)}{\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)^T \widehat{D}^{-1}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)}. \tag{4.6}$$

Proofs of Lemmas 4.1, 4.2 and Theorem 4.1 are given in Section 8. To gain further insight in the behavior of the $\widehat{\boldsymbol{p}}_\alpha$, we define the vectors

$$\boldsymbol{p}_\alpha = \frac{C\boldsymbol{\gamma}_\alpha}{||C\boldsymbol{\gamma}_\alpha||} \quad \text{for} \quad \alpha = 1, \ldots, K-1 \tag{4.7}$$

by referring to Theorem 4.1. 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 4.2 *Suppose that* Condition A – Condition 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 $\theta \in \Theta$,*

$$\widehat{P}^T P \;\xrightarrow{P}\; \left(\frac{\kappa_\beta\delta_{\alpha\beta} - \xi(\delta_{\alpha\beta} - \eta_\alpha\eta_\beta)}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta - \xi(1-\eta_\beta^2)}}\right)_{1\le\alpha,\beta\le K-1} , \tag{4.8}$$

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

Corollary 4.1 *Suppose that* Condition A – Condition E *hold, and that* $d = o(nC_d)$. *Then for all parameters* $\theta \in \Theta$,

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

Corollary 4.1 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.

For the two-class setting, we obtain a simple expression of (4.8) that does not need to use $\boldsymbol{\gamma}_1$ appeared in Theorem 4.2. Similarly to (4.4) and (4.5), we have

$$C = \sqrt{\pi_1\pi_2} D^{-1/2} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \left[\sqrt{\pi_2}, \ -\sqrt{\pi_1}\right]$$

and

$$C^T C = \lambda^* \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix} \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix}^T,$$

where $\lambda^* = \pi_1\pi_2 \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T D^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$. Also, the matrix $I_2 - \Pi^{1/2}\mathbf{1}_2\mathbf{1}_2^T\Pi^{1/2}$ can be written as

$$\begin{aligned} I_2 - \Pi^{1/2}\mathbf{1}_2\mathbf{1}_2^T\Pi^{1/2} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \pi_1 & \sqrt{\pi_1\pi_2} \\ \sqrt{\pi_1\pi_2} & \pi_2 \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix} \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix}^T. \end{aligned}$$

Therefore, we obtain

$$\frac{C^T C}{C_d} + \xi(I_2 - \Pi^{1/2}\mathbf{1}_2\mathbf{1}_2^T\Pi^{1/2}) = \left(\frac{\lambda^*}{C_d} + \xi\right) \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix} \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix}^T.$$

Hence

$$\boldsymbol{\gamma}_1 = \begin{bmatrix} \sqrt{\pi_2} \\ -\sqrt{\pi_1} \end{bmatrix}$$

is an eigenvector of $C^T C / C_d + \xi(I_2 - \Pi^{1/2}\mathbf{1}_2\mathbf{1}_2^T\Pi^{1/2})$ with $\lambda^*/C_d + \xi$.

From the assumption of Theorem 4.2, we can see that $\lambda_1/C_d = \lambda^*/C_d + \xi$ and $\boldsymbol{\gamma}_1^T \Pi^{1/2} \mathbf{1}_2 = 0$. If we add the assumption that the ratio of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and C_d converges to $\zeta > 0$:

$$\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{C_d} \quad \longrightarrow \quad \zeta \quad \text{as} \quad d \to \infty, \tag{4.9}$$

then (4.8) can be written as

$$\begin{aligned} \widehat{\boldsymbol{p}}_1^T \boldsymbol{p}_1 \quad & \xrightarrow{P} \quad \sqrt{\frac{\kappa_1 - \xi(1-\eta_1^2)}{\kappa_1}} \\ & = \quad \sqrt{\frac{\pi_1\pi_2\zeta}{\pi_1\pi_2\zeta + \xi}} \\ & = \quad \sqrt{\frac{1}{1+\xi/(\pi_1\pi_2\zeta)}}. \end{aligned} \tag{4.10}$$

(4.10) is essentially equivalent to Theorem 2 of Tamatani et al. (2012).

4.3 Asymptotic Results for Discriminant Directions

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

$$\widehat{\boldsymbol{b}}_\alpha^* = \frac{\widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_\alpha}{\sqrt{\widehat{\boldsymbol{p}}_\alpha^T \widehat{D}^{-1} \widehat{\boldsymbol{p}}_\alpha}}, \quad \boldsymbol{b}_\alpha^* = \frac{D^{-1/2}\boldsymbol{p}_\alpha}{\sqrt{\boldsymbol{p}_\alpha^T D^{-1} \boldsymbol{p}_\alpha}} \tag{4.11}$$

for $\alpha = 1, \ldots, K-1$. Then we have the following theorem and corollary:

Theorem 4.3 *Suppose that* Condition A – Condition 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* $\theta \in \Theta$,

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \le \frac{\boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\min}/\sigma_{\max})}} (1 + o_P(1)), \tag{4.12}$$

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \ge \frac{\boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\max}/\sigma_{\min})}} (1 + o_P(1)). \tag{4.13}$$

Corollary 4.2 *Let* $\widehat{B}^* = [\ \widehat{\boldsymbol{b}}_1^* \ \cdots \ \widehat{\boldsymbol{b}}_{K-1}^*\ ]$ *and* $B^* = [\ \boldsymbol{b}_1^* \ \cdots \ \boldsymbol{b}_{K-1}^*\ ]$. *Suppose that* Condition A – Condition E *hold, and that* $d = o(nC_d)$. *Then for all parameters* $\theta \in \Theta$,

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

Theorem 4.3 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

$$\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\alpha^* - \boldsymbol{b}_\alpha^{*T}\boldsymbol{b}_\alpha^*\sqrt{1 - \frac{\xi}{\kappa_\alpha}(1-\eta_\alpha^2)} \ \xrightarrow{P}\ 0. \tag{4.14}$$

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 $\alpha \neq \beta$, since

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

Remark In the two-class setting, (4.12) and (4.13) can be written as

$$\widehat{\boldsymbol{b}}_1^{*T}\boldsymbol{b}_1^* \leq \frac{1}{\sqrt{1+\xi\left(\sigma_{\min}/\sigma_{\max}\right)/(\pi_1\pi_2\zeta)}}(1+o_P(1)), \tag{4.15}$$

$$\widehat{\boldsymbol{b}}_1^{*T}\boldsymbol{b}_1^* \geq \frac{1}{\sqrt{1+\xi\left(\sigma_{\max}/\sigma_{\min}\right)/(\pi_1\pi_2\zeta)}}(1+o_P(1)) \tag{4.16}$$

by using the additional assumption (4.9). Tamatani et al. (2012) derived the asymptotic behavior of discriminant direction $\widehat{\boldsymbol{b}}_1^*$ under the constrained parameter space in their Theorem 5. However, we can see that the asymptotic behavior of discriminant directions can be evaluated on Θ from Theorem 4.3, and such restrictions of the parameter space Θ utilized in Tamatani et al. (2012) are not necessary.

4.4 Upper Bound for 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.

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

CONDITION A1. Let $X = [\ \boldsymbol{X}_{11} \ \cdots \ \boldsymbol{X}_{Kn_K}\ ]$ 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 i.i.d. random vectors distributed as $N_d(\mathbf{0}, \Sigma)$, d-dimensional normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

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 (3.3), for class $\mathcal{C}_k$ is defined as

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

where $\widehat{\Sigma}_k$ is the transformed covariance matrix of size $(K-1)\times(K-1)$ which is defined in (8.13) in Section 8, and $\widehat{\mathscr{D}}_k$ is the $(K-1)$-dimensional region given by

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

where

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

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 $\mathcal{C}_k$:

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

In addition to the region $\widehat{\mathscr{D}}_k$ we also require the following region and quantities in Theorem 4.4:

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

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

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

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

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

Furthermore, we need the following region in Corollary 4.3:

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

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

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

We have the following theorem and corollary using Theorem 4.1.

Theorem 4.4 *Suppose that* Condition A1 *and* Condition B – Condition 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 4.3 *Suppose that* Condition A1 *and* Condition B – Condition F *hold, and that $d = o(nC_d)$. Then, for all parameters $\theta \in \Theta_k$,*

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

Note that Theorem 4.4 and Corollary 4.3 extend Theorem 1 in Fan and Fan (2008) to the general multi-class setting considered in this paper.

Especially in the case of two-class setting, we can achieve a smaller upper bound for the misclassification rate than has previously been established in Fan and Fan (2008) under the assumptions implying the HDLSS consistency of $\widehat{\boldsymbol{b}}_1^*$.

The worst case misclassification rate for g is defined as

$$W_k(g) = \max_{\theta \in \Theta} W_k(g, \theta),$$

where Θ is the parameter space in (4.3). Note that we have $\Theta = \Theta_1 (= \Theta_2)$ in the case of two-class setting. The upper bound for the misclassification rate for two-class setting has been derived by Fan and Fan (2008) as follows:

Theorem 4.5 *(Fan and Fan (2008)) Suppose that* Condition A1, Condition B *and* Condition C *hold, and that $d = O(nC_d)$. Then, for $\theta \in \Theta$, the misclassification rate $W_1(\widehat{g}, \theta)$ is bounded above by*

$$W_1(\widehat{g}, \theta) \le 1 - \Phi\left(\frac{\sqrt{n_1 n_2/(dn)}\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}(1 + o_P(1)) + (n_1 - n_2)\sqrt{d/(nn_1n_2)}}{2\sqrt{\lambda_{\max}(R)}\sqrt{1 + n_1 n_2 \boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}(1 + o_P(1))/(dn)}}\right), \quad (4.17)$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

The worst case misclassification rate for $\widehat{g}$ is also derived in Fan and Fan (2008), and is given under the assumption $d = o(nC_d)$. Their result is

$$W_1(\widehat{g}) = 1 - \Phi\left(\frac{1}{2}\sqrt{\frac{n_1 n_2}{dnb_0}} C_d \{1 + o_P(1)\}\right).$$

This results seems to be derived from the bound stated in Theorem 4.5 – part (i) of their Theorem 1. Our calculations in Theorem 4.6 are based on the explicit assumption $d = o(nC_d)$, which yields a tighter bound.

Theorem 4.6 *Suppose that* Condition A1*,* Condition B *and* Condition C *hold, and that $d = o(nC_d)$. Then, for $\theta \in \Theta$, the misclassification rate $W_1(\widehat{g}, \theta)$ is bounded above by*

$$W_1(\widehat{g}, \theta) \leq 1 - \Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}}}{2\sqrt{\lambda_{\max}(R)}}(1 + o_P(1))\right),$$

with $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Moreover, for the worst case misclassification rate, we have

$$W_1(\widehat{g}) = 1 - \Phi\left(\frac{\sqrt{C_d}}{2\sqrt{b_0}}(1 + o_P(1))\right).$$

Note that Theorem 4.5 is the result for $d = O(nC_d)$, while $d = o(nC_d)$ is assumed in Theorem 4.6. We see that, again, the rate of growth of d relative to nC_d plays an important role. The upper bounds in Theorems 4.5 and 4.6 have the following relationship:

Corollary 4.4 *Suppose that* Condition A1*,* Condition B *and* Condition C *hold, and that $n_2/n_1 = c_0 + o(1)$ for $1 \leq c_0 < \infty$. Then, as $d \to \infty$, for any $\theta \in \Theta$ satisfying $d = O(nC_d)$ and, $\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}/C_d = O(1)$, where $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$,*

$$1 - \Phi\left(\frac{\sqrt{n_1 n_2/(dn)}\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}(1 + o_P(1)) + (n_1 - n_2)\sqrt{d/(nn_1n_2)}}{2\sqrt{\lambda_{\max}(R)}\sqrt{1 + n_1 n_2 \boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}(1 + o_P(1))/(dn)}}\right) \tag{4.18}$$

$$> 1 - \Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}}}{2\sqrt{\lambda_{\max}(R)}}(1 + o_P(1))\right). \tag{4.19}$$

Therefore, (4.19), the bound obtained in Theorem 4.6, is smaller than the bound (4.18) of Theorem 4.5. It is interesting to observe that the assumption $d = o(nC_d)$, which leads to the desirable hdlss consistency of $\widehat{\boldsymbol{b}}_1^*$ in Corollaly 4.2, is also responsible for the smaller error bound (4.19).

5 Feature Selection

5.1 Feature Selection in 2-class

Koch and Naito (2010) propose feature selection in a regression context, which is based on two different 'ranking vectors': the eigenvector $\widehat{\boldsymbol{p}}_*$ of the matrix $\widehat{C}_*\widehat{C}_*^T$, and the first canonical correlation vector $\widehat{\boldsymbol{b}}_* = \widehat{\Sigma}_1^{-1/2}\widehat{\boldsymbol{p}}_*$, where $\widehat{C}_*$ is the sample version of (2.6). Analogously, we consider a naive Bayes discriminant function

$$\begin{aligned}\widehat{g}(\boldsymbol{x}) &= \left(\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{D}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) \\ &= \hat{c}\left(\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_1 \\ &= \hat{c}\left(\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{\boldsymbol{b}}_1\end{aligned}$$

which includes feature selection based on $\widehat{\boldsymbol{p}}_1$ and on $\widehat{\boldsymbol{b}}_1$, where $\hat{c} = ||\widehat{D}^{-1/2}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)||$. Because $\hat{c}$ (> 0) does not affect discrimination, $\widehat{g}$ is essentially

$$\begin{aligned}\widehat{g}(\boldsymbol{x}) &= \left(\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_1 \\ &= \left(\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)\right)^T \widehat{\boldsymbol{b}}_1.\end{aligned}$$

For notational convenience we write $\widehat{\boldsymbol{q}} = (\hat{q}_1, \dots, \hat{q}_d)^T$ to denote either $\widehat{\boldsymbol{p}}_1 = (\hat{p}_1, \dots, \hat{p}_d)^T$ or $\widehat{\boldsymbol{b}}_1 = (\hat{b}_1, \dots, \hat{b}_d)^T$ as appropriate.

Let $\boldsymbol{X} = (X_1, \dots, X_d)^T$ be a random vector from one of the two classes $\mathcal{C}_\ell$. Put

$$\widehat{g}_\eta(\boldsymbol{X}) = \sum_{j=1}^{d}\left(X_j - \frac{1}{2}(\hat{\mu}_{1j} + \hat{\mu}_{2j})\right)\hat{b}_j I(|\hat{q}_j| > \eta), \tag{5.1}$$

where $\hat{\mu}_{\ell j}$ is jth components of $\widehat{\boldsymbol{\mu}}_\ell = (\hat{\mu}_{\ell 1}, \dots, \hat{\mu}_{\ell d})^T$, I is the indicator function and $\eta > 0$ is an appropriate threshold.

We interpret (5.1) in the following way. We first sort the features, that is, the variables of $\boldsymbol{X}$ in decreasing order of the absolute value of the components $\hat{q}_j$ of $\widehat{\boldsymbol{q}}$, and then consider the first m features to classify the data.

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

$$|\hat{q}_{i_1}| \geq |\hat{q}_{i_2}| \geq \cdots \geq |\hat{q}_{i_m}| \geq \cdots \geq |\hat{q}_{i_d}| \geq 0. \tag{5.2}$$

For the naive canonical correlation matrix $\widehat{C}$, $\widehat{\boldsymbol{b}}_1$ is the *naive* version of the canonical correlation vector. In Subsection 4.3, $\widehat{\boldsymbol{b}}_1$ is the direction vector for the HDLSS naive Bayes rule; $\widehat{\boldsymbol{b}}_1$ therefore plays a dual role of ranking vector for variable selection, and of direction vector for the naive Bayes discriminant function. We summarize our classification method based on feature selection with $\widehat{\boldsymbol{b}}_1$ in Steps 1–5 below.

Classification and variable ranking based on $\widehat{\boldsymbol{b}}_1$

Step 1 Calculate $\widehat{\boldsymbol{b}}_1$.

Step 2 Sort the components of $\widehat{\boldsymbol{b}}_1$ in descending order of their absolute values as in (5.2):

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

Step 3 Apply the permutation $\tau : \{1, 2, ..., d\} \to \{i_1, i_2, ..., i_d\}$ to the rows of X, and to $\widehat{\boldsymbol{b}}_1$, and then put $\widehat{\boldsymbol{b}}_1 \leftarrow \tau(\widehat{\boldsymbol{b}}_1)$ and $X \leftarrow \tau(X)$.

Step 4 Find the best truncation $\hat{m}$ of (4.3) in Fan and Fan (2008):

$$\hat{m} = \operatorname*{arg\,max}_{1 \leq m \leq d} \frac{1}{\hat{\lambda}_{\max}(R_m)} \frac{\left[\sum_{j=1}^{m} (\hat{\mu}_{1j} - \hat{\mu}_{2j})^2/\hat{\sigma}_{jj} + m(1/n_2 - 1/n_1)\right]^2}{nm/(n_1 n_2) + \sum_{j=1}^{m} (\hat{\mu}_{1j} - \hat{\mu}_{2j})^2/\hat{\sigma}_{jj}},$$

where R_m is the correlation matrix of the truncated observations.

Step 5 Classify a new datum $\boldsymbol{X}$ by

1. putting $\boldsymbol{X} \leftarrow \tau(\boldsymbol{X})$, and
2. assigning $\boldsymbol{X}$ to class $\mathcal{C}_1$ if

$$\widehat{g}_{\hat{m}}(\boldsymbol{X}) = \sum_{i=1}^{\hat{m}} \left(X_i - \frac{1}{2}(\hat{\mu}_{1i} + \hat{\mu}_{2i}) \right) \hat{b}_i > 0. \tag{5.3}$$

We refer to the classification of the five steps above as the *NAive Canonical Correlation* (NACC) approach, thus acknowledging the fact that Mardia et al. (1979) call $\widehat{\boldsymbol{b}}_1$ a canonical correlation vector. For the ranking vector $\widehat{\boldsymbol{q}} = \widehat{\boldsymbol{p}}_1$ in (5.2), the rule (5.1) becomes

$$\begin{aligned} \widehat{g}_m(\boldsymbol{X}) &= \sum_{j=1}^{m} \left(X_{i_j} - \frac{1}{2}(\hat{\mu}_{1i_j} + \hat{\mu}_{2i_j}) \right) \hat{b}_{i_j} \\ &= \frac{1}{\hat{c}} \sum_{j=1}^{m} \left(X_{i_j} - \frac{1}{2}(\hat{\mu}_{1i_j} + \hat{\mu}_{2i_j}) \right) \frac{\hat{\alpha}_{i_j}}{\hat{\sigma}_{i_{jj}}} \end{aligned}$$

for some $m \in \{1, \ldots, d\}$, where the X_{i_j} are the sorted entries of $\boldsymbol{X}$, $\hat{\alpha}_j = \hat{\mu}_{1j} - \hat{\mu}_{2j}$, and $\hat{\sigma}_{jj}$ is the jth diagonal element of $\widehat{D}$ given by

$$\hat{\sigma}_{jj} = \frac{1}{n-2} \left\{ (n_1 - 1) S_{1j}^2 + (n_2 - 1) S_{2j}^2 \right\},$$

and

$$S_{\ell j}^2 = \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} (X_{\ell ij} - \hat{\mu}_{\ell j})^2, \quad \ell = 1, 2; \; j = 1, \ldots, d.$$

A comparison of the feature selection induced by (5.2) with the *Feature Annealed Independence Rules* (FAIR), which Fan and Fan (2008) propose, shows that their selection is induced by the two sample t-statistics, namely, for jth variable

$$T_j = \frac{\hat{\alpha}_j}{\sqrt{\hat{\sigma}_{jj} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}. \tag{5.4}$$

A comparison of (4.6) and (5.4) yields that $T_j = C_n \hat{p}_j$ for all j, where the constant C_n depends on the sample size. Hence, feature selection or variable ranking based on (5.2) is essentially equivalent to FAIR, and the eigenvector $\widehat{\boldsymbol{p}}_1$ of the naive canonical correlation matrix therefore offers a natural explanation for the variable selection in FAIR.

The classifications NACC and FAIR differ in that the initial ranking is based on different vectors; NACC uses $\widehat{\boldsymbol{b}}_1$, while variable selection in FAIR is based on $\widehat{\boldsymbol{p}}_1$. As a consequence the order of the variables and the 'optimal' number of variables will differ in the two approaches.

5.2 Feature Selection in Multi-class

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 et al. (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 et al. (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

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

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 $\widehat{\boldsymbol{c}}$.

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 $X_{j_1}, \ldots, X_{j_{\hat{m}}}$.

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 $\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{p}}_{K-1}$ or $\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_{K-1}$ 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 5.1. The ranking of the variables is then inherited from the ranking vector obtained by combining these reference vectors.

Table 5.1: The ranking vector $\widehat{\boldsymbol{c}}$ associated with NACC and FAIR for K classes.

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

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 4.4. 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 (4.2) with feature selection becomes

$$\widehat{g}_{\hat{m}}\left(\boldsymbol{X}\right) = \underset{\ell\in\{1,\ldots,K\}}{\arg\min} \left(\boldsymbol{X}_{\hat{m}} - \widehat{\boldsymbol{\mu}}_{\ell,\hat{m}}\right)^T \widehat{B}_{\hat{m}}(\widehat{B}_{\hat{m}}\widehat{D}_{\hat{m}}\widehat{B}_{\hat{m}}^T)^{-1}\widehat{B}_{\hat{m}}^T \left(\boldsymbol{X}_{\hat{m}} - \widehat{\boldsymbol{\mu}}_{\ell,\hat{m}}\right), \tag{5.5}$$

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

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

is the $\hat{m} \times \hat{m}$ left-upper submatrix of $\widehat{D}$. Let $[\ \cdot\ ]_m$ denote the vector or matrix containing the first m features, that is, for $\boldsymbol{a}_\ell = [a_{\ell 1}, \ldots, a_{\ell d}]^T \in \mathbb{R}^d$ and $A = [\ \boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_K\ ] \in \mathbb{R}^{d\times K}$,

$$[\ \boldsymbol{a}_\ell\ ]_m = [a_{\ell 1}, \ldots, a_{\ell m}]^T \in \mathbb{R}^m, \quad [\ A\ ]_m = [\ [\ \boldsymbol{a}_1\ ]_m \ \cdots \ [\ \boldsymbol{a}_K\ ]_m\ ] \in \mathbb{R}^{m\times K}.$$

Then, we summarize the new classification method based on feature selection in the following steps:

Feature Selection Algorithm M-NACC (M-FAIR)

Step 1 Calculate $\widehat{P} = [\ \widehat{\boldsymbol{p}}_1\ \cdots\ \widehat{\boldsymbol{p}}_{K-1}\ ]$ and $\widehat{B} = [\ \widehat{\boldsymbol{b}}_1\ \cdots\ \widehat{\boldsymbol{b}}_{K-1}\ ]$.

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

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

where $\hat{c}_j$ is based on $\widehat{B}$ for M-NACC, and on $\widehat{P}$ for M-FAIR (see Table 5.1).

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 Find the best truncation $\hat{m}$ based on THEOREM 4.4:

$$\hat{m} = \underset{K-1 \leq m \leq d}{\arg\max} \sum_{\substack{k,\alpha=1 \\ k \neq \alpha}}^{K} \frac{n_k n_\alpha}{n^2} \hat{d}_{k\alpha}(m), \tag{5.6}$$

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

$$\hat{d}_{k\alpha}(m) = \frac{\widehat{S}_{k\alpha,m}\widehat{\Gamma}\left(\widehat{\Gamma}^T\widehat{U}_m\widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^T\widehat{Q}^T_{k\alpha,m}}{\sqrt{\lambda_{\max}(\widehat{R}_m)}\sqrt{\widehat{Q}_{k\alpha,m}\widehat{\Gamma}\left(\widehat{\Gamma}^T\widehat{U}_m\widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^T\widehat{Q}^T_{k\alpha,m}}},$$

$$\widehat{U}_m = \left[\widehat{C}\right]^T_m \left[\widehat{C}\right]_m + \frac{m}{n}(I_K - N^{1/2}\mathbf{1}_K\mathbf{1}_K^T N^{1/2}),$$

$$\widehat{Q}_{k\alpha,m} = [\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha]^T_m \widehat{D}_m^{-1/2}\left[\widehat{C}\right]_m + \frac{m}{n}\boldsymbol{q}_{k\alpha}N^{-1/2},$$

$$\widehat{S}_{k\alpha,m} = \frac{1}{2}[\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha]^T_m \widehat{D}_m^{-1/2}\left[\widehat{C}\right]_m + \frac{m}{n}\widehat{\boldsymbol{s}}_{k\alpha}N^{-1/2},$$

$$\widehat{\boldsymbol{s}}_{k\alpha} = [\hat{s}_1, \ldots, \hat{s}_K], \qquad \hat{s}_\ell = \frac{n_\ell}{n} - \frac{1}{2}\{I(\ell = k) + I(\ell = \alpha)\}$$

and $\widehat{R}_m = (\hat{\rho}_{ij})_{1 \leq i,j \leq m}$ 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 (5.5) to assign $\boldsymbol{X}$ to one of the K classes.

Note that if $\hat{m} = d$, then (5.5) is nothing other than (4.2) 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 Simulation

In this section, we illustrate the results of Section 4 and Section 5 via numerical experiments. Subsection 6.1 includes numerical studies to validate the theories for 2-class problems discussed in Subsections 4.2, 4.3 and 4.4. Subsection 6.2 provides a numerical study for discriminant directions in the multi-class setting to confirm THEOREM 4.3 and COROLLARY 4.2. In Subsection 6.3, behavior of the upper bound of misclassification rate in a 3-class setting is investigated. In Subsection 6.4, comparative studies for the performance of feature selection methods discussed in Section 5 are given. For the multi-class setting, we compare the performance of M-FAIR and M-NACC with that of the *minimum redundancy maximum relevance* (MRMR) feature selection 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 Subsection 5.2. 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 (5.6) in MRMR, as well as in our new method.

6.1 Simulation I – Discriminant Direction and Misclassification Rate in 2-class

In Simulation I, our interests focus on the misclassification rate $W_1(\widehat{g}, \theta)$, and the angle between $\widehat{\boldsymbol{b}}_1^*$ and $\boldsymbol{b}_1^*$.

We generate n_ℓ ($\ell = 1, 2$) d-dimensional observations $\boldsymbol{X}_{\ell i} \overset{\text{i.i.d.}}{\sim} N_d(\boldsymbol{\mu}_\ell, \Sigma)$ for $d = 200$ and $d = 1000$. For each value of d, we choose $n = n_1 + n_2$ and $n_1 = n_2$ such that $n \le d$. The estimate of $\widehat{\boldsymbol{b}}_1^{*T} \boldsymbol{b}_1^*$ is obtained as the sample mean over 1000 iterations. Similarly, the estimate of $W_1(\widehat{g}, \theta)$ is calculated as the average of the leave-one out cross-validation on the 1000 iterations.

In Simulation I, we take $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\mu}_2 = t\boldsymbol{1} = [t, \ldots, t]^T$, for $t > 0$. The covariance matrix $\Sigma = (\sigma_{ij})$ has an *autoregressive* (AR) covariance structure, with $\sigma_{ij} = \rho^{|i-j|}$. For this example we take $\rho = 0.6$. Then we can see that

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = dt^2$$

from $D^{-1} = I_d$. Thus, for $t = 1$, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = d$. If we take $C_d = d$ in (4.3), then the condition $d = o(nC_d)$ in COROLLARY 4.2 is satisfied. Therefore, the angle between $\widehat{\boldsymbol{b}}_1^*$ and $\boldsymbol{b}_1^*$ converges to 0.

On the other hand, if $t = 2/\sqrt{n}$, then we can evaluate

$$\frac{d}{n(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} = \frac{1}{4} \le \frac{d}{nC_d}$$

on (4.3). Thus, the condition $d = O(nC_d)$ of THEOREM 4.3 is satisfied, because we can take $C_d = 4d^\gamma$ for the sequence in CONDITION C such that $d = nd^\gamma$ for $\gamma \in (0, 1)$. It means that the angle between $\widehat{\boldsymbol{b}}_1^*$ and $\boldsymbol{b}_1^*$ does not converge to 0. Furthermore, we have $\sigma_{\min} = \sigma_{\max}$, $\xi \to 1/4$ and $\zeta \to 1$. If we set $\pi_1 = \pi_2 = 1/2$, then we have

$$\widehat{\boldsymbol{b}}_1^{*T} \boldsymbol{b}_1^* \ \xrightarrow{P}\ \frac{1}{\sqrt{2}}$$

from (4.15) and (4.16). In fact the angle between $\widehat{\boldsymbol{b}}_1^*$ and $\boldsymbol{b}_1^*$ will converge to 45 degrees.

Table 6.1 summarizes the results from Simulation I. The two columns pertaining to "consistent" relate the results under the assumptions of COROLLARY 4.2; the column 'Error' gives the estimated misclassification rate $W_1(\widehat{g}, \theta)$, and the column 'Degrees' lists the estimated angle

$$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*) = \frac{180}{\pi} \arccos(\widehat{\boldsymbol{b}}_1^{*T} \boldsymbol{b}_1^*)$$

Table 6.1: Simulation I-I results.

	Consistent		Not consistent	
	$d = 200$			
	Error	Degrees	Error	Degrees
$n = 6$	0.06733	61.18507	0.13767	64.44634
$n = 8$	0.00325	51.73060	0.04938	57.08087
$n = 10$	0.00000	45.18963	0.03200	53.64817
$n = 20$	0.00000	31.08961	0.04190	48.12596
$n = 30$	0.00000	25.33803	0.06390	47.01111
$n = 40$	0.00000	21.93963	0.08670	46.36257
$n = 50$	0.00000	19.67402	0.10786	46.05681
$n = 80$	0.00000	15.59922	0.15870	45.55593
$n = 100$	0.00000	13.89451	0.18283	45.55475
$n = 120$	0.00000	12.72426	0.20403	45.22594
$n = 150$	0.00000	11.36625	0.22650	45.19424
$n = 180$	0.00000	10.43391	0.24979	45.19436
$n = 200$	0.00000	9.88011	0.25898	45.04897
	$d = 1000$			
	Error	Degrees	Error	Degrees
$n = 30$	0.00000	25.48474	0.00037	47.20665
$n = 40$	0.00000	22.03805	0.00130	46.49089
$n = 50$	0.00000	19.73937	0.00310	46.19724
$n = 100$	0.00000	13.96957	0.02390	45.56207
$n = 150$	0.00000	11.43055	0.05001	45.36467
$n = 200$	0.00000	9.89451	0.07709	45.25663
$n = 250$	0.00000	8.85823	0.09941	45.25595
$n = 400$	0.00000	7.01443	0.15405	45.14002
$n = 500$	0.00000	6.26700	0.18030	45.06217
$n = 600$	0.00000	5.71869	0.20261	45.03024
$n = 750$	0.00000	5.12249	0.22683	45.06256
$n = 900$	0.00000	4.66969	0.24652	44.99821
$n = 1000$	0.00000	4.42869	0.25774	45.00964

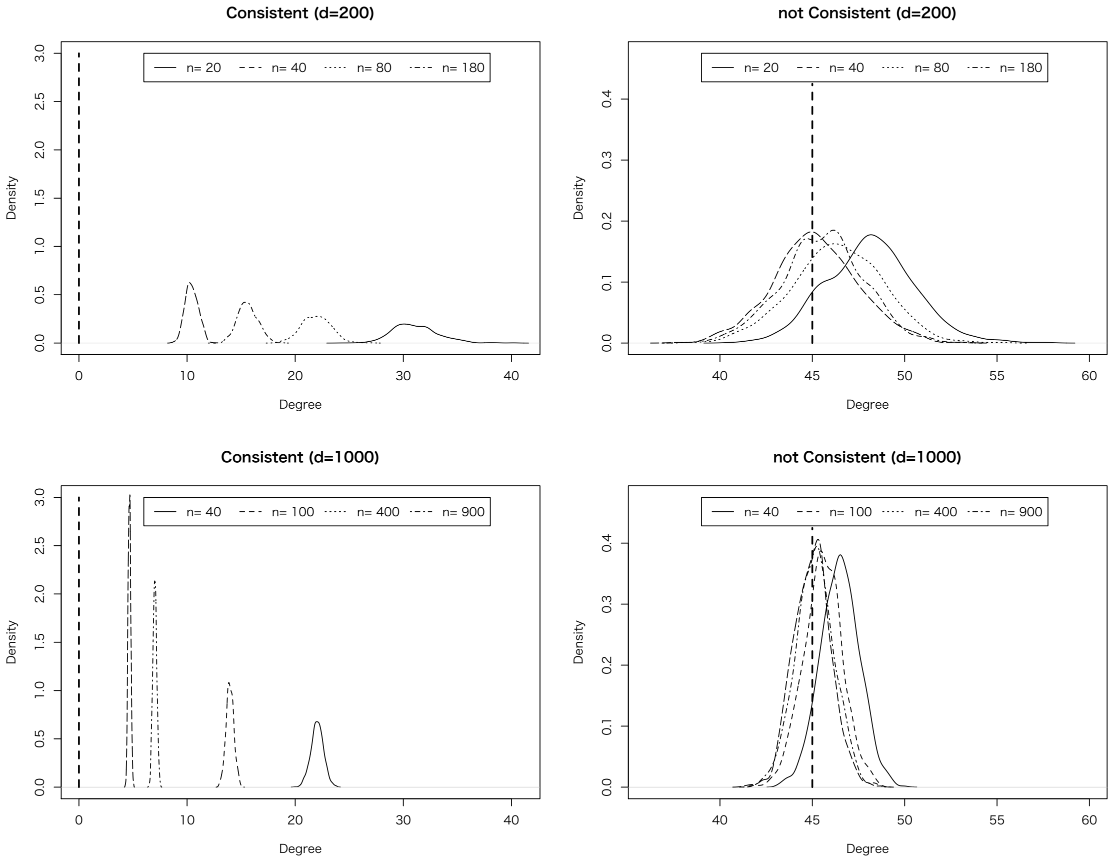


Figure 6.1: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_1^*$ and $\boldsymbol{b}_1^*$. The vertical dashed line is the asymptotic value derived in Theorem 4.3.

in degrees. The two columns "not consistent" show analogous results under the assumptions of Theorem 4.3.

We note that the angle in the "consistent" columns decreases to zero, as n and d increase. These results agree with Corollary 4.2. It is interesting to observe that the estimated angles for $n = 30,\ 50,\ 150,\ 200$ are very similar for both values of d. In contrast, for the "not consistent" columns, the angle clearly approaches 45 degrees, which agrees with Theorem 4.3. Figure 6.1 complements the results in Table 6.1: here we show kernel density estimates of the angles based on the 1000 iterations. The values of n in the top row (for $d = 200$) are $n = 20,\ 40,\ 80$ and 180, and the bottom row shows $n = 40,\ 100,\ 400$

and 900 for $d = 1000$. The left panels focus on the "consistent" columns in the Table 6.1, and the right panels show density estimates for the "not consistent" columns. These figures clearly illustrate the behavior of the angle as the sample size increases.

Returning to the "Error" columns in Table 6.1, it is noticeable that the misclassification rate is almost 0 in the "consistent" cases, and is largest when n and d are both large, that is, as $\widehat{\boldsymbol{b}}_1^*$ becomes inconsistent. For these latter results, $t = 2/\sqrt{n}$, and this choice of t makes the discrimination problem increasingly difficult as n increases.

6.2 Simulation II – Discriminant Directions in Multi-class

6.2.1 Simulation II-I

We focus on the investigation of the angle between $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\beta^*$ for simulated data from the normal distribution in the case of 3-class setting. In Simulation II-I and II-II, we generate n_ℓ d-dimensional observations $\boldsymbol{X}_{\ell i} \overset{\text{i.i.d.}}{\sim} N_d(\boldsymbol{\mu}_\ell, \Sigma)$ for $d = 600,\ 1200$ and $\ell = 1, \dots, K$. For each d and $\ell_1,\ \ell_2 \in \{1, \dots, K\}$, we choose $n = \sum_{\ell=1}^K n_\ell$ and $n_{\ell_1} = n_{\ell_2}$ such that $n \leq d$. For the simulated data we calculate $\widehat{\boldsymbol{b}}_1^*, \dots, \widehat{\boldsymbol{b}}_{K-1}^*$, the angles

$$\angle(\widehat{\boldsymbol{b}}_\alpha^*, \boldsymbol{b}_\beta^*) = \frac{180}{\pi} \arccos(\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^*) \quad (\alpha, \beta = 1, \dots, K-1)$$

and display the distributions of these angles based on 1000 iterations.

We take $\Sigma = (\sigma_{ij})$ to be the covariance of an AR structure. 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 $\min_{i \neq j}\ (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T D^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = 2d/n$ and

$$\frac{d}{n\left\{\min_{i \neq j}\ (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T D^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right\}} \leq \frac{d}{nC_d} \tag{6.1}$$

under (4.3). Thus, if we take $C_d = 2d^\gamma$ 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 4.3 is satisfied. We

Table 6.2: 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
Th.4.3	33.855	40.893	90.000	90.000
	$d = 1200$			
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
Th.4.3	33.855	40.893	90.000	90.000

obtain the following asymptotic angles in degrees $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*) \xrightarrow{P} 33.855$, $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*) \xrightarrow{P} 40.893$ and $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$, $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*) \xrightarrow{P} 90$ from (4.14) with $\sigma_{\max} = \sigma_{\min} = 1$.

Table 6.2 summarizes the results for Simulation II-I. The table shows that each angle approaches the theoretical value shown in Theorem 4.3. Figure 6.2 and Figure 6.3 depict 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 6.2 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 $n = 15,\ 30,\ 150,\ 300,\ 450$ and $d = 600$. Similarly, Figure 6.3 shows the results for $\angle(\widehat{\boldsymbol{b}}_\alpha^*, \boldsymbol{b}_\alpha^*)$ for $n = 15,\ 30,\ 150,\ 600,\ 900$ and $d = 1200$. These figures show that each angle converges as the sample size increases.

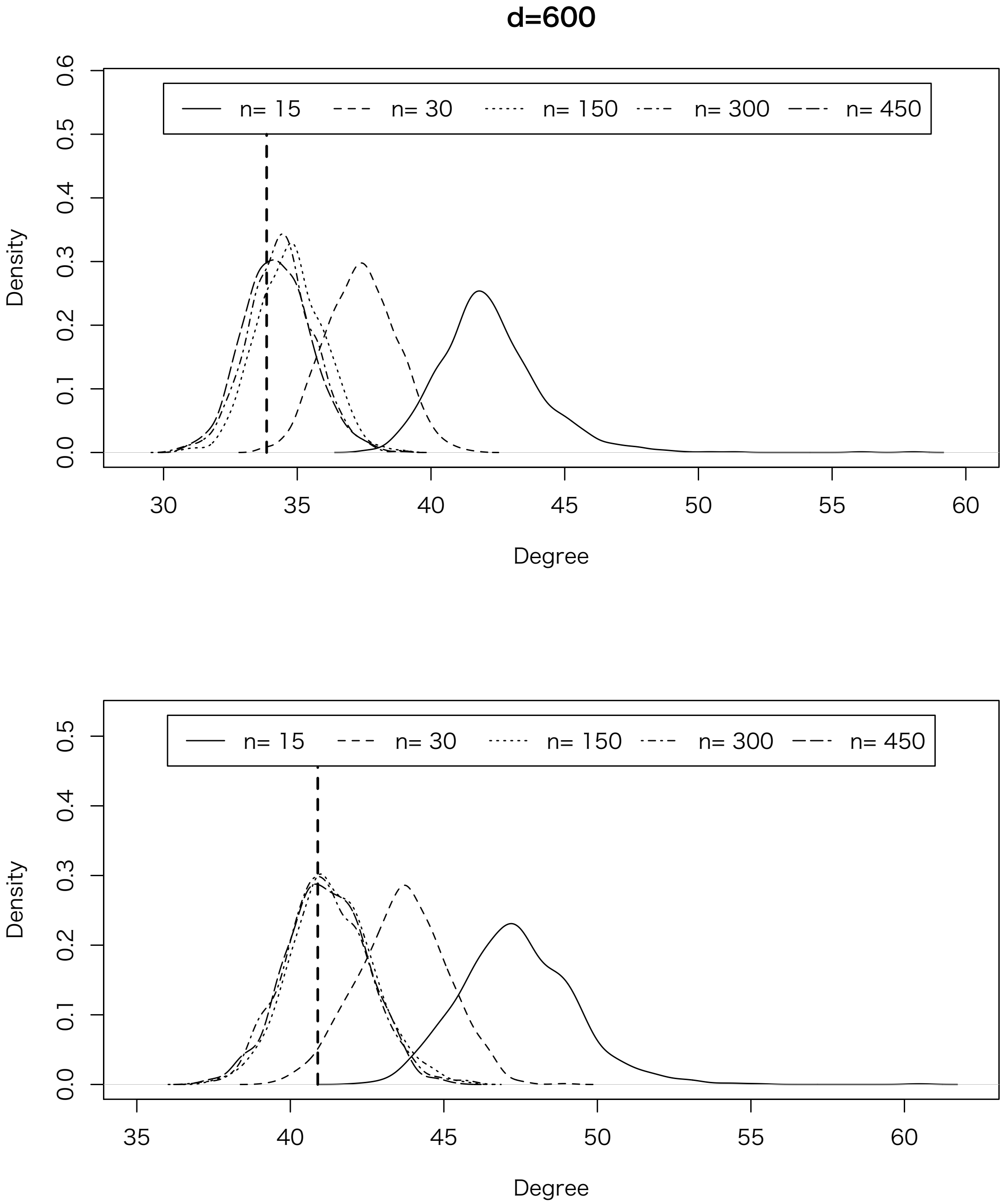


Figure 6.2: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\alpha^*$ for $d = 600$. The top panel and bottom panel are for $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$ and $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$, respectively. The vertical dashed line is the asymptotic value derived in Theorem 4.3.

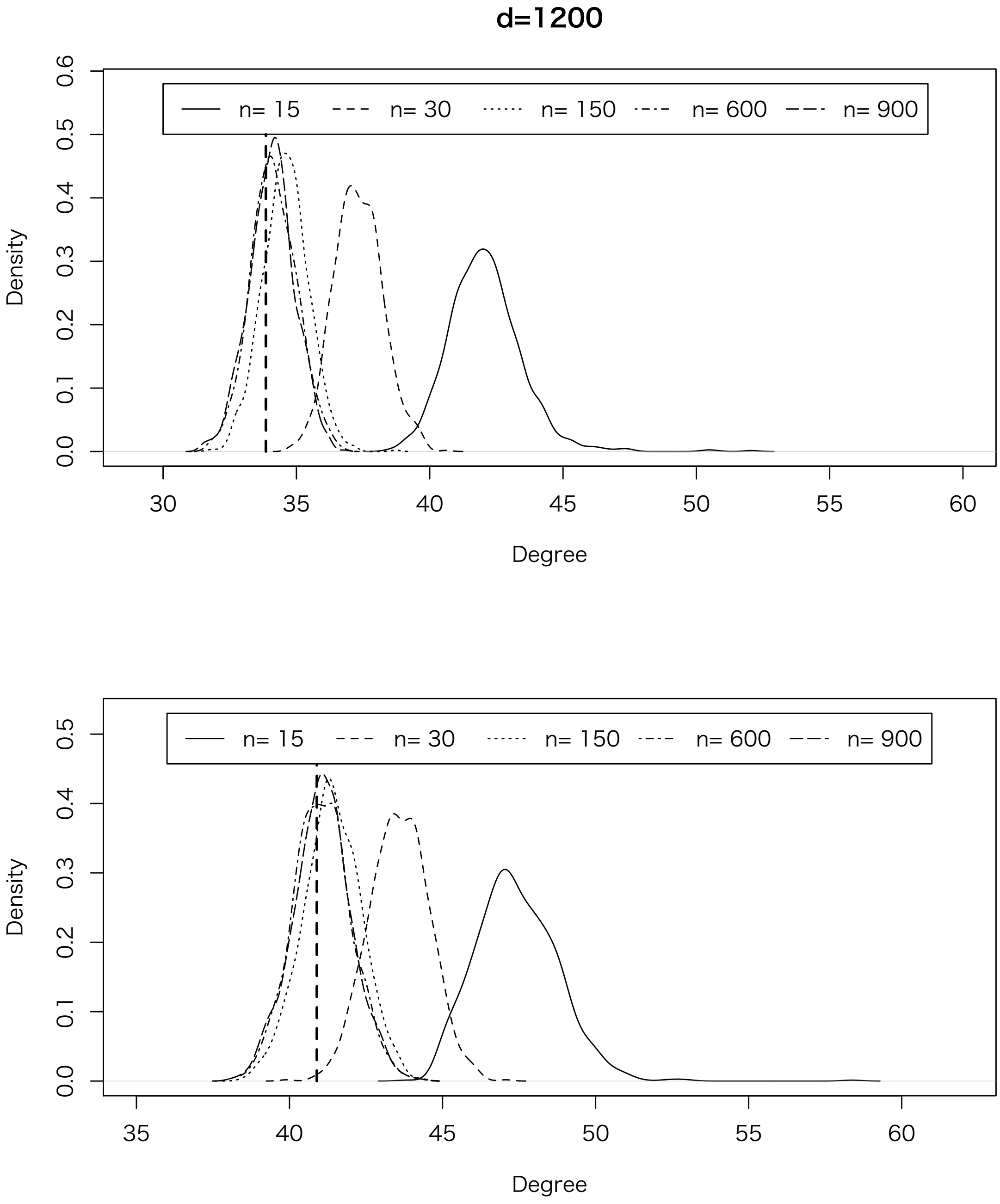


Figure 6.3: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_\alpha^*$ and $\boldsymbol{b}_\alpha^*$ for $d = 1200$. The description of each panel is the same as in Figure 6.2.

6.2.2 Simulation II-II

In Simulation II-II, we extend previous discussions to 4-class setting. The parameter settings of n, d and Σ are the same as that in Simulation II-I, but with $K = 4$. For the mean parameters $\boldsymbol{\mu}_\ell$, we utilize

$$[\boldsymbol{\mu}_1 \;\; \boldsymbol{\mu}_2 \;\; \boldsymbol{\mu}_3 \;\; \boldsymbol{\mu}_4] = \frac{10}{\sqrt{n}} \begin{bmatrix} \mathbf{1}_{d/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{d/4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{d/8} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{d/8} \end{bmatrix}.$$

In this setting, we can also confirm that if we take the sequence $C_d = 25d^\gamma$ such that $d = nd^\gamma$ for $\gamma \in (0, 1)$, then the condition $d = O(nC_d)$ is satisfied from $\min_{i \neq j}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T D^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = 25d/n$ and the inequality (6.1) on (4.3).

We obtain the following asymptotic angles in degrees $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*) \xrightarrow{P} 17.136$, $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*) \xrightarrow{P} 23.871$, $\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*) \xrightarrow{P} 29.496$ and $\angle(\widehat{\boldsymbol{b}}_\alpha^*, \boldsymbol{b}_\beta^*) \xrightarrow{P} 90$ (for $\alpha \neq \beta$) from (4.14) with $\sigma_{\max} = \sigma_{\min}$.

Table 6.3 summarizes the results for Simulation II-II. Similar to Simulation II-I, the table shows that each angle approaches the theoretical value shown in Theorem 4.3. Figure 6.4 and Figure 6.5 depict the behavior of $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$, $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$ and $\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$ using kernel density estimates. The top panel of Figure 6.4 shows the results for $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$, the middle panel shows the results for $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$, and the bottom panel shows the results for $\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$ for $n = 20,\ 40,\ 120,\ 200,\ 400$ and $d = 600$. Figure 6.5 shows the results for the same situation for $n = 20,\ 40,\ 200,\ 600,\ 1000$ and $d = 1200$. These figures also show that each angle approaches the theoretical value as the sample size increases.

6.2.3 Simulation II-III

In Simulation II-III, we use simulated data with a non-normal distribution. We consider d-dimensional observation vectors $\boldsymbol{X}_{\ell i}$ from a multivariate t-distribution

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \nu, \boldsymbol{\mu}, \Sigma) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}} |\Sigma|^{-1/2} \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)^{-(\nu+d)/2},$$

where $\nu(\geq 3)$ is the degrees of freedom and Γ is the gamma function. Note that the multivariate t-distribution approaches the normal distribution as $\nu \to \infty$.

Table 6.3: Multivariate normal distribution; Average of 1000 simulated angles between $\widehat{\boldsymbol{b}}_i^*$ and $\boldsymbol{b}_j^*$ in degrees.

	$d = 600$								
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_2^*)$
20	27.93698	32.88319	36.70527	90.28594	90.15521	90.13317	89.97885	90.28116	90.01143
40	22.20510	28.32794	33.02367	90.19837	89.80088	90.14476	90.00066	89.84188	90.14732
120	18.98815	25.93988	31.07103	90.15718	89.99979	90.13416	89.99262	90.24643	90.05794
200	18.35978	25.53957	30.68630	90.15937	90.23564	90.03319	89.99190	90.22180	89.78367
400	17.91839	25.15996	30.41560	90.20726	89.86320	90.02854	90.01554	89.85710	90.02932
600	17.70677	25.06091	30.29975	90.11280	90.07990	90.12690	90.07321	90.55446	90.27417
Th.4.3	17.13612	23.87130	29.49621	90.00000	90.00000	90.00000	90.00000	90.00000	90.00000

	$d = 1200$								
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_2^*)$
20	27.87492	32.49854	36.49906	90.05864	89.99015	90.06845	89.98721	90.00854	90.21149
40	22.15112	27.95370	32.73800	90.30325	90.09344	90.05447	90.08002	90.17766	89.92379
120	18.80498	25.34015	30.63850	90.01228	89.96309	90.01109	89.98008	89.90441	90.03415
200	18.15262	24.88861	30.23269	90.08753	90.02069	90.05367	89.95875	89.97761	89.84134
400	17.69580	24.60009	29.97659	90.07216	90.00970	90.04104	90.01775	89.93575	90.34125
600	17.54744	24.43154	29.93620	90.14421	90.01787	90.09337	89.97496	90.21238	90.16605
800	17.50330	24.34083	29.94811	89.91361	90.10382	90.01094	90.05431	90.15182	89.90779
1000	17.43285	24.34355	29.80813	90.04453	89.99016	90.08898	89.99826	89.99953	90.20697
1200	17.39830	24.27680	29.85214	89.84063	90.03168	89.91411	89.94010	89.93728	90.07480
Th.4.3	17.13612	23.87130	29.49621	90.00000	90.00000	90.00000	90.00000	90.00000	90.00000

Table 6.4: Multivariate t-distribution; Average of 1000 simulated angles between $\widehat{\boldsymbol{b}}_i^*$ and $\boldsymbol{b}_j^*$ in degrees.

					$d = 600$				
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_2^*)$
20	32.95057	37.16785	39.53338	90.32398	90.00747	90.13602	90.06861	90.40717	89.95316
40	27.19351	32.69349	36.13330	90.44932	89.84209	90.12248	90.08555	90.26261	90.22046
120	23.48950	29.92241	34.35752	90.16831	89.93733	90.04091	90.03137	90.42366	89.92071
200	22.59878	29.66543	34.36528	90.46910	89.97149	90.10762	89.95794	90.62251	90.24983
400	21.43108	28.62169	33.70222	90.13634	89.98875	90.12821	89.94322	90.38824	90.08033
600	21.33427	28.75900	33.96163	90.06981	90.09685	90.02053	89.99544	89.85872	89.59476
TH.4.3	19.59721	27.06708	33.15246	90.00000	90.00000	90.00000	90.00000	90.00000	90.00000

					$d = 1200$				
n	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_2^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_1^*)$	$\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_3^*)$	$\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_2^*)$
20	32.78811	36.79906	39.30384	90.33727	89.90701	90.15753	89.92097	90.85272	89.75525
40	27.18993	32.39060	35.94866	90.30272	90.01457	90.12140	90.03340	90.55022	89.96178
120	23.38265	29.74863	34.27585	90.29096	89.94708	90.08983	90.00765	90.50505	89.92787
200	22.44701	29.20586	33.93244	90.23571	89.85847	90.07569	90.04816	90.48160	90.13325
400	21.47640	28.30986	33.45970	89.91540	90.00603	90.06548	90.03354	90.05541	90.21667
600	21.19932	28.29647	33.41617	90.07688	89.93492	90.03259	89.93899	90.37802	90.25378
800	20.93186	28.13916	33.50695	90.10032	89.92983	89.95264	89.98382	90.26792	89.78653
1000	20.84791	28.21556	33.64170	90.12972	90.03216	90.09332	89.98724	90.47703	90.25456
1200	20.83902	28.15838	33.57061	90.24657	89.96947	89.98281	89.93986	90.18540	89.89433
TH.4.3	19.59721	27.06708	33.15246	90.00000	90.00000	90.00000	90.00000	90.00000	90.00000

For the population covariance matrix $\Sigma = (\sigma_{ij})$, we choose a moving average covariance structure $\sigma_{ij} = \rho^{|i-j|} I(|i-j| \leq 1)$ for $1 \leq i, j \leq d$, where $\rho = 0.5$. We take $\nu = 3$ and

$$[\boldsymbol{\mu}_1 \ \ \boldsymbol{\mu}_2 \ \ \boldsymbol{\mu}_3 \ \ \boldsymbol{\mu}_4] = \frac{15}{\sqrt{n}} \begin{bmatrix} \mathbf{1}_{d/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{d/4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{d/8} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{d/8} \end{bmatrix}.$$

Then we can confirm that the condition $d = O(nC_d)$ is satisfied from

$$\frac{d}{n\left\{\min_{i \neq j} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T D^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right\}} = \frac{4}{75} \leq \frac{d}{nC_d}.$$

The setting of (n_ℓ, d) and purpose of Simulation II-III are the same as Simulation II-II. We can calculate $\angle(\widehat{\boldsymbol{b}}_1^*, \boldsymbol{b}_1^*) \xrightarrow{P} 19.597$, $\angle(\widehat{\boldsymbol{b}}_2^*, \boldsymbol{b}_2^*) \xrightarrow{P} 27.067$, $\angle(\widehat{\boldsymbol{b}}_3^*, \boldsymbol{b}_3^*) \xrightarrow{P} 33.152$ and $\angle(\widehat{\boldsymbol{b}}_\alpha^*, \boldsymbol{b}_\beta^*) \xrightarrow{P} 90$ (for $\alpha \neq \beta$) from (4.14).

Table 6.4 summarizes the results for Simulation II-III. The table shows that each angle approaches the theoretical value shown in Theorem 4.3. The description of Figure 6.6 and Figure 6.7 are the same as in Figure 6.4 and Figure 6.5. Compared with Figure 6.4 and Figure 6.5, we see from degrees and density of each panel that the convergence speed of $\angle(\widehat{\boldsymbol{b}}_\alpha^*, \boldsymbol{b}_\alpha^*)$ with multivariate t-distribution is slower than that with multivariate normal distribution. However, we can see that each angle converges as the sample size increases also in this situation.

6.3 Simulation III – Upper Bound for Misclassification Rate

To appreciate the usefulness of the upper bounds which we derived in Theorem 4.4 and Corollary 4.3, 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 $\Sigma = (\sigma_{ij})$ to be an AR structure. For this example we take $\rho = 0.6$. 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 6.8 shows estimates of the misclassification rate $W_1(\widehat{g}, \theta)$ and of the upper bounds given in Theorem 4.4 and Corollary 4.3. Both figure panels show that the

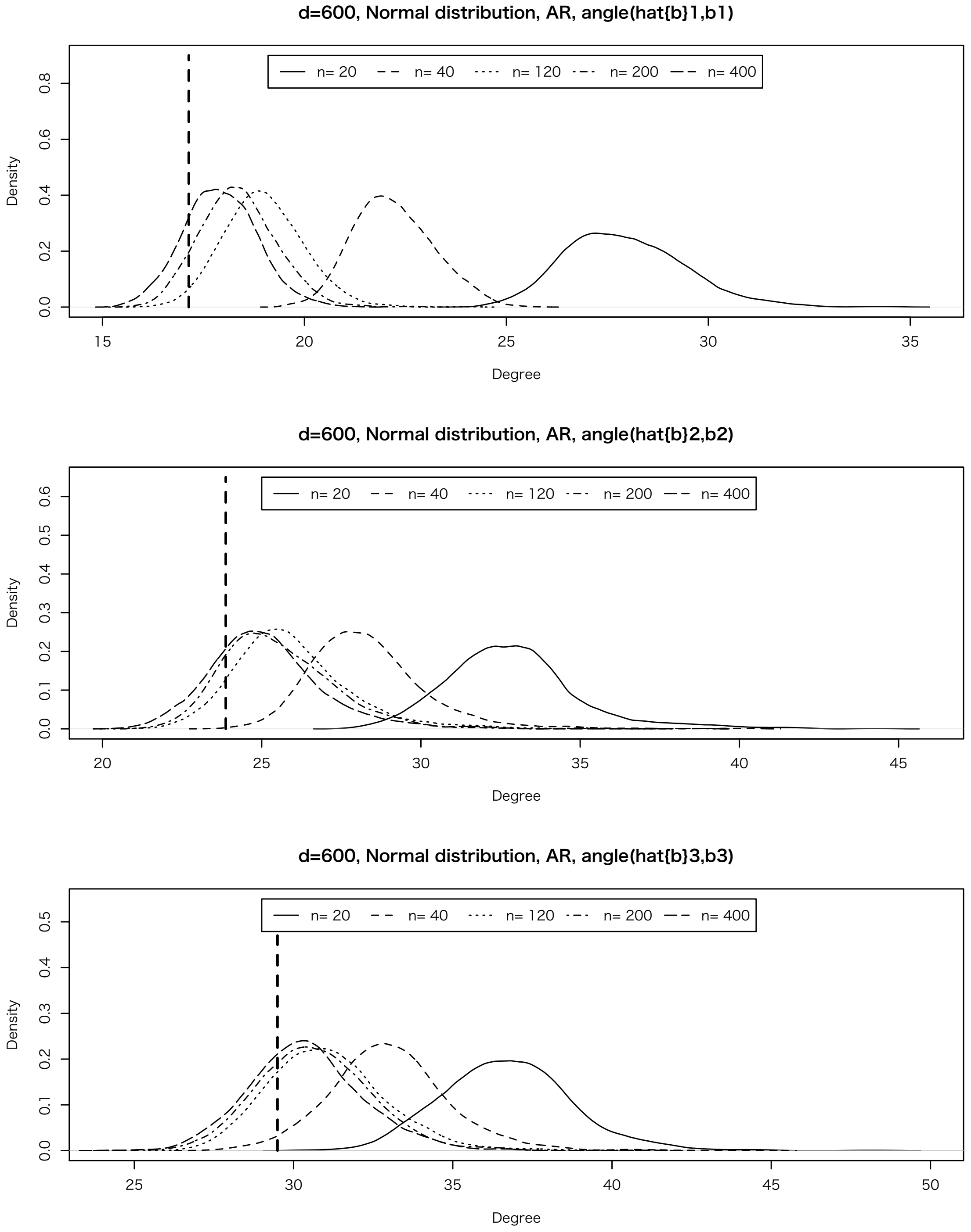


Figure 6.4: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}^*_\alpha$ and $\boldsymbol{b}^*_\alpha$ for $d = 600$. The top panel, middle panel and bottom panel are for $\angle(\widehat{\boldsymbol{b}}^*_1, \boldsymbol{b}^*_1)$, $\angle(\widehat{\boldsymbol{b}}^*_2, \boldsymbol{b}^*_2)$ and $\angle(\widehat{\boldsymbol{b}}^*_3, \boldsymbol{b}^*_3)$, respectively. The vertical dashed line is the asymptotic value derived in Theorem 4.3.

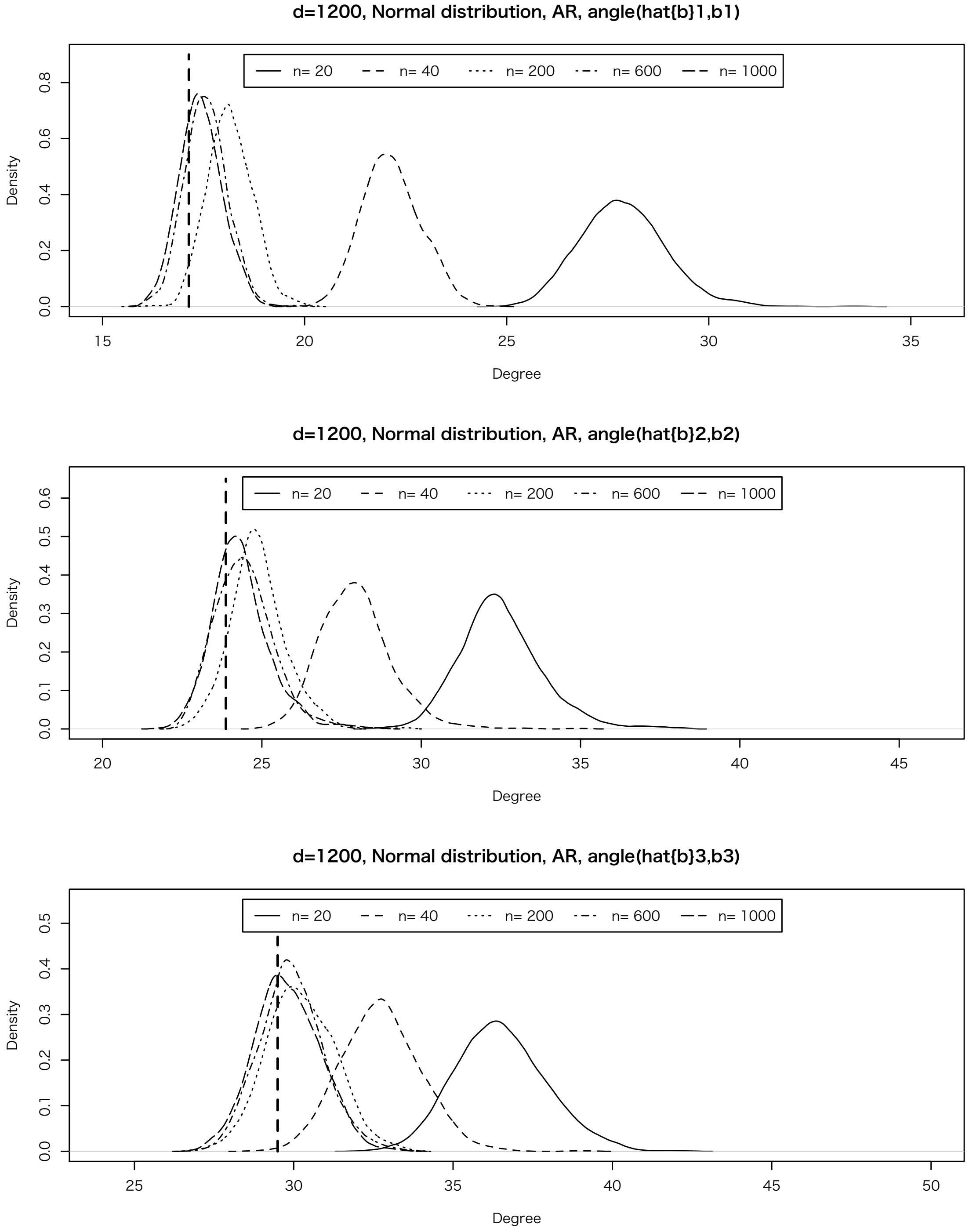
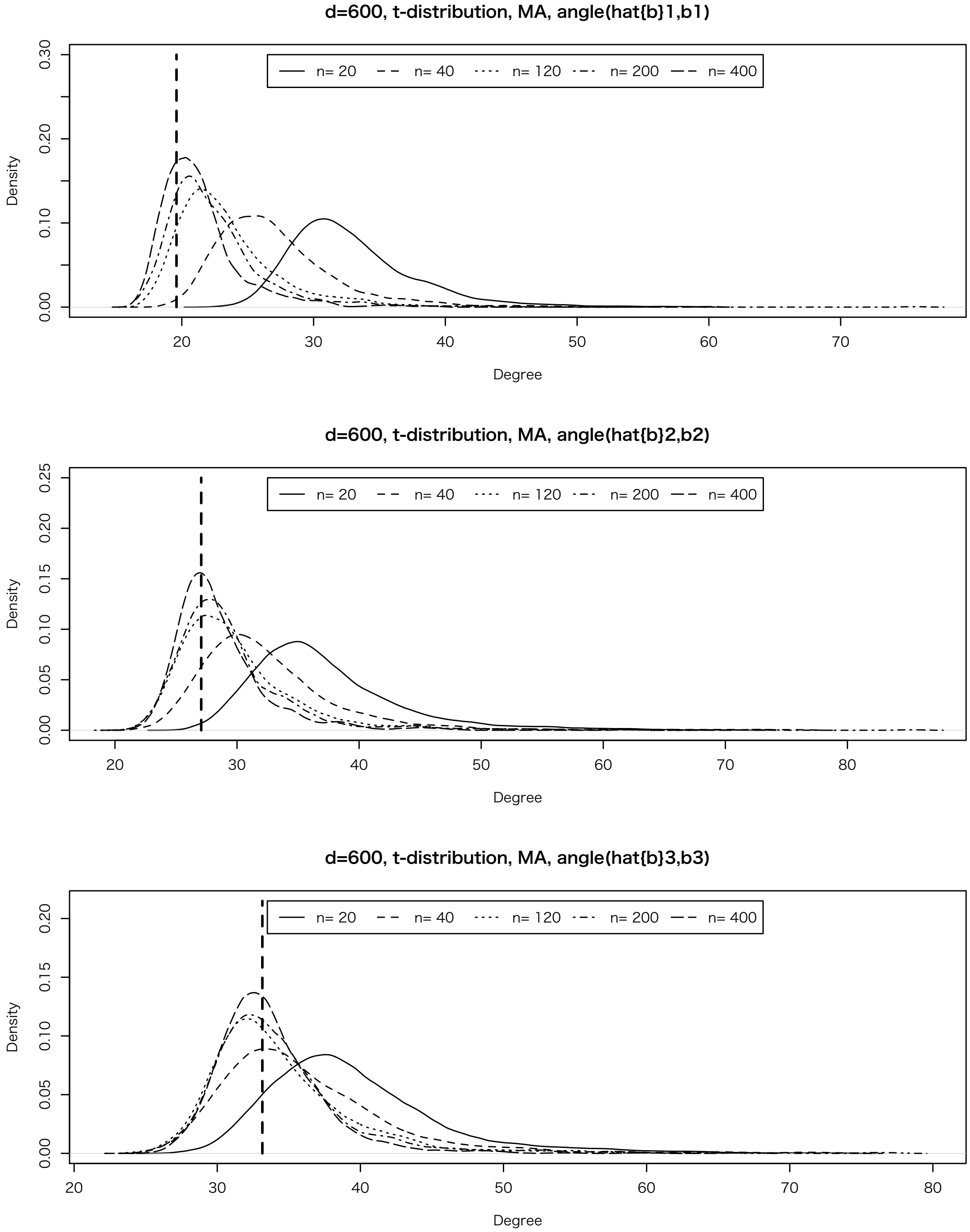


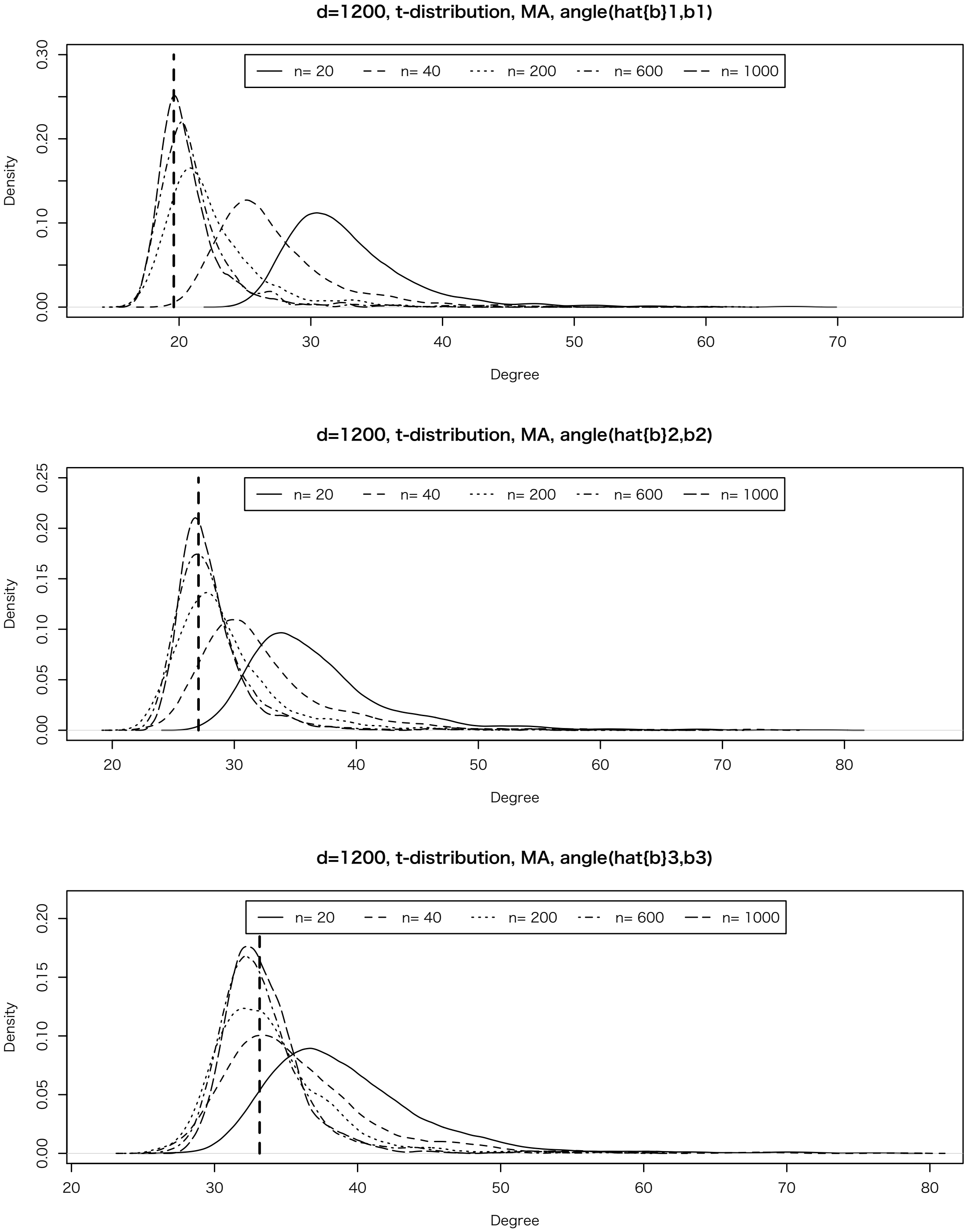
Figure 6.5: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_{\alpha}^{*}$ and $\boldsymbol{b}_{\alpha}^{*}$ for $d = 1200$. The description of each panel is the same as in Figure 6.4.


d=600, t-distribution, MA, angle(hat{b}1,b1)

d=600, t-distribution, MA, angle(hat{b}2,b2)

d=600, t-distribution, MA, angle(hat{b}3,b3)

Figure 6.6: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_{\alpha}^{*}$ and $\boldsymbol{b}_{\alpha}^{*}$ for $d = 600$. The description of each panel is the same as in Figure 6.4.


d=1200, t-distribution, MA, angle(hat{b}1,b1)

d=1200, t-distribution, MA, angle(hat{b}2,b2)

d=1200, t-distribution, MA, angle(hat{b}3,b3)

Figure 6.7: Kernel density estimates of angles in degrees between $\widehat{\boldsymbol{b}}_{\alpha}^{*}$ and $\boldsymbol{b}_{\alpha}^{*}$ for $d = 1200$. The description of each panel is the same as in Figure 6.4.

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.

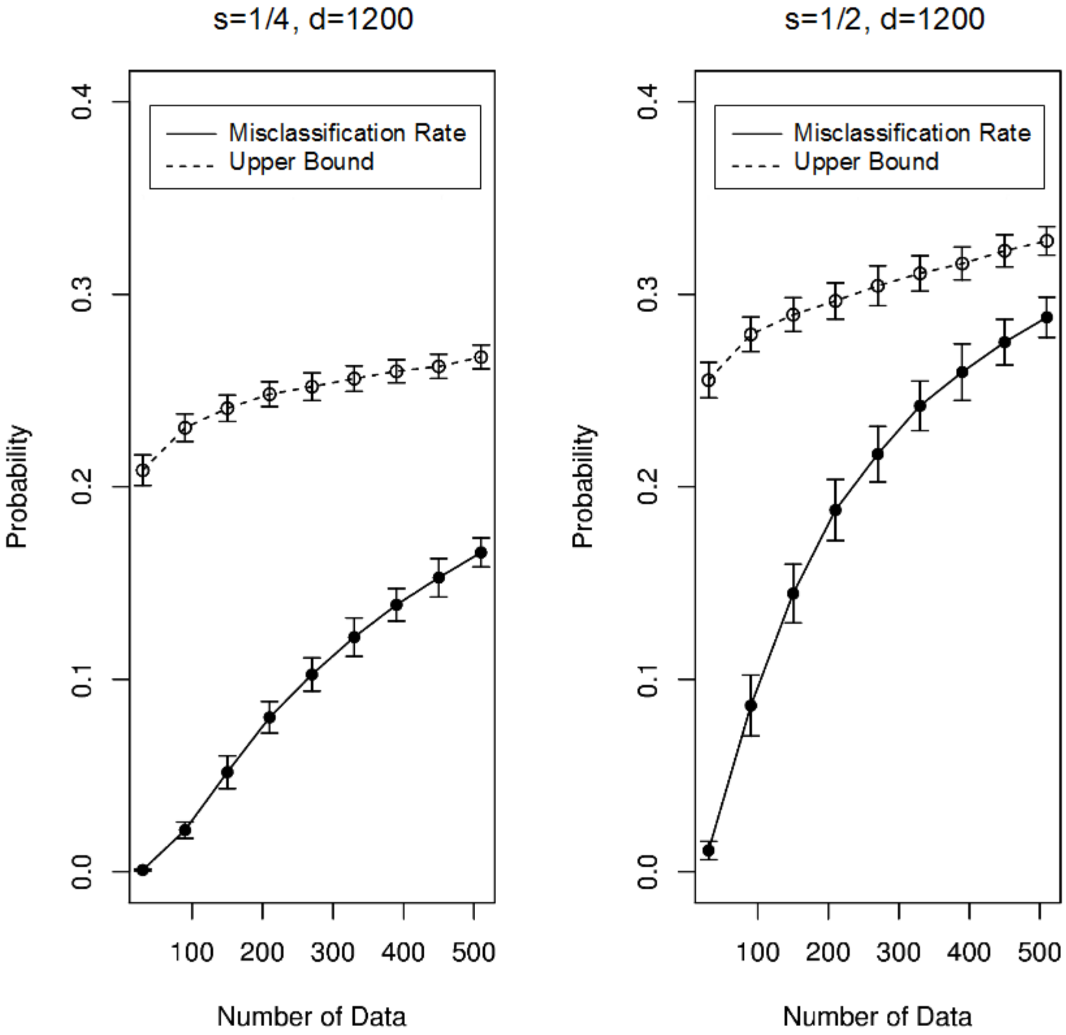


Figure 6.8: Mean upper bound for the misclassification rates over 100 simulations shown together with intervals of one standard deviation. Left panel: $d = o(nC_d)$; right panel: $d = O(nC_d)$.

6.4 Simulation IV – Feature Selection

6.4.1 Simulation IV-I

Simulation IV-I focuses on the performance of $\widehat{g}_{\hat{m}}$ of (5.3). The FAIR approach of Fan and Fan (2008) and the NACC approach share the vector $\widehat{\boldsymbol{b}}_1$ for discrimination, but base their feature selection on different ranking vectors: FAIR essentially chooses features based on (4.6), while NACC selects features based on $\widehat{\boldsymbol{b}}_1$.

For Simulation IV-I the parameters $\boldsymbol{\mu}_\ell$ and Σ_0 are

$$\boldsymbol{\mu}_2 = \mathbf{0} \in \mathbb{R}^d, \quad \boldsymbol{\mu}_1 = [\mu_{11}, \dots, \mu_{1d}]^T \in \mathbb{R}^d, \quad \mu_{1j} = \begin{cases} j/4 & j \in \{10, 20, 30\}, \\ 0 & \text{otherwise}, \end{cases}$$

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

and Σ is an AR structure with $\rho = -0.6$. The mean parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ show that only the 10th, 20th and 30th features are large, so we expect to select these features from the simulated data. Further we note that the diagonal elements of the covariance matrix Σ_0 are monotonically increasing. The observations about the large features and the behaviour of Σ_0 allow us to compare the performance of the FAIR and NACC approaches under a non-homogeneous variance structure of the features.

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

The estimates of the misclassification rates are tabulated in Table 6.5, with standard deviation in parentheses. The column $d = 200$ of Table 6.5 shows that the error probabilities of NACC are smaller than the corresponding values for FAIR except for $n = 10$, where FAIR wins. For $d = 1000$, the superiority of NACC over FAIR is apparent, especially for $n = 50, 100, 150, 200$. The results show the merit of feature selection with $\widehat{\boldsymbol{b}}_1$ over that with the vector $\widehat{\boldsymbol{p}}_1$ of FAIR.

So far we have compared the misclassification rates of the two approaches. We now look at the specific features that are selected in the simulations for (d, n) pairs. Figure

Table 6.5: Estimated value of the misclassification rate.

	$d = 200$		$d = 1000$	
	NACC	FAIR	NACC	FAIR
$n = 6$	**0.53167**	0.53500	**0.55167**	0.56167
	(0.26025)	(0.25436)	(0.23414)	(0.25031)
$n = 10$	0.51100	**0.50300**	**0.57300**	0.62500
	(0.20395)	(0.21670)	(0.24240)	(0.22847)
$n = 30$	**0.33933**	0.35700	**0.42233**	0.44433
	(0.13848)	(0.15500)	(0.17240)	(0.14688)
$n = 50$	**0.23360**	0.25940	**0.35580**	0.40420
	(0.08989)	(0.10830)	(0.13536)	(0.11691)
$n = 100$	**0.17510**	0.19280	**0.21360**	0.30550
	(0.04237)	(0.04854)	(0.08423)	(0.09312)
$n = 150$	**0.17993**	0.18640	**0.17667**	0.24360
	(0.03349)	(0.03718)	(0.04020)	(0.08258)
$n = 200$	**0.17035**	0.17295	**0.16815**	0.19155
	(0.02616)	(0.02658)	(0.02708)	(0.04728)

6.9 shows the frequency of the selected features over 100 simulations. The top panels show the results for $(d, n) = (200, 180)$, and the bottom panels show similar results for $(d, n) = (1000, 180)$. The left panels relate to NACC, and the right panels to FAIR. The figures show clearly that NACC correctly picks the large variables 10, 20 and 30 most of the time, while FAIR selects many other features, typically features with a large variance. The feature selection of NACC is based on $\widehat{\boldsymbol{b}}_1$ and thus on $(\hat{\mu}_{1j} - \hat{\mu}_{2j})/\hat{\sigma}_{jj}$. Our results suggest that feature selection with $\widehat{\boldsymbol{b}}_1$ captures the data structure better than FAIR, in particular for larger values of d.

The results of Simulation IV-I illustrate the superiority of NACC over FAIR in two ways: the misclassification rates are mostly smaller, and feature selection is more pertinent, especially for higher values of d.

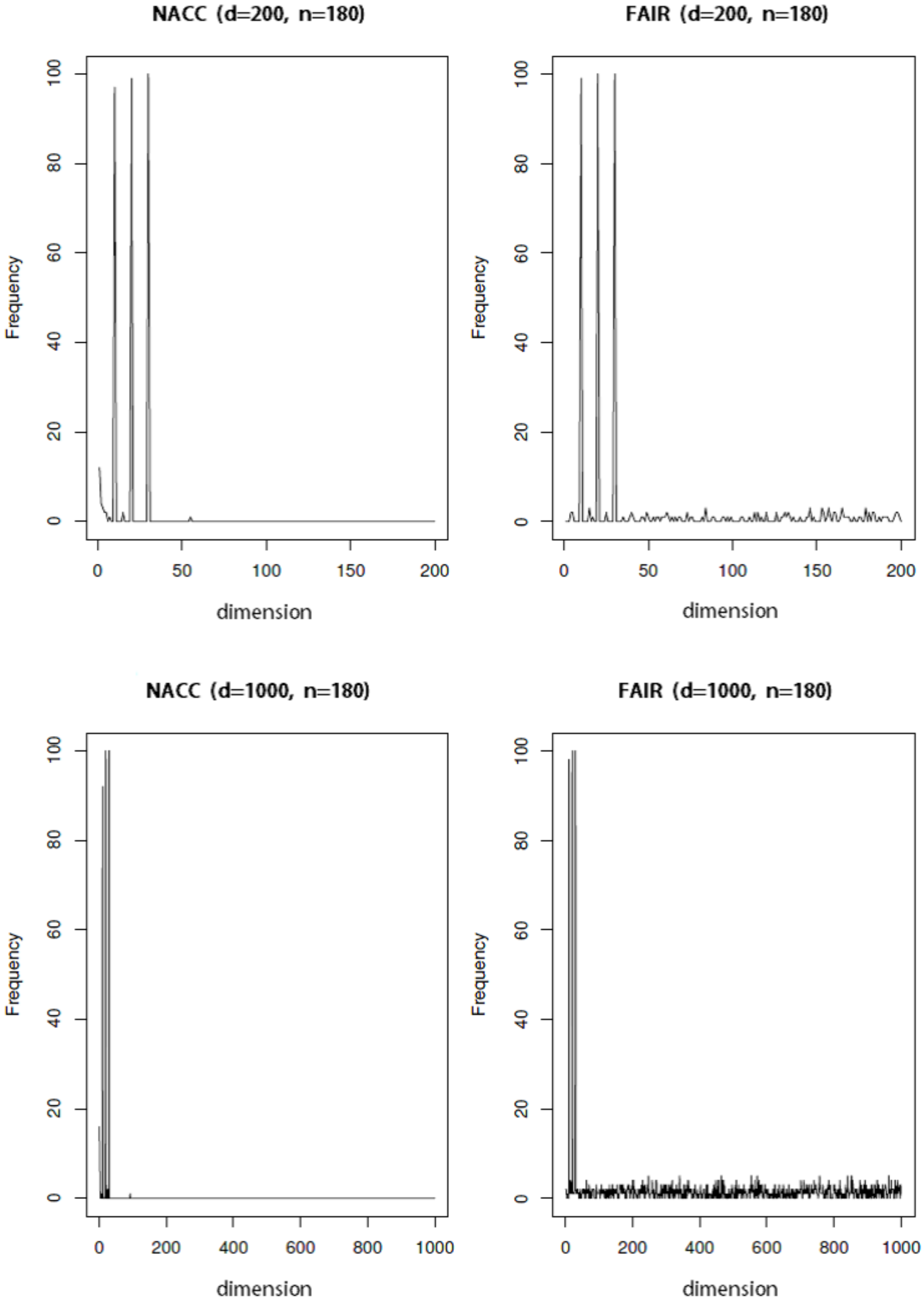


Figure 6.9: Frequency of obtained features.

6.4.2 Simulation IV-II

Simulation IV-II focuses on the performance of the proposed discriminant function with feature selection. 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:

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

where $A = \operatorname{diag}(a_{jj})$ with $a_{jj} = j$ and Σ has an AR covariance 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 (5.6). A similar comparison is given for M-FAIR and MRMR. For $d = 1000$ and $n_\ell = 50$ (with $\ell = 1, \dots, 10$), we obtained estimates for the misclassification rate based on 100 iterations.

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 $\{1, \dots, 10\}$ should be picked with a suitable feature selection. We now want to determine how successful the features selection (5.6) is in choosing the correct features. Let $\{j_1^{(t)}, \dots, j_{\hat{m}(t)}^{(t)}\}$ 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 $i \in \{1, \dots, 10\}$, let $\text{CSR}(i)$ be the *correct selection rate* for feature i and put

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

The CSR counts how often the set of selected features contains the features which should be selected. A $\text{CSR}(i)$ close to 1 means that feature i is frequently selected by the feature selection method which is good. We calculated $\text{CSR}(i)$ 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.

Table 6.6: 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

The results of the simulations are given in Table 6.6, which shows the correct selection rate as well as the estimated misclassification rate for each method. The estimate for the error based on (5.5), 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 6.6 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.

7 Application to Real Data

7.1 Lung Cancer data

In this section, we consider the lung cancer data that were analysed in Gordon et al. (2002). These data are available at `http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html`. The data have two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). There are 12553 genes, the variables, and 181 samples (31 MPM and 150 ADCA). Gordon et al. (2002) considered a training set of 16 MPM and 16 ADCA samples, and used the remaining 149 samples for testing. We use the same training and testing subsets.

For the lung cancer data, we compare FAIR with NACC. Table 7.1 shows the classification results of the two approaches. FAIR selected 14 genes, and resulted in 0 training errors and 8 test errors, while NACC selected only 7 genes which yielded 0 training errors and 8 test errors. Both approaches select features 2039 and 11368. These features may be of interest to medical experts, but we are not concerned with this aspect in the present analysis. The results show that both classification approaches have the same number of errors, so perform equally well, but NACC finds a more parsimonious set of features. Indeed, NACC requires only half the number of features in order to achieve the misclassification that FAIR achieved.

7.2 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).

Table 7.1: Results for the Lung Cancer data. Numbers in bold show common selected genes in each approach.

	NACC			FAIR		
selected genes	1955	**2039**	2928	34	1136	**2039**
	4345	8928	11238	3250	3844	4336
	11368			7249	7765	8537
				9474	11015	**11368**
				11841	12248	
No. of selected genes	7			14		
training error	0/32			0/32		
test error	8/149			8/149		

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 7.2. We observe from Table 7.2 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.

7.3 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

Table 7.2: Results for the SRBCT data. Numbers in bold show common selected genes in each approach.

	M-FAIR	**1955**	2050	1954	**1194**	1158	174	**1003**
		1389	246	107	1645	951	1980	
	M-NACC	**1955**	481	1158	1954	**1194**	1888	951
selected genes		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

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 Subsection 7.2 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 7.1. The left panel of Figure 7.1 shows the number of selected features. Figure 7.1 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 7.3. 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.

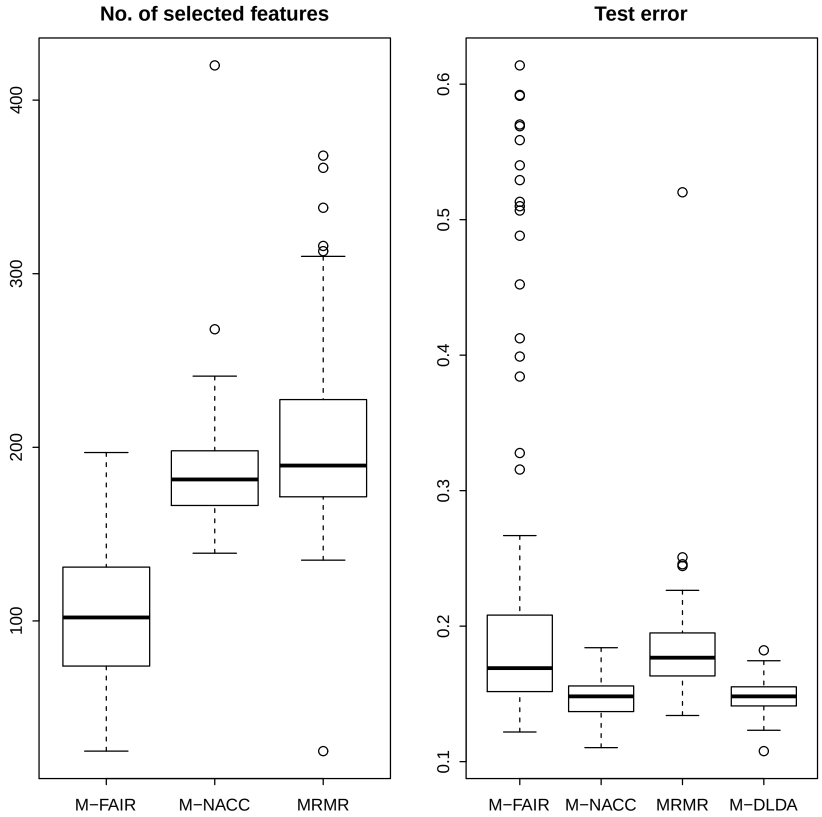


Figure 7.1: Isolet data. The left panel shows boxplots of the number of selected features for M-NACC, M-FAIR and MRMR over 100 iterations. M-FAIR is smallest, and the number of its outliers (open circles) is zero. The number of outliers for M-NACC is two, and that of MRMR is six. The right panel shows 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 7.3: 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

8 Proofs

Proof of Lemma 4.1

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

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

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 $\widetilde{M} = [\ \boldsymbol{\mu}_1 \ \cdots \ \boldsymbol{\mu}_K\ ]$ and $E_0 = [\ \overline{\boldsymbol{\varepsilon}}_1 - \overline{\boldsymbol{\varepsilon}} \ \cdots \ \overline{\boldsymbol{\varepsilon}}_K - \overline{\boldsymbol{\varepsilon}}\ ]$.

Therefore, $\widehat{C}^T\widehat{C}$ can be written as

$$
\begin{aligned}
&\widehat{C}^T\widehat{C}\\
&= \left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right)\\
&\quad+\left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}\widetilde{M}\,(N-\Pi)\,\mathbf{1}_K\mathbf{1}_K^TN^{1/2}\\
&\quad+\left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}E_0N^{1/2}+\left(N^{1/2}-\Pi^{1/2}\right) M_0^T\widehat{D}^{-1}M\Pi^{1/2}\\
&\quad+N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right)\\
&\quad+N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^TN^{1/2}\\
&\quad+N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}E_0N^{1/2}+N^{1/2}\mathbf{1}_K\mathbf{1}_K^T(N-\Pi)\widetilde{M}^T\widehat{D}^{-1}M_0\Pi^{1/2}\\
&\quad+N^{1/2}E_0^T\widehat{D}^{-1}M_0\left(N^{1/2}\Pi^{1/2}\right)+N^{1/2}E_0^T\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^TN^{1/2}\\
&\quad+N^{1/2}E_0^T\widehat{D}^{-1}E_0N^{1/2}+N^{1/2}E_0^T\widehat{D}^{-1}M_0\Pi^{1/2}\\
&\quad+\Pi^{1/2}M_0^T\widehat{D}^{-1}M_0\left(N^{1/2}-\Pi^{1/2}\right)+\Pi^{1/2}M_0\widehat{D}^{-1}\widetilde{M}(N-\Pi)\mathbf{1}_K\mathbf{1}_K^TN^{1/2}\\
&\quad+\Pi^{1/2}M_0^T\widehat{D}^{-1}E_0N^{1/2}+\Pi^{1/2}M_0\widehat{D}^{-1}M_0\Pi^{1/2}. \qquad (8.1)
\end{aligned}
$$

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

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

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

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

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

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

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

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

$$\begin{aligned} &V\left[(\overline{\varepsilon}_\ell - \overline{\varepsilon})^T D^{-1} (\overline{\varepsilon}_\ell - \overline{\varepsilon})\right] \\ &= \operatorname{tr}\left\{ (D^{-1} \otimes D^{-1}) E\left[(\overline{\varepsilon}_\ell - \overline{\varepsilon})(\overline{\varepsilon}_\ell - \overline{\varepsilon})^T \otimes (\overline{\varepsilon}_\ell - \overline{\varepsilon})(\overline{\varepsilon}_\ell - \overline{\varepsilon})^T \right] \right\} - \left\{ \operatorname{tr}(D^{-1}\Sigma^*) \right\}^2 \end{aligned} \tag{8.2}$$

by using Theorem 9.18 of Schott (1996), where $\otimes$ is Kronecker product, that is, for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$,

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp \times nq},$$

and $\Sigma^* = V[\overline{\varepsilon}_\ell - \overline{\varepsilon}] = (1/n_\ell - 1/n)\,\Sigma$. Thus, we have $\left\{\operatorname{tr}(D^{-1}\Sigma^*)\right\}^2 = d^2(1/n_\ell - 1/n)^2$. Since $D^{-1} \otimes D^{-1}$ is a diagonal matrix, (8.2) can be written as

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

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

$$D^{-1} \otimes D^{-1} E\left[\operatorname{diag}\left\{ (\overline{\varepsilon}_\ell - \overline{\varepsilon})(\overline{\varepsilon}_\ell - \overline{\varepsilon})^T \otimes (\overline{\varepsilon}_\ell - \overline{\varepsilon})(\overline{\varepsilon}_\ell - \overline{\varepsilon})^T \right\} \right] = \operatorname{diag}(v_1, \ldots, v_{d^2}), \tag{8.3}$$

where

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

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

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

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

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

In particular, we find that

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

where $\xi_{st} = E\left[\varepsilon_{11s}^2\varepsilon_{11t}^2\right]$. Therefore, the $((s-1)d+s)$th diagonal element of (8.3) becomes

$$\begin{aligned}E\left[(\overline{\varepsilon}_{\ell s} - \overline{\varepsilon}_s)^4\right] &= \frac{(n - n_\ell)(3n_\ell^2 - 3nn_\ell + n^2)}{n^3 n_\ell^3}\xi_{ss} + \frac{3}{n^4}\sum_{k\neq\ell}n_k(n_k - 1)\sigma_{ss}^2 \\ &\quad + 3\left(\frac{1}{n_\ell} - \frac{1}{n}\right)^2\frac{(n - n_\ell)(n_\ell(n_\ell + 1) + n(n_\ell - 1))}{n_\ell n^2}\sigma_{ss}^2.\end{aligned}$$

From tedious but direct calculations we have

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

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

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

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

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

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

$$\begin{aligned}
&E\left[\left(\overline{\varepsilon}_{\ell s}-\overline{\varepsilon}_s\right)^2\left(\overline{\varepsilon}_{\ell t}-\overline{\varepsilon}_t\right)^2\right]\\
&= \left(\frac{1}{n_\ell}-\frac{1}{n}\right)^4 E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)^2\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)^2\right]\\
&\quad+\left(\frac{1}{n_\ell}-\frac{1}{n}\right)^2\frac{1}{n^2}E\left[\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)^2\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)^2\right]\\
&\quad+4\left(\frac{1}{n_\ell}-\frac{1}{n}\right)^2\frac{1}{n^2}E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell is}\right)\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)\right]\\
&\quad+\left(\frac{1}{n_\ell}-\frac{1}{n}\right)^2\frac{1}{n^2}E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\left(\sum_{i=1}^{n_\ell}\varepsilon_{\ell it}\right)^2\right]\\
&\quad+\frac{1}{n^4}E\left[\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kis}\right)^2\left(\sum_{k\neq\ell}\sum_{i=1}^{n_k}\varepsilon_{kit}\right)^2\right]\\
&= \frac{(n-n_\ell)(3n_\ell^2-3n_\ell n+n^2)}{n^3n_\ell^3}\xi_{st} - \frac{(n-n_\ell)(n_\ell^2 n+3n_\ell^2-n_\ell n^2-3n_\ell n+n^2)}{n_\ell^3 n}\sigma_{ss}\sigma_{tt} + \tau\sigma_{st}^2,
\end{aligned}$$

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

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

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 (4.3), where $\odot$ is the Hadamard product; that is, if A and B are $m \times n$ matrices, then

$$A \odot B = \begin{bmatrix} a_{11} b_{11} & \cdots & a_{1n} b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} b_{m1} & \cdots & a_{mn} b_{mn} \end{bmatrix}.$$

Therefore, (8.2) can be evaluated as

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

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

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

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

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

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

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

Next, we consider the diagonal matrix

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

where

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

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

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

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

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

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 4.2

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

$$
\max \left\{ \frac{\lambda_{\alpha+1}^*}{C_d} + \xi,\ \frac{\lambda_\alpha^*}{C_d} \right\} \le \frac{\lambda_\alpha}{C_d} \le \frac{\lambda_\alpha^*}{C_d} + \xi, \tag{8.4}
$$

for $\alpha = 1, \dots, K-1$ and $0 \le \lambda_K / C_d \le \lambda_K^* / C_d + \xi = \xi$. In particular, it follows from (8.4) 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 4.1

Put $\Gamma_K = [\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K]$, 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 4.1, we obtain

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

Let $\widehat{H} = [\ \widehat{\boldsymbol{h}}_1\ \cdots\ \widehat{\boldsymbol{h}}_K\ ]$, 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 4.2, 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

$$\begin{aligned} &\Gamma_K^T\frac{\widehat{C}^T\widehat{C}}{C_d}\Gamma_K\widehat{\boldsymbol{h}}_\ell = \frac{\widehat{\lambda}_\ell}{C_d}\widehat{\boldsymbol{h}}_\ell \\ \Longrightarrow \quad & \frac{\widehat{C}^T\widehat{C}}{C_d}\left(\Gamma_K\widehat{\boldsymbol{h}}_\ell\right) = \frac{\widehat{\lambda}_\ell}{C_d}\left(\Gamma_K\widehat{\boldsymbol{h}}_\ell\right) \\ \Longrightarrow \quad & \frac{\widehat{C}\widehat{C}^T}{C_d}\left\{\frac{\widehat{C}\boldsymbol{\gamma}_\ell}{||\widehat{C}\boldsymbol{\gamma}_\ell||}(1+o_P(1))\right\} = \frac{\widehat{\lambda}_\ell}{C_d}\left\{\frac{\widehat{C}\boldsymbol{\gamma}_\ell}{||\widehat{C}\boldsymbol{\gamma}_\ell||}(1+o_P(1))\right\}. \end{aligned} \tag{8.5}$$

On the other hand,

$$\frac{\widehat{C}\widehat{C}^T}{C_d}\widehat{\boldsymbol{p}}_\ell = \frac{\widehat{\lambda}_\ell}{C_d}\widehat{\boldsymbol{p}}_\ell \tag{8.6}$$

follows from the definition in Subsection 4.1. Now, from (8.5), (8.6) and LEMMA 4.2, 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 $\operatorname{sgn}(\hat{p}_{\alpha 1}) = \operatorname{sgn}\left(\left(\widehat{C}\boldsymbol{\gamma}_\alpha/||\widehat{C}\boldsymbol{\gamma}_\alpha||\right)_1\right)$, where $(\cdot)_1$ denotes the first component of the vector. Therefore, we have

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

□

Proof of THEOREM 4.2

From THEOREM 4.1 and (4.7), the inner product of $\widehat{\boldsymbol{p}}_\alpha$ and $\boldsymbol{p}_\beta$ is given by

$$\begin{aligned} \widehat{\boldsymbol{p}}_\alpha^T\boldsymbol{p}_\beta &= \frac{\boldsymbol{\gamma}_\alpha^T\widehat{C}^TC\boldsymbol{\gamma}_\beta(1+o_P(1))}{\sqrt{\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{C}\boldsymbol{\gamma}_\alpha}\sqrt{\boldsymbol{\gamma}_\beta^TC^TC\boldsymbol{\gamma}_\beta}} \\ &= \frac{\boldsymbol{\gamma}_\alpha^T\Pi^{1/2}\widehat{M}_0^TD^{-1}M_0\Pi^{1/2}\boldsymbol{\gamma}_\beta(1+o_P(1))}{\sqrt{\boldsymbol{\gamma}_\alpha^T\widehat{C}^T\widehat{C}\boldsymbol{\gamma}_\alpha}\sqrt{\boldsymbol{\gamma}_\beta^TC^TC\boldsymbol{\gamma}_\beta}}. \end{aligned} \tag{8.7}$$

The numerator of (8.7) can be evaluated as

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

by Chebyshev's inequality. By LEMMA 4.1, $\boldsymbol{\gamma}_\alpha^T \widehat{C}^T \widehat{C} \boldsymbol{\gamma}_\alpha$ of (8.7) 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 (8.7) 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

$$
\widehat{\boldsymbol{p}}_\alpha^T \boldsymbol{p}_\beta = \frac{\kappa_\beta \delta_{\alpha\beta} - \xi(\delta_{\alpha\beta} - \eta_\alpha \eta_\beta)}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta - \xi(1 - \eta_\beta^2)}}(1 + o_P(1)). \tag{8.8}
$$

□

Proof of COROLLARY 4.1

From (8.8) in the proof of THEOREM 4.2, and the assumptions of COROLLARY 4.1, it follows that

$$
\begin{aligned}
\widehat{\boldsymbol{p}}_\alpha^T \boldsymbol{p}_\beta &= \frac{\kappa_\beta \delta_{\alpha\beta}}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta}}(1 + o_P(1)) \\
&= \begin{cases} 1 + o_P(1) & \text{if } \alpha = \beta, \\ o_P(1) & \text{if } \alpha \neq \beta, \end{cases}
\end{aligned}
$$

since $\xi \to 0$. □

Proof of THEOREM 4.3

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

$$
\begin{aligned}
\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* &= \frac{\widehat{\boldsymbol{p}}_\alpha^T \widehat{D}^{-1/2} D^{-1/2} \boldsymbol{p}_\beta}{\sqrt{\widehat{\boldsymbol{p}}_\alpha^T \widehat{D}^{-1} \widehat{\boldsymbol{p}}_\alpha}\sqrt{\boldsymbol{p}_\beta^T D^{-1} \boldsymbol{p}_\beta}} \\
&= \frac{\boldsymbol{\gamma}_\alpha^T \widehat{C}^T \widehat{D}^{-1/2} D^{-1/2} C \boldsymbol{\gamma}_\beta}{\sqrt{\boldsymbol{\gamma}_\alpha^T \widehat{C}^T \widehat{D}^{-1} \widehat{C} \boldsymbol{\gamma}_\alpha}\sqrt{\boldsymbol{\gamma}_\beta^T C^T D^{-1} C \boldsymbol{\gamma}_\beta}}(1 + o_P(1)).
\end{aligned} \tag{8.9}
$$

using THEOREM 4.1, (4.7) and (4.11). The numerator of (8.9) 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 (8.1), $\widehat{C}^T \widehat{D}^{-1} \widehat{C}$ of (8.9) 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

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

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

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \quad \ge \quad \boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \frac{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2) (1 - \sigma_{\max}/\sigma_{\min})}} (1 + o_P(1)). \tag{8.10}$$

Similarly, we obtain

$$\widehat{\boldsymbol{b}}_\alpha^{*T} \boldsymbol{b}_\beta^* \quad \le \quad \boldsymbol{b}_\alpha^{*T} \boldsymbol{b}_\beta^* \frac{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi (1 - \eta_\alpha^2) (1 - \sigma_{\min}/\sigma_{\max})}} (1 + o_P(1)). \tag{8.11}$$

□

Proof of COROLLARY 4.2

From the assumption $d = o(nC_d)$ that (8.10) and (8.11) can be evaluated as

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

and

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

respectively. Therefore, we have $\widehat{\boldsymbol{b}}_\alpha^{*T}\boldsymbol{b}_\beta^* = \boldsymbol{b}_\alpha^{*T}\boldsymbol{b}_\beta^*(1+o_P(1))$. □

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

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

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

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

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

$$\begin{aligned}
\hat{d}_{k\alpha} &= \frac{E\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]}{\sqrt{V\left[\widehat{\delta}_{k\alpha}(\boldsymbol{X})\right]}} \\
&= \frac{(\boldsymbol{\mu}_k - (\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)/2)^T\, \widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}{\sqrt{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}\Sigma\widehat{B}^T(\widehat{B}\widehat{D}\widehat{B}^T)^{-1}\widehat{B}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}}. \qquad (8.12)
\end{aligned}$$

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

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

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

$$\widehat{\Sigma}_k = \widehat{W}_k^T\Sigma\widehat{W}_k, \qquad (8.13)$$

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

$$\widehat{W}_k = \left[\begin{array}{ccc} \dfrac{\underline{\widehat{\boldsymbol{w}}}_{k1}}{\sqrt{\underline{\widehat{\boldsymbol{w}}}_{k1}^T\Sigma\underline{\widehat{\boldsymbol{w}}}_{k1}}} & \cdots & \dfrac{\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}}{\sqrt{\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}^T\Sigma\underline{\widehat{\boldsymbol{w}}}_{k(K-1)}}} \end{array}\right]$$

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

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

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

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

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

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

□

Proof of Theorem 4.4

By Theorem 4.1, $\widehat{B}$ 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 $\widehat{L} = \text{diag}\left(||\widehat{C}\boldsymbol{\gamma}_1||, \ldots, ||\widehat{C}\boldsymbol{\gamma}_{K-1}||\right)$. Using $\widehat{D} = D(1+o_P(1))$, (8.12) 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

$$\begin{aligned} I_3 &= (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M}_0 \\ &= \left[(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}), \ldots, (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}(\widehat{\boldsymbol{\mu}}_K - \widehat{\boldsymbol{\mu}})\right]. \end{aligned} \tag{8.14}$$

From Condition A, a typical component of (8.14) can be expressed as

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

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

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

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

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

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

$$\begin{aligned} &(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1} M_0 \\ &= \left[\sum_{\ell=1}^{K} \pi_\ell(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_\ell), \ldots, \sum_{\ell=1}^{K} \pi_\ell(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_\ell)\right] \\ &= \mathbf{1}_K^T \Pi F, \end{aligned}$$

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

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

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

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

where $\boldsymbol{\beta}_{k\alpha} = (0, \dots, 0, d/n_k, 0, \dots, 0, -d/n_\alpha, 0, \dots, 0)$. Next, we consider I_1. We find that

$$I_1 = \left(\boldsymbol{\mu}_k - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)\right)^T D^{-1}\widehat{M_0} = -\overline{\boldsymbol{\varepsilon}}_k^T D^{-1}\widehat{M_0} + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T D^{-1}\widehat{M_0}. \qquad (8.15)$$

Similarly, (8.15) 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

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

Finally, we consider I_2. It can be written as

$$I_2 = \widehat{M_0} D^{-1} \widehat{M_0} = \left((\widehat{\boldsymbol{\mu}}_\alpha - \widehat{\boldsymbol{\mu}})^T D^{-1}(\widehat{\boldsymbol{\mu}}_\beta - \widehat{\boldsymbol{\mu}})\right)_{1 \le \alpha, \beta \le K}. \tag{8.16}$$

Each component of (8.16) can be decomposed as

$$(\widehat{\boldsymbol{\mu}}_\alpha - \widehat{\boldsymbol{\mu}})^T D^{-1}(\widehat{\boldsymbol{\mu}}_\beta - \widehat{\boldsymbol{\mu}}) = \left\{(\boldsymbol{\mu}_\alpha - \boldsymbol{\mu})^T D^{-1}(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}) + J_1 + J_2\right\}(1 + o_P(1)) + J_3, \tag{8.17}$$

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

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

Consequently, (8.17) 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

$$\begin{aligned}
&\widehat{d}_{k\alpha} \\
&\geq \frac{S_{k\alpha}\Gamma\left[\Gamma^T\left\{C^T C + (d/n)\left(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2}\right)\right\}\Gamma\right]^{-1}\Gamma^T Q_{k\alpha}^T(1+o_P(1))}{\sqrt{\lambda_{\max}(R)}\sqrt{Q_{k\alpha}\Gamma\left[\Gamma^T\left\{C^T C + (d/n)\left(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2}\right)\right\}\Gamma\right]^{-1}\Gamma^T Q_{k\alpha}^T}}. \qquad (8.18)
\end{aligned}$$

This completes the proof of Theorem 4.4. □

Proof of Corollary 4.3

Using $M_{k\alpha} = \mathbf{1}_K^T O(C_d)$ and $C^T C = \mathbf{1}_K\mathbf{1}_K^T O(C_d)$, $S_{k\alpha}$ of (8.18) becomes

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

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

Proof of Theorem 4.6

The right side of (4.17) can be written as

$$\begin{aligned}
&1-\Phi\left(\frac{\sqrt{n_1n_2/(dn)}\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}(1+o_P(1))+(n_1-n_2)\sqrt{d/(nn_1n_2)}}{2\sqrt{\lambda_{\max}(R)}\sqrt{1+n_1n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}(1+o_P(1))/(dn)}}\right)\\
&=1-\Phi\left(\frac{\sqrt{n_1n_2/(dn)}\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}\left[(1+o_P(1))+\{d/(n\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\}\{(n_1-n_2)/n_1\}\right]}{2\sqrt{\lambda_{\max}(R)}\sqrt{n_1n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}/(dn)}\sqrt{(n/n_1)\{d/(n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\}+(1+o_P(1))}}\right)\\
&=1-\Phi\left(\frac{\sqrt{n_1n_2/(dn)}\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}\left[(1+o_P(1))+o(1)O(1)\right]}{2\sqrt{\lambda_{\max}(R)}\sqrt{n_1n_2/(dn)}\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}\sqrt{O(1)o(1)+(1+o_P(1))}}\right)\\
&=1-\Phi\left(\frac{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}{2\sqrt{\lambda_{\max}(R)}}(1+o_P(1))\right)
\end{aligned}$$

by the assumption $d=o(nC_d)$. Therefore, we have

$$W_1(\widehat{g})=\max_{\theta\in\Theta}W_1(\widehat{g},\theta)=1-\Phi\left(\frac{C_d}{2\sqrt{b_0}}(1+o_P(1))\right).$$

□

Proof of Corollary 4.4

We derive (4.18) as follows:

$$\begin{aligned}
&1-\Phi\left(\frac{\sqrt{n_1n_2/(dn)}\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}(1+o_P(1))+(n_1-n_2)\sqrt{d/(nn_1n_2)}}{2\sqrt{\lambda_{\max}(R)}\sqrt{1+n_1n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}(1+o_P(1))/(dn)}}\right)\\
&=1-\Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}[(1+o_P(1))+\{d/(n_2C_d)\}(C_d/\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})(1-n_2/n_1)]}{2\sqrt{\lambda_{\max}(R)}\sqrt{(n/n_1)\{d/(n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\}+(1+o_P(1))}}\right)\\
&=1-\Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}[(1+o_P(1))+\{d/(n_2C_d)\}(C_d/\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\{1-(c_0+o(1))\}]}{2\sqrt{\lambda_{\max}(R)}\sqrt{(n/n_1)\{d/(n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\}+(1+o_P(1))}}\right)\\
&\geq 1-\Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}[(1+o_P(1))+\{d/(n_2C_d)\}(C_d/\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})o(1)]}{2\sqrt{\lambda_{\max}(R)}\sqrt{(n/n_1)\{d/(n_2\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha})\}+(1+o_P(1))}}\right)\\
&=1-\Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}\{(1+o_P(1))+O(1)O(1)o(1)\}}{2\sqrt{\lambda_{\max}(R)}\sqrt{O(1)O(1)O(1)+(1+o_P(1))}}\right)\\
&>1-\Phi\left(\frac{\sqrt{\boldsymbol{\alpha}^TD^{-1}\boldsymbol{\alpha}}}{2\sqrt{\lambda_{\max}(R)}}(1+o_P(1))\right).
\end{aligned}$$

□

9 Conclusion

In this paper, we discussed the asymptotic theories of the multi-class linear discriminant function in a HDLSS context. In Section 3, we constructed the linear discriminant function based on naive canonical correlation in the context of multi-class problem. In Section 4, we derived the asymptotic behavior of eigenvectors of the naive canonical correlation matrix corresponding to positive eigenvalues. In the asymptotic theory, both the dimension d and the sample size n grow, and provided d does not grow too fast, we showed that all eigenvectors and discriminant directions are HDLSS consistent. Under suitable conditions, we were able to derive an upper bound for the worst case misclassification rate in the multi-class setting. In Section 5, we proposed a feature selection method for HDLSS data, called NACC approach, using a discriminant direction. Further, for the general multi-class setting, we proposed and discussed two methods for feature selection, called 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 Subsection 6.4 that NACC and M-NACC performed better than FAIR and M-FAIR respectively. Applications to real data sets demonstrate that NACC and M-NACC performed well.

In recent years, other discriminant method for HDLSS data has been studied by many authors: Marron et al. (2007) proposed Distance-Weighted Discrimination (DWD) which improves the performance of Support Vector Machine (SVM) in the HDLSS setting, Fan et al. (2012) proposed Regularized Optimal Affine Discriminant (ROAD) which uses the L_1-constraint in the Fisher's criterion (2.3), Ahn et al. (2012) proposed a hierarchical clustering algorithm based on the MDP distance by referring to Ahn and Marron (2010). Other possible research directions include extensions of our theoretical results to the "kernel method" in linear discrimination described in Mika et al. (1999).

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 k_0-banded matrix has been considered in Bickel and Levina (2008), and their results are expected to apply to linear discriminant function in HDLSS settings. Furthermore, we can explore issues of discriminant function based on invertible k_0-banded matrix: asymptotic behavior of misclassification rate, the selection criteria of k_0, the algorithm for preprocessing correlation of HDLSS data before discrimination.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Kanta Naito for all of his help and lots of encouragement to me during the bachelor's, the master's and doctoral courses. I would also like to thank Prof. Inge Koch for her many valuable comments and suggestions. Prof. Toshihiro Nakanishi and Prof. Daishi Kuroiwa gave me helpful comments about the draft of this thesis.

This work was supported by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows. I am grateful to the JSPS for making this study possible by the financial.

Finally, I would like to thank my father and mother for their understanding, support and encouragement during this study.

References

Ahn, J., Lee, M. H. and Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance, *Statistica Sinica*, Vol. 22, p. 443.

Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination, *Biometrika*, Vol. 97, pp. 254–259.

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*, Vol. 94, pp. 760–766.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data, *Sequential analysis*, Vol. 30, pp. 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*, Vol. 10, pp. 989–1010.

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

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, Vol. 1: springer New York.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*: Springer.

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

Duda, R. O., Hart, P. E. and Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

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*, Vol. 97, pp. 77–87.

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

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 74, pp. 745–771.

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

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *science*, Vol. 286, pp. 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, *Cancer Research*, Vol. 62, pp. 4963–4967.

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*, Vol. 29, pp. 295–327.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context, *The Annals of Statistics*, Vol. 37, pp. 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, *Nat. Med.*, Vol. 7, pp. 673–679.

Koch, I. and Naito, K. (2010). Prediction of multivariate responses with a selected number of principal components, *Computational Statistics & Data Analysis*, Vol. 54, pp. 1791–1807.

Mardia, K. V., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*: Academic Press.

Marron, J. S., Todd, M. J. and Ahn, J. (2007). Distance-weighted discrimination, *Journal of the American Statistical Association*, Vol. 102, pp. 1267–1271.

Mika, S., Rätsch, G., Weston, J. and Schölkopf, B. (1999). Fisher discriminant analysis with kernels, *IEEE Neural Networks for Signal Processing Workshop*, pp. 41–48.

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 27, pp. 1226–1238.

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

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

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

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer cell*, Vol. 1, pp. 203–209.

Srivastava, M. S. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data, *J. Japan Statist. Soc*, Vol. 37, pp. 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*, Vol. 111, pp. 350–367.

Tamatani, M., Naito, K. and Koch, I. (2013). Multi-class discriminant function based on canonical correlation in high dimension low sample size, *Bulletin of Informatics and Cybernetics*, Vol. 45, pp. 67–101.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences*, Vol. 99, pp. 6567–6572.

Weinberger, K., Blitzer, J. and Saul, L. (2006). Distance Metric Learning for Large Margin Nearest Neighbor Classification, in Y. Weiss, B. Schölkopf and J. Platt eds. *Advances in Neural Information Processing Systems 18*, Cambridge, MA: MIT Press, pp. 1473–1480.