学位申請論文

Estimation of misclassification probability for Euclidean distance in high-dimensional data

（高次元データにおけるユークリッド距離に基づく誤判別確率の推定）

２０１９年９月

渡邉　弘己

Acknowledgments

The author would like to express my gratitude to Professor Takashi Seo for his much support and advice. The author would like to appreciate Associate Professor Masashi Hyodo for his valuable idea and discussions. The author would like to thank Mr.Yuki Yamada for his help. The author would like to thank Lecturer Nobumichi Shutoh for his help. The author would like to thank all members of Health Informatics and Biostatistics Laboratory in Oita University of Nursing and Health Sciences. Finally, the author would like to thank his family.

Contents

Chapter 1

Introduction

In classical multivariate analysis, one of the representative classification methods is linear discriminant analysis (LDA). LDA is a generalization of Fisher's linear discriminant rule. The original dichotomous discriminant rule was developed by Fisher [13]. LDA is widely applied to practical situations. In order to quantitatively evaluate the degree of its classification accuracy, we need to investigate misclassification probability. Even in the problem of estimating the misclassification probability of LDA, it is too complicated to directly estimate, so the research that approximate with a simple expression has been conducted. For a review of these results, see, e.g., Okamoto [30, 31], Lachenbruch [26], and Siotani [37]. Because it is difficult to introduce all these studies, we introduce two representative approximations in Chapter 2.

By the way, the most of theoretical properties of LDA is derived under the following assumptions.

- Large sample framework:Sample size is much larger than dimension.
- Multivariate normality:Both training data and test data are random samples from multivariate normal population.
- Homogeneity of covariance matrices(homoscedasticity):The population covariance matrices of each group are all the same.

But, they do not necessarily conform to the data observed in recent years. For example, the data set has become increasingly larger in dimension p than total sample size in many applications such as genome projects (Xing et al. [44]), text categorization (Yang and Pederson [46]), image retrieval (Rui et al. [34]), and customer relationship management (Ng and Liu [28]). However, the recent increase of dimensionality of data poses a severe problem to many existing classical multivariate statistical methods with respect to efficiency and effectiveness, because many classical multivariate statistical theories are based on large sample framework. It becomes essential to revise the classical multivariate statistical theory in order to make them useful in wide range of relations between p and n and to extend multivariate statistical theory in high-dimensional situations. For high-dimensional data classification problem, due to the small number of samples and large number of variables, classical LDA has poor performance corresponding to the singularity and instability of the sample covariance matrix. Several efficient methods for this problem

have been proposed. When $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, Saranadasa [35] considered the approach based on one way MANOVA problem. Bickel and Levina [7] considered using diagonal linear discrminant rule. Srivastava [39] considered using the Moore-Penrose inverse matrix. Hyodo et al. [23] considered the modified W-rule. When $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, Dudoit et al. [12] considered using the inverse matrix defined by only diagonal elements of each group sample covariance matrix. Aoshima and Yata [2] proposed a quadoratic classifier and showed that the classifier has misclassification rates which are no more than a prespecified value. Hall et al. [18] and Marron et al. [27] considered distance weighted classifiers. Hall et al. [18, 19] and Chan and Hall [9] considered distance based classifiers. Recentry, Aoshima and Yata [3] discussed a scale adjusted-type distance-based classifier given by Chan and Hall [9]. In this discrimination rule, it is essential to unbiased-estimate the Euclidean norm of the population mean difference, so we call it the Euclidean distance rule.

In this paper, we focus on the estimation of Euclidean distance rule's misclassification probability. Aoshima and Yata [3] proved the consistency and asymptotic normality of the discriminant function which is used for this discrimination rule. Using their asymptotic normality and Chen and Qin's estimator (See, Chen and Qin [10]), we can propose an estimator of the misclassification probability. This estimator has consistency under certain assumptions, however the estimator has a large bias if the assumptions dose not hold. In Chapter 3, we introduce the existing theoretical properties of Euclidean distance discrimination function and we will be described in details of this point. Since practical use in a wider range is the object of this paper, we aim to construct new estimator of the misclassification probability under a relaxed mathematical assumption as much as possible. We deal with the estimation of the misclassification probability in the two-sample problem in Chapter 4 and the estimation of the misclassification probability in the k-sample problem in Chapter 5, respectively. In Chapter 4, we propose a Lachenbruch type approximation of the misclassification probability and constructe a plug-in type estimator and showed its consistency. In Chapter 5, by extending the asymptotic normality of the discriminant function in two samples to multiple samples, we provide an approximation of misclassification probability based on multinormal distribution.

The discrimination accuracy of Euclidean distance rule which is dealt with in Chapters 4 and 5 depends on the magnitude of the square Euclidean norm of the population mean difference. In Chapter 6, we propose an interval estimation method of the square Euclidean norm of the population mean difference.

Chapter 2

Classical linear discriminant analysis

2.1 Construction of discriminant rules

We treat the problem of classifying a $p \times 1$ observation vector $\mathbf{x}$ as coming from one of two populations G_1 and G_2. We consider the pure decision case. It is make an assignment of an entity with feature vector $\mathbf{x}$ to one of the groups. Let $r(\mathbf{x})$ denote an allocation rule formed for this purpose, where $r(\mathbf{x}) = i$ implies that an entity with feature vector $\mathbf{x}$ is to be assigned to the i-th group $G_i,\ i \in \{1, 2\}$. Under the mixture model approach to discriminant analysis, it is assumed that the entity has been drawn from the two groups G_1 and G_2 in proportions π_1 and π_2, respectively, where

$$\sum_{i=1}^{2} \pi_i = 1,\ 0 < \pi_1,\ \pi_2 < 1.$$

Let c_{ij} denote the cost of allocation when an entity from G_j is allocated to group Π_i, where $c_{ij} = 0$ for $i = j$ that is zero cost for a correct allocation. Then expected loss for the rule $r(\boldsymbol{x})$ is

$$R(r) = e(2|1)\pi_1 c_{21} + e(1|2)\pi_2 c_{12},$$

where

$$e(2|1) = \Pr\left(r(\mathbf{x}) = 2|\mathbf{x} \sim G_1\right) \quad \textit{and} \quad e(1|2) = \Pr\left(r(\mathbf{x}) = 1|\mathbf{x} \sim G_2\right).$$

Here, $e(i|j)$ for $i \neq j$ is called the misclassification probability. An optimal rule of allocation can be defined by taking it to be the one that minimizes the risk $R(r)$ at each value $\mathbf{x}$ of the feature vector (see, e.g., Anderson [1]). The rule that minimizes the risk $R(r)$ is said to be a Bayes rule. Under a multivariate normal model for the group-conditional distributions of the feature vector $\mathbf{x}$ on entity, it is assumed that $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),\ i \in \{1, 2\}$. Then the i-th group-conditional density $f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is given by

$$f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)}{2}\right\}.$$

Therefore, $R(r)$ can be expressed as

$$\begin{aligned} R(r) =& \pi_1 c_{21} \int_{R_2} f(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) d\mathbf{x} + \pi_2 c_{12} \int_{R_1} f(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) d\mathbf{x} \\ =& \int_{R_2} \{\pi_1 c_{21} f(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) - \pi_2 c_{12} f(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\} d\mathbf{x} + \pi_2 c_{12}. \end{aligned}$$

In this setting, the optimal or Bayes rule r_0 assigns an entity with feature vector $\mathbf{x}$ to G_1 if

$$R_1 : \log \frac{f(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{f(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})} \geq \log \frac{\pi_2 c_{12}}{\pi_1 c_{21}}.$$

Otherwise, the entity is assigned to Π_2 if

$$R_2 : \log \frac{f(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{f(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})} < \log \frac{\pi_2 c_{12}}{\pi_1 c_{21}}.$$

When we assume that $\pi_1 = \pi_2$ and $c_{12} = c_{21}$, the optimal or Bayes rule r_0 is given by

$$r_0(\mathbf{x}) = \begin{cases} 1, & L_0 \geq 0, \\ 2, & L_0 < 0, \end{cases}$$

where

$$L_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \{\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}.$$

The misclassification probability resulting from this rule is the same for observations from either population, and it is $\Phi(-\Delta_{\boldsymbol{\Sigma}^{-1}}/2)$, where $\Phi(\cdot)$ denotes a cumulative distribution function of the standard normal distribution and $\Delta^2_{\boldsymbol{\Sigma}^{-1}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the Mahalanobis squared distance between the two populations. So, risk for Bayes rule r_0 is given by

$$R(r_0) = \Phi\left(-\frac{1}{2}\Delta_{\boldsymbol{\Sigma}^{-1}}\right).$$

However, in practical use, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are generally taken to be unknown and must be estimated from the available training data $\mathbf{x}_{ij}$, $j \in \{1, 2, \ldots, n_i\}$ as given by G_i, $i \in \{1, 2\}$. With the estimative approach to discriminant analysis, the Bayes rule r_0 is estimated simply by plugging in estimate $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, such as the maximum likelihood estimate, for $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ in the group-conditional densities. The maximum likelihood estimates of $\boldsymbol{\mu}_i$, $i = 1, 2$ and $\boldsymbol{\Sigma}$ computed from the training data are given by the sample mean $\overline{\mathbf{x}}_i$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, respectively, where

$$\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}, \ i = 1, 2$$

and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^\top.$$

Here, $n = n_1 + n_2$. In the subsequent work, we follow the usual practice of estimating $\mathbf{\Sigma}$ by the unbiased estimator

$$\mathbf{S} = \frac{n}{n-2}\widehat{\mathbf{\Sigma}},$$

With the parameter sets $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma})$ estimated as above, the plug-in sample version $\hat{r}_0$ is given by

$$\hat{r}_0(\boldsymbol{x}) = \begin{cases} 1, & L \geq 0, \\ 2, & L < 0, \end{cases}$$

where

$$L = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}\{\mathbf{x} - \tfrac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)\}.$$

This method is called linear discriminant rule (LDR). It is essentially the same as Fisher's linear discriminant analysis (See, Fisher [13]) without the explicit adoption of a normal population. In addition, it should be noted that the statistics L is obtained by replacing unknown parameters in the Bayes rule L_0 with their consistent estimators. Thus, LDR has Bayes risk consistent rule. Namely, the conditional risk for W converges in probability to one of the Bayes rule $R(r_0)$, as p :fix and $n_1, n_2 \to \infty$. Altogether, LDR has a good property in large sample case.

2.2 Error rate of discriminant rule

We adopt the misclassification probability as a standard of discrimination performance. The misclassification probability means that the probability of misclassifying $\mathbf{x}$ into G_2 (G_1) when it actually belongs to G_1 (G_2).

Definition 2.2.1 (The misclassification probability of LDR)**.** *Let* $\mathbf{x}_{1j} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma})$, $\mathbf{x}_{2j'} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma})$ *for* $j \in \{1, 2, \ldots, n_1\}$, $j' \in \{1, 2, \ldots, n_2\}$, *and* $\mathbf{x}, \mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ *be mutually independent. The misclassification probabilities of LDR are defined by*

$$e(2|1) = \Pr(L < 0) \text{ when } \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}),$$
$$e(1|2) = \Pr(L > 0) \text{ when } \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}).$$

It is generally difficult to obtain an explicit expression for the misclassification probability. So, there are many works for asymptotic properties of misclassification probability for LDR. The asymptotic properties under a framework such that n_1 and n_2 are large and p is fixed has been studied. Okamoto [30, 31] obtained an asymptotic expansion for the distribution of L up to terms of $O(n^{-2})$. Siotani and Wang [38] extended their results to terms of $O(n^{-3})$. Since $e(1|2)$ can be obtained from same logic of $e(2|1)$, we treat only discuss $e(2|1)$. The following theorem gives the asymptotic expansion of misclassification probability.

Theorem 2.2.1 (Okamoto [30, 31])**.** *Let* $\mathbf{x}, \mathbf{x}_{1j} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma})$, $\mathbf{x}_{2j'} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma})$ *for* $j \in \{1, 2, \ldots, n_1\}$, $j' \in \{1, 2, \ldots, n_2\}$, *and* $\mathbf{x}, \mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ *be mutually*

independent. We assume that $n_1 = O(n_2)$ *and* $n_2 = O(n_1)$ *as* $n_1 \to \infty$, $n_2 \to \infty$. *Then it holds that*

$$\begin{aligned} e(2|1) =& \Phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2) + \frac{\phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2)}{4}\left[\frac{1}{4n_1\Delta_{\mathbf{\Sigma}^{-1}}}\{\Delta^2_{\mathbf{\Sigma}^{-1}} + 12(p-1)\}\right. \\ &\left. + \frac{1}{4n_2\Delta_{\mathbf{\Sigma}^{-1}}}\{\Delta^2_{\mathbf{\Sigma}^{-1}} - 4(p-1)\} + \frac{1}{n-2}(p-1)\Delta_{\mathbf{\Sigma}^{-1}}\right] \end{aligned}$$

as $n_1 \to \infty$, $n_2 \to \infty$, *where* $\phi(\cdot)$ *denotes a probability density function of the standard normal distribution.*

The asymptotic expansion in Theorem 2.2.1 contains an unknown parameter $\Delta_{\mathbf{\Sigma}^{-1}}$. As a simple method, we use the sample Mahalanobis distance

$$\widehat{\Delta}_{\mathbf{\Sigma}^{-1}} = \sqrt{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)}$$

as the estimator of $\Delta_{\mathbf{\Sigma}^{-1}}$. However, $\Phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)$ has a bias. In fact,

$$\begin{aligned} \mathrm{E}\{\Phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)\} =& \Phi(\Delta_{\mathbf{\Sigma}^{-1}}/2) + \phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2)\left[\frac{n}{n_1 n_2}\left\{\Delta_{\mathbf{\Sigma}^{-1}} - \frac{4(p-1)}{\Delta_{\mathbf{\Sigma}^{-1}}}\right\}\right. \\ &\left. + \frac{\Delta_{\mathbf{\Sigma}^{-1}}\{\Delta^2_{\mathbf{\Sigma}^{-1}} - 4(2p+1)\}}{2(n-2)}\right] + O(n^{-2}). \end{aligned}$$

By combining this result with Theorem 2.2.1, we obtain

$$\begin{aligned} e(2|1) - \mathrm{E}\{\Phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)\} =& \phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2)\left[\frac{n}{n_1 n_2}\left\{\Delta_{\mathbf{\Sigma}^{-1}} - \frac{4(p-1)}{\Delta_{\mathbf{\Sigma}^{-1}}}\right\}\right. \\ &\left. + \frac{\Delta_{\mathbf{\Sigma}^{-1}}\{\Delta^2_{\mathbf{\Sigma}^{-1}} - 4(2p+1)\}}{2(n-2)}\right] + O(n^{-2}). \end{aligned} \tag{2.1}$$

Also, we note that

$$\begin{aligned} &\mathrm{E}\left[\frac{n\phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)}{n_1 n_2}\left\{\widehat{\Delta}_{\mathbf{\Sigma}^{-1}} - \frac{4(p-1)}{\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}}\right\} + \frac{\phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}\{\widehat{\Delta}^2_{\mathbf{\Sigma}^{-1}} - 4(2p+1)\}}{2(n-2)}\right] \\ =& \frac{n\phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2)}{n_1 n_2}\left\{\Delta_{\mathbf{\Sigma}^{-1}} - \frac{4(p-1)}{\Delta_{\mathbf{\Sigma}^{-1}}}\right\} + \frac{\phi(-\Delta_{\mathbf{\Sigma}^{-1}}/2)\Delta_{\mathbf{\Sigma}^{-1}}\{\Delta^2_{\mathbf{\Sigma}^{-1}} - 4(2p+1)\}}{2(n-2)} + o(n^{-1}). \end{aligned} \tag{2.2}$$

From (2.1) and (2.2), we can obtain an estimator of $e(2|1)$:

$$\begin{aligned} \widehat{e(2|1)} =& \Phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2) + \frac{n\phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)}{n_1 n_2}\left\{\widehat{\Delta}_{\mathbf{\Sigma}^{-1}} - \frac{4(p-1)}{\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}}\right\} \\ &+ \frac{\phi(-\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}/2)\widehat{\Delta}_{\mathbf{\Sigma}^{-1}}\{\widehat{\Delta}^2_{\mathbf{\Sigma}^{-1}} - 4(2p+1)\}}{2(n-2)}. \end{aligned}$$

This estimator is second-order unbiased estimator of $e(2|1)$ i.e., $\mathrm{E}(\widehat{e(2|1)}) = e(2|1) + o(n^{-1})$.

The asymptotic properties under a framework that n_1, n_2 and p are all large (but $p < n-2$) have also been studied (see, e.g., Lachenbruch [26], Raudys [33] and Fujikoshi and Seo [15]). In addition, Fujikoshi [14] gave an explicit formula of error bounds for asymptotic approximation of EPMC for LDA. Lachenbruch [26] proposed an approximation

$$e(2|1) \approx \Phi\{-\mathrm{var}(L)^{-1/2}\mathrm{E}(L)\}.$$

Note that

$$\mathrm{E}(L) = \frac{n-2}{2(m-p)}\left\{\Delta^2_{\boldsymbol{\Sigma}^{-1}} + \frac{(n_1-n_2)p}{n_1 n_2}\right\} \ (m>1),$$
$$\mathrm{var}(L) = \frac{(n-2)^2(n-3)}{m(m-1)(m-3)}\left\{\Delta^2_{\boldsymbol{\Sigma}^{-1}} + \frac{np}{n_1 n_2}\right\} \ (m>3),$$

where $m = n-p-2$. Fujikoshi [14] obtained error bounds for this approximation to $e(2|1)$.

Theorem 2.2.2 (Fujikoshi [14]). *Let* $\mathbf{x}, \mathbf{x}_{1j} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathbf{x}_{2j'} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ *for* $j \in \{1,2,\ldots,n_1\}$, $j' \in \{1,2,\ldots,n_2\}$, *and* $\mathbf{x}, \mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ *be mutually independent. If* $m > 7$, *then it holds that*

$$|e(2|1) - \Phi\{-\mathrm{var}(L)^{-1/2}\mathrm{E}(L)\}| \le b,$$

where

$$b = \beta_{2,0}\mathrm{var}(L)^{-1}v_1 + \beta_{2,2}\mathrm{var}(L)^{-2}v_2 + \beta_{2,1}\mathrm{var}(L)^{-3/2}(v_1 v_2)^{1/2}.$$

Here, $\beta_{2,0} = 0.121$, $\beta_{2,1} = 0.2$ *and* $\beta_{2,2} = 0.5$ *and*

$$\begin{aligned}
v_1 =& \frac{(n-2)^2}{2m(m-1)(m-3)}\left[\frac{\Delta^4_{\boldsymbol{\Sigma}^{-1}}}{m-1} + \frac{2(n-3)\Delta^2_{\boldsymbol{\Sigma}^{-1}}}{mn_2}\left\{1 + \frac{n_1-n_2}{(m-1)n_1}\right\}\right.\\
&\left. + \frac{2(n-3)p}{n_1 n_2}\left\{\frac{1}{m} + \frac{(n_1-n_2)^2}{2(m-1)n_1 n_2}\right\}\right],\\
v_2 =& \frac{2(n-3)(n-2)^4}{m(m-1)^2(m-3)^2}\left[\frac{1}{m}\left\{1 + \frac{8(m-4)}{(m-5)(m-7)}\right\}\right.\\
&\times\left\{\frac{p-1}{m}\left(\Delta^2_{\boldsymbol{\Sigma}^{-1}} + \frac{pn}{n_1 n_2}\right)^2 + \frac{n(n-5)}{n_1 n_2}\left(2\Delta^2_{\boldsymbol{\Sigma}^{-1}} + \frac{pn}{n_1 n_2}\right)\right\}\\
&\left. + \frac{4(n-3)(m-4)}{m(m-5)(m-7)}\left(\Delta^2_{\boldsymbol{\Sigma}^{-1}} + \frac{pn}{n_1 n_2}\right)^2\right].
\end{aligned}$$

We get the following corollary from error bound.

Corollary 2.2.1. *Let* c *and* d *be fixed constants which satisfy* $c, d \in (0,1)$. *We assume* $p = \lfloor cn \rfloor$ *and* $n_1 = \lfloor dn \rfloor$. *Then as* $n \to \infty$, *it holds that*

$$e(2|1) = \Phi\{-\mathrm{var}(L)^{-1/2}\mathrm{E}(L)\} + O(n^{-1}).$$

We assume p *is fixed positive integer and* $n_1 = \lfloor dn \rfloor$. *Then as* $n \to \infty$, *it holds that*

$$e(2|1) = \Phi\{-\mathrm{var}(L)^{-1/2}\mathrm{E}(L)\} + O(n^{-1}).$$

From Corollary 2.2.1, we understand that Lachenbruch's approximation is valid not only in the conventional large sample frame work but also in high-dimensional framework.

Chapter 3

High-dimensional discriminant analysis

3.1 Statistical model in high-dimensional settings

In traditional multivariate analysis, the theory is developed assuming some or all of the following conditions.

- The population distribution is a multivariate normal distribution.
- All population covariance matrices are equal.

In fact, there is no guarantee that high-dimensional data follows a multivariate normal distribution, it is difficult problem to ascertain that. Furthermore, the homogeneity of the population covariance matrix is said to be a severe assumption as the dimension increases. From the above it can be said that discussing under the assumption of traditional multivariate analysis is unrealistic in high-dimensional data analysis.

After this chapter, we will discuss the theory under the following statistical model which relax these unrealistic assumptions. Assuming fixed $i \in \{1, 2, \ldots, k\}$, we let

$$\mathbf{x} = \boldsymbol{\Sigma}_i^{1/2}\mathbf{z} + \boldsymbol{\mu}_i.$$

We further assume that

$$\forall_{i\in\{1,2,\ldots,k\},\ j\in\{1,2,\ldots,n_i\}} \quad \mathbf{x}_{ij} = \boldsymbol{\Sigma}_i^{1/2}\mathbf{z}_{ij} + \boldsymbol{\mu}_i.$$

Here, $\boldsymbol{\Sigma}_i$ is positive-semi-definite, and the random vectors

$$\mathbf{z}, \mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{1n_1}, \mathbf{z}_{21}, \mathbf{z}_{22}, \ldots, \mathbf{z}_{2n_2}, \ldots, \mathbf{z}_{k1}, \mathbf{z}_{k2}, \ldots, \mathbf{z}_{kn_k}$$

are independent and identically distributed (i.i.d.) random vectors such that $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{z}) = \mathbf{I}_p$.

3.2 Some properties of scale adjusted-type distance-based classifier

Recentry, Aoshima and Yata [3] discussed a scale adjusted-type distance-based classifier given by Chan and Hall [9]. They introduced the following discriminant rule: one classifies an individual into G_1 if $W > 0$ and into G_2 otherwise, where

$$W = \{2\mathbf{x} - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)\}^\top (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) + \frac{\mathrm{tr}(\mathbf{S}_1)}{n_1} - \frac{\mathrm{tr}(\mathbf{S}_2)}{n_2}. \tag{3.1}$$

Note that (3.1) can be expressed as

$$W = \left\{ \|\mathbf{x} - \overline{\mathbf{x}}_2\|^2 - \frac{\mathrm{tr}(\mathbf{S}_2)}{n_2} \right\} - \left\{ \|\mathbf{x} - \overline{\mathbf{x}}_1\|^2 - \frac{\mathrm{tr}(\mathbf{S}_1)}{n_1} \right\}.$$

Then

$$\mathrm{E}(W) = \begin{cases} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 & \text{when } \mathbf{x} \sim G_1, \\ -\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 & \text{when } \mathbf{x} \sim G_2. \end{cases}$$

Aoshima and Yata [3] developed a scale adjusted-type distance-based classifier that can ensure high accuracy in misclassification rates under some assumptions.

Theorem 3.2.1 (Aoshima and Yata [3]). *We assume the following conditions:*

- $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4 = o(1)$ *as* $p \to \infty$ *for* $i \in \{1, 2\}$,
- $\max_{j\in\{1,2\}} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)/(n_i\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4) = o(1)$ *as* $p \to \infty$ *either when* n_i *is fixed or* $n_i \to \infty$ *for* $i \in \{1, 2\}$.

Then, we have as $p \to \infty$ *that*

$$\frac{W}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2} = \frac{(-1)^{i-1}}{2} + o_p(1)$$

when $\mathbf{x} \sim G_i$ *for* $i \in \{1, 2\}$. *Also, we have as* $p \to \infty$ *that* $e(2|1) = o(1)$ *and* $e(1|2) = o(1)$.

However, these methods are not sufficient methods.

Theorem 3.2.2 (Aoshima and Yata [3]). *Let*

$$\delta_i = 2\left[\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_{i'} + \sum_{\ell=1}^{2} \mathrm{tr}(\boldsymbol{\Sigma}_1^2)/\{2n_\ell(n_\ell - 1)\} \right]^{1/2}$$

for $i \neq i', i, i' \in \{1, 2\}$. *We assume the following conditions:*

- *The fourth moments of each variable in* $\mathbf{z}$ *are uniformuly bounded,* $\mathrm{E}(z_q^2 z_s^2) = 1$ *and* $\mathrm{E}(z_q z_s z_t z_u) = 0$ *for all* $q \neq s, t, u$.

- *The fourth moments of each variable in* $\mathbf{z}_{ij}$ *are uniformuly bounded,* $\mathrm{E}(z_{ijq}^2 z_{ijs}^2) = 1$ *and* $\mathrm{E}(z_{ijq} z_{ijs} z_{ijt} z_{iju}) = 0$ *for all* $q \neq s, t, u$.
- $\mathrm{tr}(\boldsymbol{\Sigma}_i^4)/\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^2 = o(1)$ *and* $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \in (0, \infty)$ *as* $p \to \infty$ *for* $i \in \{1, 2\}$,
- $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\delta_i^2 = o(1)$ *as* $p \to \infty$, $n_1 \to \infty$, *and* $n_2 \to \infty$ *for* $i \in \{1, 2\}$.

Here, for function $f(\cdot)$, *"*$f(p) \in (0, \infty)$ *as* $p \to \infty$*" implies* $\liminf_{p\to\infty} f(p) > 0$ *and* $\limsup_{p\to\infty} f(p) < \infty$. *Then, we have as* $p \to \infty$, $n_1 \to \infty$, *and* $n_2 \to \infty$ *that*

$$(W - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2)/\delta_1 \rightsquigarrow \mathcal{N}(0, 1) \text{ when } \mathbf{x} \sim G_1,$$
$$(W + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2)/\delta_2 \rightsquigarrow \mathcal{N}(0, 1) \text{ when } \mathbf{x} \sim G_2.$$

Also, we have as $p \to \infty$, $n_1 \to \infty$, *and* $n_2 \to \infty$ *that*

$$e(2|1) = \Phi(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/\delta_1) + o(1) \quad \text{and} \quad e(1|2) = \Phi(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/\delta_2) + o(1).$$

Chapter 4

Estimation of misclassification probability for two-class in high-dimensional data

In this chapter, we consider a discriminant problem that allocates a given object $\mathbf{x}$ to one of two populations, G_1 and G_2 in high-dimensional data. Here $\mathbf{x}$ is a continuous random vector (such as an observation vector) represented by a set of features $(x_1, x_2, \ldots, x_p)$. We assume a training data set $(\mathbf{x}_{11},\ \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1},\ \mathbf{x}_{21},\ \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2})$, where $\mathbf{x}_{\ell j}$ is a p-dimensional continuous observation vector from the ℓ-th population G_ℓ, and we calculate

$$\forall_{\ell\in\{1,2\}} \quad \overline{\mathbf{x}}_\ell = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathbf{x}_{\ell j},\ \mathbf{S}_\ell = \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)^\top.$$

We use the following Euclidean distance discriminant function used in Chan and Hall [9] and Aoshima and Yata [3]:

$$W = \|\mathbf{x} - \overline{\mathbf{x}}_2\|^2 - \|\mathbf{x} - \overline{\mathbf{x}}_1\|^2 - \frac{\mathrm{tr}(\mathbf{S}_2)}{n_2} + \frac{\mathrm{tr}(\mathbf{S}_1)}{n_1} \tag{4.1}$$

And we also use a distance discriminant rule that assigns a new observation $\mathbf{x}$ to G_1 if $W > 0$, and to G_2 otherwise. We propose a consistent and asymptotically unbiased estimator of misclassification probability, and compare the MSEs of the estimator that we introduce and the estimator based on leave-one-out cross-validation (CV) in numerical experiment.

4.1 Statistical model

Assuming fixed $g, g' \in \{1, 2\}$ and $g' \neq g$, we let $\mathbf{x} = \mathbf{\Sigma}_g^{1/2}\mathbf{z} + \boldsymbol{\mu}_g$. We further assume that $\forall_{\ell\in\{1,2\},\ j\in\{1,2,\ldots,n_\ell\}}\quad \mathbf{x}_{\ell j} = \mathbf{\Sigma}_\ell^{1/2}\mathbf{z}_{\ell j} + \boldsymbol{\mu}_\ell$. Here, $\mathbf{\Sigma}_\ell$ is positive-semi-definite. The random vectors $\mathbf{z}$, $\mathbf{z}_{11}$,$\mathbf{z}_{12}$,...,$\mathbf{z}_{1n_1}$, $\mathbf{z}_{21}$,$\mathbf{z}_{22}$,...,$\mathbf{z}_{2n_2}$ are independent and identically distributed (i.i.d.) random vectors such that $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{z}) = \mathbf{I}_p$. We denote $\mathbf{z} = (z_1, z_2, \ldots, z_p)^\top$, and consider two cases, (C1) and (C2), as follows.

(C1) $\mathrm{E}(z_i^4) = \kappa_4 + 3 < \infty,\ \mathrm{E}(z_{i_1}^2 z_{i_2}^2) = 1,$ and $\mathrm{E}(z_{i_1} z_{i_2} z_{i_3} z_{i_4}) = 0$ $(i_1 \neq i_2, i_3, i_4)$.

(C2) $z_1, z_2, \ldots, z_p$ are mutually independent, and $\mathrm{E}(z_i^4) = \kappa_4 + 3 < \infty$.

The condition (C1) means that each $\{z_i\}_{i=1}^p$ has a kind of pseudo-independence among its components. Obviously, if (C2) holds, then (C1) is trivially true. Note that (C1) and (C2) include multivariate normal populations.

4.2 Normal approximation of misclassification probability

In this section, we consider the normal approximation of the misclassification probability. It is given by

$$e(g'|g) \approx \Phi\left(-\mu/\sigma_g\right). \tag{4.2}$$

Lemma 4.2.1 (Watanabe, Hyodo, Yamada, and Seo [43]). *μ and σ_g^2 can be written as*

$$\begin{aligned}
&\mu = \mathrm{E}\{(-1)^{g+1} W\} = \boldsymbol{\delta}^\top \boldsymbol{\delta} \qquad \textit{and} \\
&\sigma_g^2 = \mathrm{var}(W) = 4\left\{ \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g \boldsymbol{\delta} + \frac{1}{n_g}\mathrm{tr}(\boldsymbol{\Sigma}_g^2) + \frac{1}{n_g'}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) + \frac{1}{n_g'}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{g'} \boldsymbol{\delta} \right\} \\
&\qquad\qquad + 2\sum_{\ell=1}^{2} \frac{1}{n_\ell(n_\ell - 1)}\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)
\end{aligned}$$

respectively, where $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Proof. Let $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}_g$ and $\mathbf{y}_{\ell j} = \mathbf{x}_{\ell j} - \boldsymbol{\mu}_\ell$ for $\ell \in \{g, g'\}$. Then, $(-1)^{g+1} W$ can be expressed as $(-1)^{g+1} W = \boldsymbol{\delta}^\top \boldsymbol{\delta} + W_1 + W_2$, where

$$\begin{aligned}
W_1 =& 2(-1)^{g+1} \boldsymbol{\delta}^\top \mathbf{y} + 2(\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})^\top \mathbf{y}, \\
W_2 =& 2(-1)^g \boldsymbol{\delta}^\top \overline{\mathbf{y}}_{g'} + \frac{1}{n_{g'}(n_{g'} - 1)} \sum_{j_1, j_2 = 1,\ j_1 \neq j_2}^{n_{g'}} \mathbf{y}_{g' j_1}^\top \mathbf{y}_{g' j_2} \\
&- \frac{1}{n_g(n_g - 1)} \sum_{j_1, j_2 = 1,\ j_1 \neq j_2}^{n_g} \mathbf{y}_{g j_1}^\top \mathbf{y}_{g j_2}.
\end{aligned}$$

Here, $\overline{\mathbf{y}}_\ell = \overline{\mathbf{x}}_\ell - \boldsymbol{\mu}_\ell$. Since $\mathrm{E}(W_1) = \mathrm{E}(W_2) = 0$, we obtain $\mu = \mathrm{E}\{(-1)^{g+1} W\} = \boldsymbol{\delta}^\top \boldsymbol{\delta}$. Since it can be shown that

$$\begin{aligned}
\mathrm{var}\,(W_1) &= 4\left\{ \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g \boldsymbol{\delta} + \frac{1}{n_g}\mathrm{tr}(\boldsymbol{\Sigma}_g^2) + \frac{1}{n_{g'}}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \right\}, \\
\mathrm{var}\,(W_2) &= 2\left\{ \frac{1}{n_g(n_g - 1)}\mathrm{tr}(\boldsymbol{\Sigma}_g^2) + \frac{1}{n_{g'}(n_{g'} - 1)}\mathrm{tr}(\boldsymbol{\Sigma}_{g'}^2) + \frac{2}{n_{g'}}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{g'} \boldsymbol{\delta} \right\},
\end{aligned}$$

and $\mathrm{cov}(W_1, W_2) = 0$, we also obtain the variance of W.

□

The normal approximation is justified under some assumptions. For each $\ell \in \{1,2\}$, let n_ℓ be a function of p, i.e., $n_\ell = n_\ell(p)$. For any $\ell \in \{1,2\}$, let $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_\ell \boldsymbol{\delta}$ and $\mathrm{tr}\{(\boldsymbol{\Sigma}_\ell^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell^{1/2}) \odot (\boldsymbol{\Sigma}_\ell^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell^{1/2})\}$ be a function of p. For any $\ell, \ell', r \in \{1,2\}$, let $\mathrm{tr}\{(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})^r\}$ be a function of p. Then we use the following conditions:

(A0) For all $\ell \in \{1,2\}$, $\lim_{p\to\infty} n_\ell(p) = \infty$,

(A1) For all $\ell \in \{1,2\}$, $\mathrm{tr}(\boldsymbol{\Sigma}_\ell^4)/\left\{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\right\}^2 = o(1)$, $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2) \in (0,\infty)$,

(A2) $\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{g'}\boldsymbol{\delta} = o(n_{g'}\sigma_g^2)$,

(A3) $\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta} = o(\delta_g^2)$,

(A4) $\mathrm{tr}\{(\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g^{1/2}) \odot (\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g^{1/2})\} = o(\sigma_g^4)$.

Here, "$A \odot B$" denotes Hadamard product of same size matrices A and B. For a function $f(\cdot)$, "$f(p) \in (0,\infty)$ as $p \to \infty$" implies $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$. In practical use, the assumption $\mathrm{tr}(\boldsymbol{\Sigma}_\ell^4)/\left\{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\right\}^2 = o(1)$ in (A1) is often not appropriate. This assumption can be called as the non strongly spiked eigenvalue (NSSE) model in Aoshima and Yata [5]. However, it is natural to assume the strongly spiked eigenvalue (SSE) model for microarray data analysis. When NSSE assumption is not satisfied, we recommend a data transformation technique which is proposed in Aoshima and Yata [5]. This transformation reduce the discussion under SSE model to the discussion under NSSE model.

The following theorem represents the asymptotic normality of $(-1)^{g+1}W$.

Theorem 4.2.1 (Watanabe, Hyodo, Yamada, and Seo [43])**.** *We assume (A0)–(A2). Then* (i) *and* (ii) *hold.*

(i) *Under (C1) and (A3),* $\{(-1)^{g+1}W - \mu\}/\delta_g \rightsquigarrow \mathcal{N}(0,1)$ *as* $p \to \infty$.

(ii) *Under (C2) and (A4),* $\{(-1)^{g+1}W - \mu\}/\sigma_g \rightsquigarrow \mathcal{N}(0,1)$ *as* $p \to \infty$.

Here, $\rightsquigarrow$ *denotes that the convergence in distribution.*

Proof. Statement (i) has been demonstrated by Aoshima and Yata [3], and we prove statement (ii).

Under conditions (C2) and (A0)–(A2), W_2 in the proof of Lemma 4.3.1 is negligible. Thus $\{(-1)^{g+1}W - \mu\}/\sigma_g = \sum_{i=1}^p \epsilon_i + o_p(1)$, where $\epsilon_i = 2\{(-1)^{g+1}\boldsymbol{\delta} + (\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})\}^\top\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{e}_i z_i/\sigma_g$. Defining $\mathcal{F}_0 = \sigma\{\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2\}$ and $\mathcal{F}_{i-1} = \sigma\{\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2, z_1, z_2, \ldots, z_{i-1}\}$ $(2 \le i)$, it is straightforward to show that $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{E}(\epsilon_i|\mathcal{F}_{i-1}) = 0$, where $\sigma\{S\}$ means σ-field of sets on Ω generated from S. Thus, ϵ_i is a martingale difference sequence. To show the asymptotic normality of $\sum_{i=1}^p \epsilon_i$, we adapt the martingale difference central-limit theorem (see Shiryaev [36] or Hall and Heyde [17]). Now let $\sigma_{g\cdot i}^2 = \mathrm{E}(\epsilon_i^2|\mathcal{F}_{i-1})$. To apply the martingale central-limit theorem, we need to show that (a):$\sum_{i=1}^p \sigma_{g\cdot i}^2 = 1 + o_p(1)$ and (b):$\sum_{i=1}^p \mathrm{E}(\epsilon_i^4) = o(1)$.

To show (a), we evaluate $\sigma_{g\cdot i}^2 = 4[\{(-1)^{g+1}\boldsymbol{\delta} + (\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})\}^\top \boldsymbol{\Sigma}_g^{1/2}\boldsymbol{e}_i]^2/\sigma_g^2$, and

$$\sum_{i=1}^{p}\sigma_{g\cdot i}^2 = \frac{4}{\sigma_g^2}\left\{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta} + 2(-1)^{g+1}R_1 + R_2\right\}.$$

Here, $R_1 = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g(\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})$ and $R_2 = (\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})^\top\boldsymbol{\Sigma}_g(\overline{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'})$. Since $\mathrm{E}(R_1) = 0$ and $\mathrm{E}(R_2) = \mathrm{tr}(\boldsymbol{\Sigma}_g^2)/n_g + \mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_{g'})/n_{g'}$, we obtain

$$\mathrm{E}\left(\sum_{i=1}^{p}\sigma_{g\cdot i}^2\right) = \frac{4}{\sigma_g^2}\left\{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta} + \frac{1}{n_g}\mathrm{tr}(\boldsymbol{\Sigma}_g^2) + \frac{1}{n_{g'}}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right\} = 1 + o(1).$$

To check (a), we need to show that $\mathrm{var}(R_1) = o(\sigma_g^4)$ and $\mathrm{var}(R_2) = o(\sigma_g^4)$. These variances are given as follows:

$$\begin{aligned}
\mathrm{var}(R_1) &= O\left(\frac{1}{n_g}\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta} + \frac{1}{n_{g'}}\sqrt{\mathrm{tr}\{(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_{g'})^2\}}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}\right),\\
\mathrm{var}(R_2) &= O\left(\frac{1}{n_g^2}\mathrm{tr}(\boldsymbol{\Sigma}_g^4) + \frac{1}{n_{g'}^2}\mathrm{tr}\{(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_{g'})^2\}\right).
\end{aligned}$$

Hence, under (A1), $\mathrm{var}(R_1) = o(\sigma_g^4)$ and $\mathrm{var}(R_2) = o(\sigma_g^4)$. Thus, under (A1), (a) holds.

To show (b), we decompose ϵ_i into the sum of three parts, $\epsilon_i = 2\{(-1)^{g+1}\epsilon_{1i} + \epsilon_{i2} - \epsilon_{i3}\}/\sigma_g$, where $\epsilon_{i1} = \boldsymbol{\delta}^\top\Sigma_g^{1/2}\boldsymbol{e}_i z_i$, $\epsilon_{i2} = \overline{\mathbf{y}}_g^\top\Sigma_g^{1/2}\boldsymbol{e}_i z_i$, and $\epsilon_{i3} = \overline{\mathbf{y}}_{g'}^\top\Sigma_g^{1/2}\boldsymbol{e}_i z_i$. Then, we need to show that $\sum_{i=1}^p \mathrm{E}(\epsilon_{i\ell}^4) = o(\sigma_g^4)$ for $\ell \in \{1,2,3\}$. These expectations are given as follows:

$$\begin{aligned}
&\sum_{i=1}^{p}\mathrm{E}(\epsilon_{i1}^4) = O\left(\mathrm{tr}\{(\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g^{1/2})\odot(\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g^{1/2})\}\right),\\
&\sum_{i=1}^{p}\mathrm{E}(\epsilon_{i2}^4) = O\left(\frac{1}{n_g^2}\mathrm{tr}(\boldsymbol{\Sigma}_g^4)\right),\ \sum_{i=1}^{p}\mathrm{E}(\epsilon_{i3}^4) = O\left(\frac{1}{n_{g'}^2}\mathrm{tr}\{(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_{g'})^2\}\right).
\end{aligned}$$

Thus, $\sum_{i=1}^p \mathrm{E}(\epsilon_{i1}^4) = o(\sigma_g^4)$ under (A4). Also, under (A1), $\sum_{i=1}^p \mathrm{E}(\epsilon_{i2}^4) = o(\sigma_g^4)$ and $\sum_{i=1}^p \mathrm{E}(\epsilon_{i3}^4) = o(\sigma_g^4)$. These results complete the proof.

□

Note that under (A2) and (A3), $\sigma_g = \delta_g + o(\delta_g)$. Thus we obtain the following corollary.

Corollary 4.2.1. *Under (C1) and (A0)–(A3),* $\{(-1)^{g+1}W - \mu\}/\sigma_g \rightsquigarrow \mathcal{N}(0,1)$ *as* $p \to \infty$.

From Theorem 4.2.1 and Corollary 4.2.1, we propose the following proposition. This result represents the accuracy of approximation (4.2).

Proposition 4.2.1. *We assume (A0)–(A2) and* $\mu/\sigma_g = O(1)$. *Then* (i) *and* (ii) *hold.*

(i) $e(g'|g) - \Phi\left(-\mu/\delta_g\right) = \begin{cases} o(1) & \text{under } (C1) \text{ and } (A3).\\ O(1) & \text{under } (C2) \text{ and } (A4).\end{cases}$

(ii) $e(g'|g) - \Phi\left(-\mu/\sigma_g\right) = \begin{cases} o(1) & \textit{under } (C1) \textit{ and } (A3). \\ o(1) & \textit{under } (C2) \textit{ and } (A4). \end{cases}$

Remark 4.2.1 (Watanabe, Hyodo, Yamada, and Seo [43]). *We assume (C1) or (C2). Under* $\mu/\sigma_g \to \infty$, $e(g'|g) = o(1)$.

From Remark 4.2.1, we assume a sufficient condition that guarantees a non-zero limit value of the misclassification probability, i.e., $\mu/\sigma_g = O(1)$.

4.3 Estimator of misclassification probability

Based on Proposition 4.2.1, we approximate the misclassification probability as $\Phi(-\mu/\sigma_g)$. To estimate the unknown values in μ and σ_g, we apply unbiased estimators.

Let $\ell, \ell' \in \{1, 2\}$ and $\ell \neq \ell'$. Preliminarily, we introduce the unbiased estimators of μ, $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$, $\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)$ and $\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}$ as follows:

$$
\begin{aligned}
\widehat{\mu} =& (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) - \frac{1}{n_1}\mathrm{tr}(\mathbf{S}_1) - \frac{1}{n_2}\mathrm{tr}(\mathbf{S}_2), \\
\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)} =& \mathrm{tr}(\mathbf{S}_1\mathbf{S}_2), \\
\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)} =& \frac{(n_\ell - 1)\left[(n_\ell - 1)(n_\ell - 2)\mathrm{tr}(\mathbf{S}_\ell^2) + \{\mathrm{tr}(\mathbf{S}_\ell)\}^2 - n_\ell K_\ell\right]}{n_\ell(n_\ell - 2)(n_\ell - 3)}, \\
\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}} =& (\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{\ell'})^\top\mathbf{S}_\ell(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{\ell'}) - \frac{2U_\ell}{(n_\ell - 1)(n_\ell - 2)} - \frac{\mathrm{tr}(\mathbf{S}_1\mathbf{S}_2)}{n_{\ell'}} \\
& + \frac{2n_\ell K_\ell - (n_\ell - 1)\left\{\mathrm{tr}(\mathbf{S}_\ell)\right\}^2 - (n_\ell - 1)^2\mathrm{tr}(\mathbf{S}_\ell^2)}{n_\ell(n_\ell - 2)(n_\ell - 3)},
\end{aligned}
$$

where

$$
\begin{aligned}
K_\ell =& \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}\|\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell\|^2 \quad \textit{and} \\
U_\ell =& (\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{\ell'})^\top\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)^\top(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell).
\end{aligned}
$$

The unbiased estimator $\widehat{\mu}$ has been used in the L^2 norm based on two-sample test (see for example Chen and Qin [10] or Aoshima and Yata [2]). The unbiased estimator $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)}$ was proposed by Himeno and Yamada [20]. The unbiased estimator $\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}$ is newly derived. To show the consistency of the plug-in estimator based on the normal approximation, we investigate the leading variance term of these estimators.

Lemma 4.3.1 (Watanabe, Hyodo, Yamada, and Seo [43]). *We assume (C1) or (C2). Then (i)–(iv) hold.*

(i) *Under (A0)–(A2),* $\mathrm{var}\left(\widehat{\mu}\right) = o(\sigma_g^2)$,

(ii) *Under (A0) and (A1),* $\mathrm{var}\left(\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}\right) = o(n_{g'}^2\sigma_g^4)$,

(iii) *Under (A0) and (A1),* $\mathrm{var}\left(\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)}\right) = o(n_\ell^2\sigma_\ell^4)$,

(iv) *Under (A0) and (A1),* $\mathrm{var}\left(\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}\right) = o(\sigma_\ell^4)$.

Proof. (i) is obtained by Section 6.1 in Chen and Qin [10]. (iii) is obtained by Lemma 1 in Himeno and Yamada [20]. (ii) is obtained by same route as (iii). We present only the proof of (iv). Let $\mathbf{y}_{\ell j} = \mathbf{x}_{\ell j} - \boldsymbol{\mu}_\ell$ and $\mathbf{y}_{\ell' j} = \mathbf{x}_{\ell' j} - \boldsymbol{\mu}_{\ell'}$. The statistic $\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}$ can be expressed as $\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}} = \sum_{\alpha=1}^{12} A_\alpha$, where

$$
\begin{aligned}
A_1 &= \frac{1}{n_\ell(n_\ell-1)(n_\ell-2)} \sum_{\substack{j_1,j_2,j_3=1 \\ j_1\neq j_2, j_2\neq j_3, j_3\neq j_1}}^{n_\ell} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_2} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_3}, \\
A_2 &= -\frac{1}{n_\ell(n_\ell-1)(n_\ell-2)(n_\ell-3)} \sum_{\substack{j_1,j_2,j_3,j_4=1 \\ j_1\neq j_2\neq j_3\neq j_4 \\ j_3\neq j_1\neq j_4\neq j_2}}^{n_\ell} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_2} \mathbf{y}_{\ell j_3}^\top \mathbf{y}_{\ell j_4}, \\
A_3 &= -\frac{2}{n_\ell(n_\ell-1)} \sum_{\substack{j_1,j_2=1 \\ j_1\neq j_2}}^{n_\ell} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_2} \mathbf{y}_{\ell j_1}^\top \overline{\mathbf{y}}_{\ell'}, \\
A_4 &= \frac{2}{n_\ell(n_\ell-1)(n_\ell-2)} \sum_{\substack{j_1,j_2,j_3=1 \\ j_1\neq j_2, j_2\neq j_3, j_3\neq j_1}}^{n_\ell} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_2} \mathbf{y}_{\ell j_3}^\top \overline{\mathbf{y}}_{\ell'}, \\
A_5 &= \frac{2}{n_\ell n_{\ell'}(n_{\ell'}-1)} \sum_{j_1=1}^{n_\ell} \sum_{\substack{j_2,j_3=1 \\ j_2\neq j_3}}^{n_{\ell'}} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell' j_2} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell' j_3}, \\
A_6 &= -\frac{2}{n_\ell(n_\ell-1)n_{\ell'}(n_{\ell'}-1)} \sum_{\substack{j_1,j_2=1 \\ j_1\neq j_2}}^{n_\ell} \sum_{\substack{j_3,j_4=1 \\ j_3\neq j_4}}^{n_{\ell'}} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell' j_3} \mathbf{y}_{\ell j_2}^\top \mathbf{y}_{\ell' j_4}, \\
A_7 &= \frac{2}{n_\ell(n_\ell-1)} \sum_{\substack{j_1,j_2=1 \\ j_1\neq j_2}}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j_1} \mathbf{y}_{\ell j_1}^\top \mathbf{y}_{\ell j_2}, \\
A_8 &= -\frac{2}{n_\ell(n_\ell-1)(n_\ell-2)} \sum_{\substack{j_1,j_2,j_3=1 \\ j_1\neq j_2, j_2\neq j_3, j_3\neq j_1}}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j_1} \mathbf{y}_{\ell j_2}^\top \mathbf{y}_{\ell j_3}, \\
A_9 &= -\frac{2}{n_\ell} \sum_{j=1}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j} \mathbf{y}_{\ell j}^\top \overline{\mathbf{y}}_{\ell'},
\end{aligned}
$$

$$A_{10} = \frac{2}{n_\ell(n_\ell-1)} \sum_{\substack{j_1,j_2=1 \\ j_1 \neq j_2}}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j_1} \mathbf{y}_{\ell j_2}^\top \overline{\mathbf{y}}_{\ell'},$$

$$A_{11} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j} \mathbf{y}_{\ell j}^\top (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}),$$

$$A_{12} = -\frac{1}{n_\ell(n_\ell-1)} \sum_{\substack{j_1,j_2=1 \\ j_1 \neq j_2}}^{n_\ell} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\top \mathbf{y}_{\ell j_1} \mathbf{y}_{\ell j_2}^\top (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}).$$

The expectations of A_α are derived as $\mathrm{E}(A_\alpha) = 0$ ($\alpha \neq 11$) and $\mathrm{E}(A_{11}) = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_\ell \boldsymbol{\delta}$. The variances of A_α are derived as follows:

$$\begin{aligned}
\mathrm{var}(A_1) =& O\left(\frac{\{\mathrm{tr}\boldsymbol{\Sigma}_\ell^2\}^2}{n_\ell^3} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^4)}{n_\ell^2} \right), \mathrm{var}(A_2) = O\left(\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\}^2}{n_\ell^4} \right), \\
\mathrm{var}(A_3) =& O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})}{n_\ell^2 n_{\ell'}} + \frac{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^4)}\sqrt{\mathrm{tr}\{(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})^2\}}}{n_\ell n_{\ell'}} \right), \\
\mathrm{var}(A_4) =& O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})}{n_\ell^3 n_{\ell'}} \right), \mathrm{var}(A_5) = O\left(\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})\}^2}{n_\ell n_{\ell'}^2} + \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})^2\}}{n_{\ell'}^2} \right), \\
\mathrm{var}(A_6) =& O\left(\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})\}^2}{n_\ell^2 n_{\ell'}^2} \right), \mathrm{var}(A_7) = O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_\ell^2} + \frac{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^4)}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_\ell} \right), \\
\mathrm{var}(A_8) =& O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_\ell^3} \right), \\
\mathrm{var}(A_9) =& O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_\ell n_{\ell'}} + \frac{\sqrt{\mathrm{tr}\{(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})^2\}}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_{\ell'}} \right), \\
\mathrm{var}(A_{10}) =& O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{\Sigma}_{\ell'})\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta}}{n_\ell^2 n_{\ell'}} \right), \mathrm{var}(A_{11}) = O\left(\frac{(\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta})^2}{n_\ell} \right), \\
\mathrm{var}(A_{12}) =& O\left(\frac{(\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_\ell\boldsymbol{\delta})^2}{n_\ell^2} \right).
\end{aligned}$$

Thus $\mathrm{var}(A_\alpha) = o(\sigma_g^4)$ for all $\alpha \in \{1, 2, \ldots, 12\}$. □

These estimators provide the following estimator of σ_g^2:

$$\begin{aligned}
\widehat{\sigma}_g^2 =& 4\left\{ \max(0, \widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}}) + \frac{1}{n_g}\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)} + \frac{1}{n_{g'}}\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)} + \frac{1}{n_{g'}}\max(0, \widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{g'}\boldsymbol{\delta}}) \right\} \\
&+ 2\sum_{\ell=1}^{2} \frac{1}{n_\ell(n_\ell-1)}\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)}.
\end{aligned}$$

Replacing the unknown values μ and σ_g^2 by their estimators, we propose $\widehat{e(g'|g)} = \Phi\left(-\widehat{\mu}/\widehat{\sigma}_g\right)$. The consistency of the estimator $\widehat{e(g'|g)}$ is demonstrated in the following proposition.

Proposition 4.3.1 (Watanabe, Hyodo, Yamada, and Seo [43]). *We assume (A0)–(A2) and* $\mu/\sigma_g = O(1)$. *Then*

$$\widehat{e(g'|g)} = \begin{cases} e(g'|g) + o_p(1) & \textit{under } (C1) \textit{ and } (A3). \\ e(g'|g) + o_p(1) & \textit{under } (C2) \textit{ and } (A4). \end{cases}$$

Proof. We assume (C1) or (C2). From Lemma 4.3.1, under (A0)–(A2),

$$\widehat{\mu} = \mu + o_p(\sigma_g), \tag{4.3}$$

$$\frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}}{n_{g'}} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_{g'}} + o_p(\sigma_g^2), \ \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}}{n_g} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{n_g} + o_p(\sigma_g^2). \tag{4.4}$$

We also note that $|\max(0, \widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}}) - \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}| \leq |\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}} - \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}|$ a.s. From this result and (iv) in Lemma 4.3.1, we get

$$\mathrm{E}\{(\max(0, \widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}}) - \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta})^2\} \leq \mathrm{var}(\widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}}) = o(\sigma_g^4).$$

Hence,

$$\max(0, \widehat{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta}}) = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\delta} + o_p(\sigma_g^2). \tag{4.5}$$

From (4.3), $\widehat{\mu} = \mu + o_p(\sigma_g)$. From (4.4) and (4.5), $\widehat{\sigma}_g^2 = \sigma_g^2 + o_p(\sigma_g^2)$. Thus, under (A0)–(A2),

$$\widehat{w}_g = w_g + o_p(w_g), \tag{4.6}$$

where $w_g = -\mu/\sigma_g$ and $\widehat{w}_g = -\widehat{\mu}/\widehat{\sigma}_g$.

We note that $|e(g'|g) - \Phi(\widehat{w}_g)| \leq |e(g'|g) - \Phi(w_g)| + |\Phi(\widehat{w}_g) - \Phi(w_g)|$. From Proposition 4.2.1, $|e(g'|g) - \Phi(w_g)| = o(1)$. Hence, it is sufficient to show that $|\Phi(\widehat{w}_g) - \Phi(w_g)| = o_p(1)$. From (4.6), we obtain $\widehat{w}_g = w_g + o_p(1)$. By the continuous mapping theorem, we then get $|\Phi(\widehat{w}_g) - \Phi(w_g)| = o_p(1)$. □

From $|\widehat{e(g'|g)} - e(g'|g)| < 1$ and Proposition 4.3.1, we obtain the following corollary.

Corollary 4.3.1 (Watanabe, Hyodo, Yamada, and Seo [43]). *We assume (A0)–(A2) and* $\mu/\sigma_g = O(1)$. *Then*

$$\mathrm{E}\{\widehat{e(g'|g)}\} = \begin{cases} e(g'|g) + o(1) & \textit{under } (C1) \textit{ and } (A3). \\ e(g'|g) + o(1) & \textit{under } (C2) \textit{ and } (A4). \end{cases}$$

4.4 Leave-one-out cross-validation method

In this section, we consider the leave-one-out CV method, which is popularly used for estimating prediction errors in small samples. For $j \in \{1, 2, \ldots, n_g\}$, we consider the set

$$\mathbb{X}_g^{(-j)} = (\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gj-1}, \mathbf{x}_{gj+1}, \ldots, \mathbf{x}_{gn_g}).$$

This set denotes the leave-one-out learning set, which is a collection of data with observation $\mathbf{x}_{gj}$ removed. In a prediction problem, CV calculates the probability of misclassifying a sample from all other observations in the sample. We define the discriminant function by

$$W_g^{(-j)} = \left\{2\mathbf{x}_{gj} - \left(\overline{\mathbf{x}}_{g(-j)} + \overline{\mathbf{x}}_{g'}\right)\right\}^\top \left(\overline{\mathbf{x}}_{g(-j)} - \overline{\mathbf{x}}_{g'}\right) + \left\{\frac{1}{n_g - 1}\mathrm{tr}(\mathbf{S}_{g(-j)}) - \frac{1}{n_{g'}}\mathrm{tr}(\mathbf{S}_{g'})\right\},$$

where $\overline{\mathbf{x}}_{g(-j)}$ and $\mathbf{S}_{g(-j)}$ are calculated by the procedures in (4.1) using the learning set $\mathbb{X}_g^{(-j)}$. The CV-based estimator is then given by

$$c(g'|g) = \frac{1}{n_g}\sum_{j=1}^{n_g} I(W_g^{(-j)} < 0),$$

where the function $I(A)$ is the indicator function defined as

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{if } A \text{ is false.} \end{cases}$$

By straightforward calculation, we obtain

$$\begin{aligned} \mathrm{E}\{c(g'|g)\} =& \Pr(W_g^{(-1)} < 0) \quad \textit{and} \\ \mathrm{var}\{c(g'|g)\} =& \Pr(W_g^{(-1)} < 0, W_g^{(-2)} < 0) - \left\{\Pr(W_g^{(-1)} < 0)\right\}^2 \\ &+ \frac{1}{n_g}\left\{\Pr(W_g^{(-1)} < 0) - \Pr(W_g^{(-1)} < 0, W_g^{(-2)} < 0)\right\}. \end{aligned}$$

Note that $c(g'|g)$ is consistent when

$$\begin{aligned} &\Pr(W_g^{(-1)} < 0) - \Pr(W_g < 0) = o(1) \quad \textit{and} \\ &\Pr(W_g^{(-1)} < 0, W_g^{(-2)} < 0) - \left\{\Pr(W_g^{(-1)} < 0)\right\}^2 = o(1). \end{aligned}$$

To confirm this statement in a high-dimensional setting, we must investigate the distribution of $W_g^{(-1)}$ and the joint distribution of $(W_g^{(-1)}, W_g^{(-2)})^\top$. The joint asymptotic normality of the random vector $(W_g^{(-1)}, W_g^{(-2)})^\top$ is given by the following lemmas:

Lemma 4.4.1 (Watanabe, Hyodo, Yamada, and Seo [43])**.** *Under (C1) and (A0)–(A3),*

$$\left((W_g^{(-1)} - \mu)/\delta_g,\ (W_g^{(-2)} - \mu)/\delta_g\right)^\top \rightsquigarrow \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2).$$

Proof. We assume (C1). Let $k, k' \in \{1, 2\}$ and $k \neq k'$. Then we decompose $W_g^{(-k)} - \mu$ as

$W_{g1}^{(-k)} + W_{g2}^{(-k)}$, where

$$\begin{aligned}
W_{g1}^{(-k)} =& 2\left\{(-1)^{g+1}\boldsymbol{\delta} - \overline{\mathbf{y}}_{g'} + \frac{n_g - 2}{n_g - 1}\overline{\mathbf{y}}_{g(-1,-2)}\right\}^\top \mathbf{y}_{gk}, \\
W_{g2}^{(-k)} =& \frac{2}{n_g - 1}\mathbf{y}_{g1}^\top \mathbf{y}_{g2} - \frac{2}{n_g - 1}\overline{\mathbf{y}}_{g(-1,-2)}^\top \mathbf{y}_{gk'} + 2(-1)^g \boldsymbol{\delta}^\top \overline{\mathbf{y}}_{g'} \\
&+ \frac{1}{n_{g'}(n_{g'} - 1)} \sum_{j_1, j_2 = 1,\ j_1 \neq j_2}^{n_{g'}} \mathbf{y}_{g'j_1}^\top \mathbf{y}_{g'j_2} \\
&- \frac{1}{(n_g - 1)(n_g - 2)} \sum_{j_1, j_2 = 1,\ j_1 \neq j_2,\ j_1, j_2 \neq 1,2}^{n_g} \mathbf{y}_{gj_1}^\top \mathbf{y}_{gj_2}.
\end{aligned}$$

Then it holds that $W_g^{(-k)} - \mu = W_{g1}^{(-k)} + o_p(\sigma_g)$ under (A0)–(A2).

For non-random constants c_1 and c_2, we define $M = c_1 W_{g1}^{(-1)} + c_2 W_{g1}^{(-2)}$. Then

$$\{c_1(W_g^{(-1)} - \mu) + c_2(W_g^{(-2)} - \mu)\}/\sigma_g = M/\sigma_g + o_p(1).$$

The asymptotic normality of M would imply Lemma 4.4.1.

Especially, under (C1) and (A0)–(A3), $M/\sigma_g = \sum_{j=1}^{n_g + n_{g'} - 2} \varepsilon_j + o_p(1)$, where

$$\varepsilon_j = \begin{cases} \dfrac{2\mathbf{y}_{gj+2}^\top (c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2})}{\delta_g (n_g - 1)} & \forall j \in \{1, 2, \ldots, n_g - 2\}, \\ -\dfrac{2\mathbf{y}_{g'j-n_g+2}^\top (c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2})}{\delta_g n_{g'}} & \forall j \in \{n_g - 1, n_g, \ldots, n_g + n_{g'} - 2\}. \end{cases}$$

Define

$$\begin{aligned}
&\mathcal{F}_j = \sigma\{\mathbf{y}_{g1}, \ldots, \mathbf{y}_{gj+2}\} \quad (0 \leq j \leq n_g - 2), \\
&\mathcal{F}_j = \sigma\{\mathbf{y}_{g1}, \ldots, \mathbf{y}_{gn_g}, \mathbf{y}_{g'1}, \ldots, \mathbf{y}_{g'j-n_g+2}\} \quad (n_g - 1 \leq j).
\end{aligned}$$

Then it is straightforward to show that $\mathrm{E}(\varepsilon_j) = 0$ and $\mathrm{E}(\varepsilon_j | \mathcal{F}_{j-1}) = 0$. To apply the martingale central-limit theorem, we need to show that

$$\sum_{j=1}^{n_g + n_{g'} - 2} \mathrm{E}(\varepsilon_j^2 | \mathcal{F}_{j-1}) = c_1^2 + c_2^2 + o_p(1), \quad \sum_{j=1}^{n_g + n_{g'} - 2} \mathrm{E}(\varepsilon_j^4) = o(1). \tag{4.7}$$

First, we check the first part in (4.7). Note that

$$\sum_{j=1}^{n_g + n_{g'} - 2} \mathrm{E}(\varepsilon_j^2 | \mathcal{F}_{j-1}) - (c_1^2 + c_2^2) = \frac{V_1 + V_2}{\delta_g^2} + o_p(1),$$

where

$$\begin{aligned}
V_1 &= \frac{4(n_g - 2)}{(n_g - 1)^2} \left\{(c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2})^\top \boldsymbol{\Sigma}_g (c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2}) - (c_1^2 + c_2^2)\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\right\}, \\
V_2 &= \frac{4}{n_{g'}} \left\{(c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2})^\top \boldsymbol{\Sigma}_{g'} (c_1 \mathbf{y}_{g1} + c_2 \mathbf{y}_{g2}) - (c_1^2 + c_2^2)\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_{g'})\right\}.
\end{aligned}$$

Under (A1),

$$\mathrm{var}(V_1) = O\left(\frac{\mathrm{tr}(\mathbf{\Sigma}_g^4)}{n_g^2}\right) = o(\delta_g^4),\ \mathrm{var}(V_2) = O\left(\frac{\mathrm{tr}\{(\mathbf{\Sigma}_g\mathbf{\Sigma}_{g'})\}^2}{n_{g'}^2}\right) = o(\delta_g^4).$$

Thus, the first part of (4.7) holds.

Next, we show the second part of (4.7). Note that, under (A0),

$$\mathrm{E}(\varepsilon_j^4) = \begin{cases} O\left(\frac{1}{n_g^2}\right) & \forall j \in \{1, 2, \ldots, n_g - 2\} \\ O\left(\frac{1}{n_{g'}^2}\right) & \forall j \in \{n_g - 1, n_g, \ldots, n_g + n_{g'} - 2\}. \end{cases}$$

Thus the second part of (4.7) holds. From these results, the proof is complete.

□

Lemma 4.4.2 (Watanabe, Hyodo, Yamada, and Seo [43])**.** *Under (C2),(A0)–(A2), and (A4),*

$$\left((W_g^{(-1)} - \mu)/\sigma_g,\ (W_g^{(-2)} - \mu)/\sigma_g\right)^\top \rightsquigarrow \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2).$$

Proof. Under (C2) and (A0)–(A2), the random variable M in proof of Lemma 4.4.1 can be factorized as $M/\sigma_g = \sum_{i=1}^p \xi_i$, where

$$\begin{aligned}\xi_i =& \frac{2}{\sigma_g} c_1 \left\{(-1)^{g+1}\boldsymbol{\delta} - \overline{\mathbf{y}}_{g'} + \frac{n_g - 2}{n_g - 1}\widetilde{\mathbf{y}}_g\right\}^\top \mathbf{\Sigma}_g^{1/2}\boldsymbol{e}_i z_{gi1} \\ &+ \frac{2}{\sigma_g} c_2 \left\{(-1)^{g+1}\boldsymbol{\delta} - \overline{\mathbf{y}}_{g'} + \frac{n_g - 2}{n_g - 1}\widetilde{\mathbf{y}}_g\right\}^\top \mathbf{\Sigma}_g^{1/2}\boldsymbol{e}_i z_{gi2},\end{aligned}$$

here $z_{gi1} = \boldsymbol{e}_i^\top \mathbf{z}_{g1}$, $z_{gi2} = \boldsymbol{e}_i^\top \mathbf{z}_{g2}$, and $\widetilde{\mathbf{y}}_g = \overline{\mathbf{y}}_{g(-1,-2)}$. The asymptotic normality of M would imply Lemma 4.4.2. Define

$$\begin{aligned}&\mathcal{F}_0 = \sigma\{\overline{\mathbf{y}}_{g'}, \widetilde{\mathbf{y}}_g\}, \\ &\mathcal{F}_{i-1} = \sigma\{\overline{\mathbf{y}}_{g'}, \widetilde{\mathbf{y}}_g, z_{g11}, \ldots, z_{gi-11}, z_{g12}, \ldots, z_{gi-12}\}\ (2 \le i).\end{aligned}$$

Thus, ξ_i is a martingale difference sequence. To apply the martingale central-limit theorem, we need to show that

$$\sum_{i=1}^p \mathrm{E}(\xi_i^2|\mathcal{F}_{i-1}) = c_1^2 + c_2^2 + o_p(1),\ \sum_{i=1}^p \mathrm{E}(\xi_i^4) = o(1). \tag{4.8}$$

To this end, we show the first part of (4.8). Note that

$$\sum_{i=1}^p \mathrm{E}(\xi_i^2|\mathcal{F}_{i-1}) - (c_1^2 + c_2^2) = 4(c_1^2 + c_2^2)\frac{2(-1)^{g+1}P_1 + P_2}{\sigma_g^2} + o_p(1),$$

where

$$
\begin{aligned}
P_1 =& \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g \left(\frac{n_g - 2}{n_g - 1} \widetilde{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'} \right), \\
P_2 =& \left(\frac{n_g - 2}{n_g - 1} \widetilde{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'} \right)^\top \boldsymbol{\Sigma}_g \left(\frac{n_g - 2}{n_g - 1} \widetilde{\mathbf{y}}_g - \overline{\mathbf{y}}_{g'} \right) \\
& - \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{n_g} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)}{n_{g'}} \right\}.
\end{aligned}
$$

These variances are evaluated as

$$
\begin{aligned}
\mathrm{var}(P_1) &= O\left(\frac{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)} \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g \boldsymbol{\delta}}{n_g} + \frac{\sqrt{\mathrm{tr}\{(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_{g'})^2\}} \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g \boldsymbol{\delta}}{n_{g'}} \right) = o(\sigma_g^4), \\
\mathrm{var}(P_2) &= O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)}{n_g^2} + \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_{g'})^2\}}{n_{g'}^2} \right) = o(\sigma_g^4).
\end{aligned}
$$

Thus, under (A1), the first part of (4.8) holds.

We decompose ξ_i into the sum of three parts, $\xi_i = 2\{(-1)^{g+1}\xi_{1i} + (n_g - 2)/(n_g - 1)\xi_{i2} - \xi_{i3}\}/\sigma_g$, where

$$
\begin{aligned}
&\xi_{i1} = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g^{1/2} \boldsymbol{e}_i (c_1 z_{gi1} + c_2 z_{gi2}),\ \xi_{i2} = \widetilde{\mathbf{y}}_g^\top \boldsymbol{\Sigma}_g^{1/2} \boldsymbol{e}_i (c_1 z_{gi1} + c_2 z_{gi2}), \\
&\xi_{i3} = \overline{\mathbf{y}}_{g'}^\top \boldsymbol{\Sigma}_g^{1/2} \boldsymbol{e}_i (c_1 z_{gi1} + c_2 z_{gi2}).
\end{aligned}
$$

Then, we need to show that $\sum_{i=1}^p \mathrm{E}(\xi_{i\ell}^4) = o(\sigma_g^4)$ for $\ell \in \{1, 2, 3\}$. These expectations are given as follows:

$$
\begin{aligned}
&\sum_{i=1}^p \mathrm{E}(\xi_{i1}^4) = O\left(\mathrm{tr}\{(\boldsymbol{\Sigma}_g^{1/2} \boldsymbol{\delta}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g^{1/2}) \odot (\boldsymbol{\Sigma}_g^{1/2} \boldsymbol{\delta}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_g^{1/2})\} \right), \\
&\sum_{i=1}^p \mathrm{E}(\xi_{i2}^4) = O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)}{n_g^2} \right),\ \sum_{i=1}^p \mathrm{E}(\xi_{i3}^4) = O\left(\frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_{g'})^2\}}{n_{g'}^2} \right).
\end{aligned}
$$

Thus, under (A4), $\sum_{i=1}^p \mathrm{E}(\xi_{i1}^4) = o(\sigma_g^4)$. Also, under (A1), $\sum_{i=1}^p \mathrm{E}(\xi_{i2}^4) = o(\sigma_g^4)$ and $\sum_{i=1}^p \mathrm{E}(\xi_{i2}^4) = o(\sigma_g^4)$. This proves the second part of (4.8). From these results, the proof is complete.

□

From Lemmas 4.4.1 and 4.4.2, we obtain the following proposition.

Proposition 4.4.1 (Watanabe, Hyodo, Yamada, and Seo [43])**.** *We assume (A0)–(A2), and* $\mu/\sigma_g = O(1)$*. Then*

$$
c(g'|g) = \begin{cases} e(g'|g) + o_p(1) & \textit{under } (C1) \textit{ and } (A3), \\ e(g'|g) + o_p(1) & \textit{under } (C2) \textit{ and } (A4). \end{cases}
$$

4.5 Numerical experiment

In Monte Carlo simulations, we investigated the numerical performance of the approximation based on Proposition 4.2.1, and compared the consistencies of the estimators $\widehat{e(2|1)}$ and $c(2|1)$.

4.5.1 Accuracy of normal approximations

First, we investigate the accuracy of the normal approximations:

$$\text{(I)} \ : \ e(2|1) \approx \Phi\left(-\mu/\delta_1\right), \quad \text{(II)} \ : \ e(2|1) \approx \Phi\left(-\mu/\sigma_1\right).$$

Approximation (I) was proposed in Aoshima and Yata [3], and approximation (II) is proposed in the Theorem 4.2.1 (ii). The asymptotic property of these approximations is shown in (i) and (ii) of Proposition 4.2.1. The misclassification probability $e(2|1)$ was calculated in 100,000 replications of the Monte Carlo simulations. In each step, the datasets were generated as

$$\forall_{j\in\{1,\ldots,n_1\}} \quad \mathbf{x}_{1j} = \boldsymbol{\Sigma}_1^{1/2}\mathbf{z}_{1j} + \boldsymbol{\mu}_1, \quad \forall_{j\in\{1,\ldots,n_2\}} \quad \mathbf{x}_{2j} = \boldsymbol{\Sigma}_2^{1/2}\mathbf{z}_{2j} + \boldsymbol{\mu}_2,$$

where $\boldsymbol{\mu}_1 = \mathbf{0}$. In $\boldsymbol{\mu}_2$, the first $\lfloor\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}\rfloor$ elements are $\sqrt{3}n_1^{-1/4}$, and all other elements are 0. Moreover,

$$\boldsymbol{\Sigma}_1 = \mathbf{B}\left(0.3^{|i-j|}\right)\mathbf{B}, \ \boldsymbol{\Sigma}_2 = 1.2\mathbf{B}\left(0.3^{|i-j|}\right)\mathbf{B}.$$

Here,

$$\mathbf{B} = \mathrm{diag}\left(\left(\frac{1}{2}+\frac{1}{p+1}\right)^{\frac{1}{2}}, \left(\frac{1}{2}+\frac{2}{p+1}\right)^{\frac{1}{2}}, \ldots, \left(\frac{1}{2}+\frac{p}{p+1}\right)^{\frac{1}{2}}\right).$$

We considered the following four distributions of $\mathbf{z}_{gj} = (z_{gij})$. Note that the fourth moment of z_{gij} exists.

(A)Standard normal distribution : $z_{gij} \sim \mathcal{N}(0,1)$,

(B)Standardized chi $-$ squared distribution with 10 degrees of freedom :

$z_{gij} = (u_{gij} - 10)/\sqrt{20}$ for $u_{gij} \sim \chi^2_{10}$,

(C)Standardized t distribution with 10 degrees of freedom :

$z_{gij} = u_{gij}/\sqrt{5/4}$ for $u_{gij} \sim t_{10}$,

(D)Standardized skew normal distribution :

$z_{gij} = \{1 - 9/(5\pi)\}^{-1/2}(u_{gij} - 3/\sqrt{5\pi})$ for $u_{gij} \sim \mathcal{SN}(3)$.

Setting $p \in \{50, 100, 200, 400, 800\}$ and $(n_1, n_2) \in \{(20,40), (30,30), (40,20), (40,80), (60,60), (80,40)\}$, we compared the $e(2|1)$ values calculated by the simulation, approximation (I), and approximation (II). The results are shown in Table 4.1. Comparing the tabulated approximations, we observe that in most cases, approximation (II) more closely approaches $e(2|1)$ than approximation (I). In addition, approximation (II) exhibits high stability when we vary the population distribution.

Table 4.1: Comparison of approximations

			(n_1, n_2)					
p			(20,40)	(30,30)	(40,20)	(40,80)	(60,60)	(80,40)
50	$e(2\|1)$	(A)	0.2071	0.2354	0.2696	0.2270	0.2529	0.2846
		(B)	0.2057	0.2332	0.2686	0.2246	0.2530	0.2847
		(C)	0.2025	0.2332	0.2685	0.2272	0.2507	0.2856
		(D)	0.2038	0.2355	0.2678	0.2251	0.2555	0.2842
	approx	(I)	0.1212	0.1590	0.2117	0.1199	0.1579	0.2102
		(II)	0.2072	0.2351	0.2682	0.2268	0.2542	0.2830
100	$e(2\|1)$	(A)	0.1908	0.2243	0.2616	0.2072	0.2385	0.2739
		(B)	0.1898	0.2185	0.2598	0.2071	0.2360	0.2694
		(C)	0.1915	0.2238	0.2584	0.2096	0.2369	0.2710
		(D)	0.1884	0.2234	0.2623	0.2087	0.2386	0.2708
	approx	(I)	0.1283	0.1662	0.2186	0.1269	0.1651	0.2171
		(II)	0.1922	0.2224	0.2595	0.2084	0.2382	0.2712
200	$e(2\|1)$	(A)	0.1689	0.2049	0.2478	0.1842	0.2178	0.2562
		(B)	0.1685	0.2033	0.2439	0.1861	0.2151	0.2529
		(C)	0.1686	0.2024	0.2456	0.1859	0.2154	0.2547
		(D)	0.1695	0.2045	0.2451	0.1846	0.2162	0.2530
	approx	(I)	0.1218	0.1597	0.2123	0.1205	0.1585	0.2108
		(II)	0.1709	0.2031	0.2439	0.1842	0.2162	0.2534
400	$e(2\|1)$	(A)	0.1634	0.1982	0.2393	0.1745	0.2092	0.2471
		(B)	0.1623	0.1986	0.2427	0.1742	0.2061	0.2495
		(C)	0.1641	0.1996	0.2402	0.1747	0.2079	0.2481
		(D)	0.1624	0.1949	0.2438	0.1727	0.2072	0.2493
	approx	(I)	0.1286	0.1666	0.2189	0.1273	0.1654	0.2174
		(II)	0.1639	0.1976	0.2412	0.1740	0.2075	0.2479
800	$e(2\|1)$	(A)	0.1468	0.1849	0.2299	0.1569	0.1893	0.2350
		(B)	0.1492	0.1821	0.2289	0.1537	0.1909	0.2331
		(C)	0.1500	0.1850	0.2299	0.1561	0.1927	0.2355
		(D)	0.1490	0.1835	0.2296	0.1543	0.1905	0.2345
	approx	(I)	0.1220	0.1598	0.2125	0.1207	0.1587	0.2110
		(II)	0.1484	0.1832	0.2293	0.1560	0.1908	0.2343

4.5.2 Accuracy of the estimators

The MSEs of both estimators are listed in Table 4.2. In all cases, the estimator $\widehat{e(2|1)}$ gives a smaller MSE than the estimator $c(2|1)$. Based on these simulation experiments, we therefore recommend estimator $\widehat{e(2|1)}$.

4.6 Conclusion

We proposed consistent and asymptotically unbiased estimators of misclassification probabilities in high-dimensional settings. Our proposed estimator was obtained by using a normal approximation of the misclassification probability. We confirmed the consistency of the proposed estimator under variance heterogeneity and non-normality (Proposition 4.3.1). We also showed the consistency of an estimator based on the leave-one-out CV method (Proposition 4.4.1). The MSEs of the two estimators were compared in numerical simulations. The estimator based on the normal approximation proved more accurate than the estimator based on leave-one-out CV.

Table 4.2: Comparison of MSE$\times 10^3$ of estimators

			(n_1, n_2)					
p			(20,40)	(30,30)	(40,20)	(40,80)	(60,60)	(80,40)
50	(A)	$\widehat{e(2\|1)}$	**3.961**	**3.600**	**4.413**	**1.982**	**1.768**	**2.130**
		$c(2\|1)$	6.515	5.493	5.956	3.361	2.786	2.980
	(B)	$\widehat{e(2\|1)}$	**3.939**	**3.636**	**4.394**	**1.996**	**1.784**	**2.127**
		$c(2\|1)$	6.486	5.517	5.911	3.373	2.833	2.990
	(C)	$\widehat{e(2\|1)}$	**4.024**	**3.637**	**4.435**	**2.012**	**1.819**	**2.132**
		$c(2\|1)$	6.501	5.481	5.948	3.366	2.787	2.982
	(D)	$\widehat{e(2\|1)}$	**3.939**	**3.569**	**4.432**	**1.985**	**1.776**	**2.120**
		$c(2\|1)$	6.476	5.449	5.977	3.371	2.807	2.965
100	(A)	$\widehat{e(2\|1)}$	**3.748**	**3.555**	**4.459**	**1.963**	**1.786**	**2.156**
		$c(2\|1)$	6.296	5.445	6.047	3.293	2.772	2.988
	(B)	$\widehat{e(2\|1)}$	**3.786**	**3.520**	**4.448**	**1.982**	**1.803**	**2.177**
		$c(2\|1)$	6.308	5.375	5.960	3.266	2.780	2.981
	(C)	$\widehat{e(2\|1)}$	**3.758**	**3.665**	**4.464**	**1.984**	**1.802**	**2.163**
		$c(2\|1)$	6.280	5.389	5.955	3.316	2.770	2.991
	(D)	$\widehat{e(2\|1)}$	**3.796**	**3.556**	**4.388**	**1.984**	**1.777**	**2.163**
		$c(2\|1)$	6.352	5.389	5.926	3.321	2.771	2.953
200	(A)	$\widehat{e(2\|1)}$	**3.453**	**3.337**	**4.264**	**1.842**	**1.690**	**2.105**
		$c(2\|1)$	6.023	5.179	5.831	3.123	2.621	2.912
	(B)	$\widehat{e(2\|1)}$	**3.376**	**3.509**	**4.303**	**1.885**	**1.730**	**2.120**
		$c(2\|1)$	5.997	5.180	5.766	3.106	2.643	2.870
	(C)	$\widehat{e(2\|1)}$	**3.432**	**3.365**	**4.327**	**1.852**	**1.707**	**2.109**
		$c(2\|1)$	5.950	5.161	5.842	3.122	2.619	2.883
	(D)	$\widehat{e(2\|1)}$	**3.431**	**3.374**	**4.275**	**1.866**	**1.718**	**2.120**
		$c(2\|1)$	5.977	5.168	5.766	3.082	2.640	2.893
400	(A)	$\widehat{e(2\|1)}$	**3.703**	**3.274**	**4.346**	**1.781**	**1.696**	**2.155**
		$c(2\|1)$	5.922	5.176	5.890	3.043	2.624	2.935
	(B)	$\widehat{e(2\|1)}$	**3.283**	**3.279**	**4.270**	**1.793**	**1.714**	**2.150**
		$c(2\|1)$	5.910	5.194	5.836	3.039	2.627	2.925
	(C)	$\widehat{e(2\|1)}$	**3.273**	**3.285**	**4.263**	**1.776**	**1.708**	**2.121**
		$c(2\|1)$	5.894	5.187	5.876	3.050	2.635	2.903
	(D)	$\widehat{e(2\|1)}$	**3.312**	**3.359**	**4.406**	**1.785**	**1.706**	**2.138**
		$c(2\|1)$	5.956	5.234	5.834	3.055	2.622	2.922
800	(A)	$\widehat{e(2\|1)}$	**2.972**	**3.034**	**4.118**	**1.616**	**1.620**	**2.079**
		$c(2\|1)$	5.628	4.928	5.723	2.853	2.525	2.840
	(B)	$\widehat{e(2\|1)}$	**2.961**	**3.106**	**4.124**	**1.638**	**1.627**	**2.090**
		$c(2\|1)$	5.575	4.953	5.672	2.868	2.540	2.864
	(C)	$\widehat{e(2\|1)}$	**2.917**	**3.053**	**4.117**	**1.620**	**1.603**	**2.083**
		$c(2\|1)$	5.538	4.955	5.707	2.835	2.513	2.846
	(D)	$\widehat{e(2\|1)}$	**2.927**	**3.060**	**4.128**	**1.618**	**1.607**	**2.092**
		$c(2\|1)$	5.575	4.958	5.725	2.832	2.504	2.862

Chapter 5

Estimation of misclassification probability for multi-class in high-dimensional data

In this chapter, we extend the contents of Chapter 4 to multiple groups $G_1, G_2, \ldots, G_k$. Training data set is extend to $(\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}, \ldots, \mathbf{x}_{k1}, \mathbf{x}_{k2}, \ldots, \mathbf{x}_{kn_k})$, where $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}$ is a p-dimensional random sample from the i-th population G_i. For $p \leq \sum_{i=1}^{k} n_i - k$, a natural extension of Fisher linear discriminant exists using multiple discriminant analysis (See, Johnson and Wichern [24]). However, when $p > \sum_{i=1}^{k} n_i - k$, it cannot be used due to the singularity of pooled sample covariance matrix. In this case, the Euclidean distance-based classifier is often used. Let $\mathbf{x}$ be test data generated from one of the several populations $G_1, G_2, \ldots, G_k$. The Euclidean distance-based discriminant function is defined as

$$W_{ji} \;=\; \|\mathbf{x} - \overline{\mathbf{x}}_j\|^2 - \|\mathbf{x} - \overline{\mathbf{x}}_i\|^2 - \frac{\mathrm{tr}(\mathbf{S}_j)}{n_j} + \frac{\mathrm{tr}(\mathbf{S}_i)}{n_i}$$

for $i \neq j,\ i, j \in \{1, 2, \ldots, k\}$, where $\mathbf{S}_i$ and $\mathbf{S}_j$ are the sample covariance matrices. Using this function, the classification rule for test data $\mathbf{x}$ is given by

$$\mathbf{x} \in \mathcal{R}_q \Rightarrow \mathbf{x} \sim G_q,$$

where the region $\mathcal{R}_q$ ($q \in \{1, 2, \ldots, k\}$) is defined by

$$\mathcal{R}_q = \{\mathbf{x} \in \mathbb{R}^p \;;\; W_{jq} > 0,\ j = 1, 2, \ldots, k,\ j \neq q\},$$

where the notation "$\mathbf{x} \sim G_\ell$" means $\mathbf{x}$ generated from G_ℓ. Then, the misclassification probability of an observation from G_q is

$$e_q = 1 - \Pr(\mathbf{x} \in \mathcal{R}_q | \mathbf{x} \sim G_q).$$

However, it is generally difficult to obtain an exact value for e_q. In Chapter 4, we obtained a plug-in estimator contained in the approximate value of e_q for two-class classification. To extend that to multi-class, we show the asymptotic multivariate normality for $(W_{1q}, \ldots, W_{q-1q}, W_{q+1q}, \ldots, W_{kq})^\top$. By using this result, it is possible to construct an approximation that is not the upper bound of e_q. Further, we propose the plug-in type estimator of misclassification probability of e_q using asymptotic multivariate normality.

5.1 Statistical model

For simplicity of notation, we only deal with e_1. Then, $q = 1$ and $j \in \{2, 3, \ldots, k\}$. As a natural extension in the two-class classification, we assume that the data set is generated by the following model:

$$\mathbf{x} = \mathbf{\Sigma}_1^{1/2}\mathbf{z}_{10} + \boldsymbol{\mu}_1, \ \forall_{\ell\in\{1,2,\ldots,k\},t\in\{1,2,\ldots,n_\ell\}} \ \mathbf{x}_{\ell t} = \mathbf{\Sigma}_\ell^{1/2}\mathbf{z}_{\ell t} + \boldsymbol{\mu}_\ell,$$

where $\boldsymbol{e}_s^\top \mathbf{z}_{\ell t}$ are iid random variables s.t. fourth moment is bounded. Under this model, the population mean vector and covariance matrix of $\mathbf{x}_{\ell 1}$ are $\mathrm{E}(\mathbf{x}_{\ell 1}) = \boldsymbol{\mu}_\ell$ and $\mathrm{var}(\mathbf{x}_{\ell 1}) = \mathbf{\Sigma}_\ell$, respectively. Let $\forall_{i,i'\in\{1,2,\ldots,k\}} \ \boldsymbol{\delta}_{ii'} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}$. The mean of W_{j1}, variance of W_{j1}, and covariance of $(W_{j1}, W_{j'1})$ for $j, j' \in \{2, 3, \ldots, k\}$ are

$$\begin{aligned}
\mu_j &= \mathrm{E}(W_{j1}) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_j\|^2 = \|\boldsymbol{\delta}_{1j}\|^2, \\
\sigma_j^2 &= \mathrm{var}(W_{j1}) = 4\left\{\Delta_{1j} + \frac{\mathrm{tr}(\mathbf{\Sigma}_1^2)}{n_1} + \frac{\mathrm{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_j) + \Delta_{j1}}{n_j}\right\} + \frac{2\mathrm{tr}(\mathbf{\Sigma}_1^2)}{n_1(n_1-1)} \\
&\qquad + \frac{2\mathrm{tr}(\mathbf{\Sigma}_j^2)}{n_j(n_j-1)}, \\
\sigma_{jj'} &= \mathrm{cov}(W_{j1}, W_{j'1}) = 4\left(\Delta_{1jj'} + \frac{\mathrm{tr}(\mathbf{\Sigma}_1^2)}{n_1}\right),
\end{aligned}$$

respectively.Here, $\Delta_{i_1 i_2} = \|\mathbf{\Sigma}_{i_1}^{1/2} \ \boldsymbol{\delta}_{i_1 i_2}\|^2$ and $\Delta_{i_1 i_2 i_3} = \ \boldsymbol{\delta}_{i_1 i_2}^\top \mathbf{\Sigma}_{i_1} \ \boldsymbol{\delta}_{i_1 i_3}$ for $i_1, i_2, i_3 \in \mathbb{N}$. μ_j and σ_j^2 are the same as μ and σ_g^2 in Lemma 4.3.1.

5.2 Normal approximation of misclassification probability

To obtain asymptotic normality, we make asymptotic frameworks for some parameters. Let n_j, $\mathrm{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_j)$, $\mathrm{tr}\{(\mathbf{\Sigma}_1\mathbf{\Sigma}_j)^2\}$, $\mathrm{tr}\{(\Sigma_1^{1/2}\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Sigma_1^{1/2}) \odot (\Sigma_1^{1/2}\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Sigma_1^{1/2})\}$, $\mathrm{tr}(\mathbf{\Sigma}_1^2)$, $\mathrm{tr}(\mathbf{\Sigma}_1^4)$, $\mathrm{tr}(\mathbf{\Sigma}_j^2)$ and Δ_{1j} be functions of p for $j \in \{2, 3, \ldots, k\}$. Then, we assume (B1)–(B3).

(B1) $\min\{n_1, n_2, \ldots, n_k\} \to \infty$ and $n_j/n_1 \in (0, \infty)$.

(B2) $\mathrm{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_j)/\{\mathrm{tr}(\mathbf{\Sigma}_1^2)\}^2 \in (0, \infty)$, $\Delta_{j1}/\Delta_{1j} \in (0, \infty)$, $\{\mathrm{tr}(\mathbf{\Sigma}_j^2)\}^2/\{\mathrm{tr}(\mathbf{\Sigma}_1^2)\}^2 \in (0, \infty)$.

(B3) $\mathrm{tr}(\mathbf{\Sigma}_1^4) = o(\{\mathrm{tr}(\mathbf{\Sigma}_1^2)\}^2)$, $\sqrt{\mathrm{tr}\{(\mathbf{\Sigma}_1\mathbf{\Sigma}_j)^2\}} = o(\mathrm{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_j))$,
$\mathrm{tr}\{(\Sigma_1^{1/2}\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Sigma_1^{1/2}) \odot (\Sigma_1^{1/2}\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Sigma_1^{1/2})\} = o(\Delta_{1j})$,

where for a function $f(\cdot)$,"$f(p) \in (0, \infty)$ as $p \to \infty$" implies $\liminf_{p\to\infty} f(p) > 0$, $\limsup_{p\to\infty} f(p) < \infty$.

We consider the standardized Euclidean discriminant functions as follows:

$$T_j = \frac{W_{j1} - \mu_j}{\sigma_j}, \qquad \text{for } \ j \in \{2, 3, \ldots, k\}.$$

We show the asymptotic normality of $\boldsymbol{T} = (T_2, T_3, \ldots, T_k)^\top$. Then, mean vector $\mathbf{E}(\boldsymbol{T}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{T}) = (\rho_{jj'}) =: \mathbf{R}$, where $\rho_{jj'} = \sigma_{jj'}/(\sigma_j\sigma_{j'})$. The multivariate asymptotic normality of $\boldsymbol{T}$ is given by the following theorem.

Theorem 5.2.1 (Watanabe, Hyodo, and Seo [42]). *Under (B1)–(B3),* $\boldsymbol{T} \rightsquigarrow \mathcal{N}_{k-1}(\mathbf{0}, \lim_{p\to\infty} \mathbf{R})$.

Proof. From Cramér–Wold theorem(See, Cramér and Wold [11]), it is sufficient to show $\boldsymbol{\beta}^\top \boldsymbol{T} \rightsquigarrow \mathcal{N}(0, \lim_{p\to\infty} \boldsymbol{\beta}^\top \mathbf{R}\boldsymbol{\beta})$ for any $k-1$ dimensional nonrandom vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{k-1})^\top \in \mathbb{R}^{k-1}/\{\mathbf{0}\}$. We define

$$\epsilon_s = 2\sum_{j=2}^{k} \beta_j \sigma_j^{-1} (\boldsymbol{\delta}_{1j}^\top \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{e}_s + \overline{\mathbf{z}}_1^\top \boldsymbol{\Sigma}_1 \boldsymbol{e}_s - \overline{\mathbf{z}}_j^\top \boldsymbol{\Sigma}_j^{1/2} \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{e}_s) z_s,$$

where $z_s = \boldsymbol{e}_s^\top \mathbf{z}_{10}$ and $\overline{\mathbf{z}}_i = n_i^{-1} \sum_{t=1}^{n_i} \mathbf{z}_{it}$ for $i \in \{1, 2, \ldots, k\}$. Under (B1) and (B2), $\boldsymbol{\beta}^\top \boldsymbol{T} = \sum_{s=1}^{p} \epsilon_s + o_p(1)$. Let $\mathcal{F}_0 = \sigma\{\overline{\mathbf{z}}_1, \overline{\mathbf{z}}_2, \ldots, \overline{\mathbf{z}}_k\}$ and $\mathcal{F}_{s-1} = \sigma\{\overline{\mathbf{z}}_1, \overline{\mathbf{z}}_2, \ldots, \overline{\mathbf{z}}_k, z_1, z_2, \ldots, z_{s-1}\}$ for $s \geq 2$. Then, (ϵ_s) is a martingale difference sequence. Under (B1) and (B2), there exists $\lim_{p\to\infty} \boldsymbol{\beta}^\top \mathbf{R}\boldsymbol{\beta} (\in (0, \infty))$. Let $\sigma^2 = \lim_{p\to\infty} \boldsymbol{\beta}^\top \mathbf{R}\boldsymbol{\beta}$. Also, under (B1) and (B3),

$$\sum_{s=1}^{p} \mathrm{E}(\epsilon_s^2 | \mathcal{F}_{s-1}) = \sigma^2 + o_p(1), \quad \sum_{s=1}^{p} \mathrm{E}(\epsilon_s^4) = o(1).$$

Applying the martingale central limit theorem(See, Hall and Heyde [17]), we prove asymptotic normality of $\boldsymbol{\beta}^\top \boldsymbol{T}$. □

5.3 Estimator of misclassification probability

Using Theorem 5.2.1, we propose the asymptotic approximation of misclassification probability as follows:

$$\tilde{e}_1 = 1 - F(\boldsymbol{r}, \mathbf{R}), \tag{5.1}$$

where $F(\boldsymbol{r}, \mathbf{R}) = \int_{\mathcal{D}} (2\pi)^{-(k-1)/2} |\mathbf{R}|^{-1/2} e^{-\boldsymbol{w}^\top \mathbf{R}^{-1} \boldsymbol{w}/2} d\boldsymbol{w}$. Here,

$$\mathcal{D} = \{\boldsymbol{w} \in \mathbb{R}^{k-1} \ ; \ \boldsymbol{e}_1^\top \boldsymbol{w} + r_1, \boldsymbol{e}_2^\top \boldsymbol{w} + r_2, \ldots, \boldsymbol{e}_{k-1}^\top \boldsymbol{w} + r_{k-1} > 0\}$$

and $\boldsymbol{r} = (\mu_2/\sigma_2, \mu_3/\sigma_3, \ldots, \mu_k/\sigma_k)^\top$.

The approximation (5.1) includes the unknown values $\|\boldsymbol{\delta}_{1j}\|^2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_j)$, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$, Δ_{1j},

Δ_{j1} and $\Delta_{1jj'}$. We prepare unbiased estimators of these unknown values as follows:

$$\begin{aligned}
\widehat{\mu}_j &= \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_j\|^2 - \frac{\mathrm{tr}(\mathbf{S}_1)}{n_1} - \frac{\mathrm{tr}(\mathbf{S}_j)}{n_j},\ \widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_j)} = \mathrm{tr}(\mathbf{S}_1\mathbf{S}_j),\\
\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} &= \frac{(n_i-1)[(n_i-1)(n_i-2)\mathrm{tr}(\mathbf{S}_i^2) + \{\mathrm{tr}(\mathbf{S}_i)\}^2 - n_iK_i]}{n_i(n_i-2)(n_i-3)},\\
\widehat{\Delta}_{1j} &= V_{1jj} - \frac{2U_{1j}}{(n_1-1)(n_1-2)} - \frac{\mathrm{tr}(\mathbf{S}_1\mathbf{S}_j)}{n_1}\\
&\quad + \frac{2n_1K_1 - (n_1-1)\{\mathrm{tr}(\mathbf{S}_1)\}^2 - (n_1-1)^2\mathrm{tr}(\mathbf{S}_1^2)}{n_1(n_1-2)(n_1-3)},\\
\widehat{\Delta}_{j1} &= V_{j11} - \frac{2U_{j1}}{(n_j-1)(n_j-2)} - \frac{\mathrm{tr}(\mathbf{S}_j\mathbf{S}_1)}{n_j}\\
&\quad + \frac{2n_jK_j - (n_j-1)\{\mathrm{tr}(\mathbf{S}_j)\}^2 - (n_j-1)^2\mathrm{tr}(\mathbf{S}_j^2)}{n_j(n_j-2)(n_j-3)},\\
\widehat{\Delta}_{1jj'} &= V_{1jj'} - \frac{U_{1j}+U_{1j'}}{(n_1-1)(n_1-2)}\\
&\quad + \frac{2n_1K_1 - (n_1-1)\{\mathrm{tr}(\mathbf{S}_1)\}^2 - (n_1-1)^2\mathrm{tr}(\mathbf{S}_1^2)}{n_1(n_1-2)(n_1-3)},
\end{aligned}$$

where for $i_1, i_2, i_3 \in \{1, 2, \ldots, k\}$,

$$\begin{aligned}
K_{i_1} &= \frac{1}{n_{i_1}-1}\sum_{t=1}^{n_{i_1}} \|\mathbf{x}_{i_1t} - \overline{\mathbf{x}}_{i_1}\|^2,\\
V_{i_1i_2i_3} &= (\overline{\mathbf{x}}_{i_1} - \overline{\mathbf{x}}_{i_2})^\top \mathbf{S}_{i_1}(\overline{\mathbf{x}}_{i_1} - \overline{\mathbf{x}}_{i_3}),\\
U_{i_1i_2} &= (\overline{\mathbf{x}}_{i_1} - \overline{\mathbf{x}}_{i_2})^\top \sum_{t=1}^{n_{i_1}} (\mathbf{x}_{i_1t} - \overline{\mathbf{x}}_{i_1})(\mathbf{x}_{i_1t} - \overline{\mathbf{x}}_{i_1})^\top (\mathbf{x}_{i_1t} - \overline{\mathbf{x}}_{i_1}).
\end{aligned}$$

$\widehat{\mu}_j$, $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_j)}$, $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$, $\widehat{\Delta}_{1j}$, and $\widehat{\Delta}_{j1}$ are the same as the estimators in the section 4.3. The unbiased estimator $\widehat{\Delta}_{1jj'}$ is newly obtained in this section. However, some estimators do not always take appropriate values. We note that $\Delta_{1j}, \Delta_{1j'}, \Delta_{j1} > 0$ and $\Delta_{1jj'} \in [-\sqrt{\Delta_{1j}\Delta_{1j'}}, \sqrt{\Delta_{1j}\Delta_{1j'}}]$. We truncate estimators of these parameters so that they take appropriate values. Thus, we obtain $\widetilde{\Delta}_{1j} = \max\{0, \widehat{\Delta}_{1j}\}$, $\widetilde{\Delta}_{1j'} = \max\{0, \widehat{\Delta}_{1j'}\}$, $\widetilde{\Delta}_{j1} = \max\{0, \widehat{\Delta}_{j1}\}$, and

$$\widetilde{\Delta}_{1jj'} = \min\left\{\sqrt{\widetilde{\Delta}_{1j}\widetilde{\Delta}_{1j'}}, \max\left\{\widehat{\Delta_{1jj'}}, -\sqrt{\widetilde{\Delta}_{1j}\widetilde{\Delta}_{1j'}}\right\}\right\}.$$

We estimate σ_j^2 and $\sigma_{jj'}$ using the following estimators:

$$\begin{aligned}
\widehat{\sigma}_j^2 &= 4\left(\widetilde{\Delta}_{1j} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}}{n_1} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_j)} + \widetilde{\Delta}_{j1}}{n_j}\right) + 2\sum_{i\in\{1,j\}} \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}}{n_i(n_i-1)},\\
\widehat{\sigma}_{jj'} &= 4\left(\widetilde{\Delta}_{1jj'} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}}{n_1}\right).
\end{aligned}$$

Let $\widehat{\rho}_{jj'} = \widehat{\sigma}_{jj'}/(\widehat{\sigma}_j\widehat{\sigma}_{j'})$. Replacing the unknown value $\rho_{jj'}$ in $\mathbf{R}$, we obtain the estimator $\widehat{\mathbf{R}} = (\widehat{\rho}_{jj'})$. We note that the matrix $\widehat{\mathbf{R}}$ is always a positive matrix. Moreover, we estimate $\boldsymbol{r}$: $\widehat{\boldsymbol{r}} = (\widehat{\mu}_2/\widehat{\sigma}_2, \widehat{\mu}_3/\widehat{\sigma}_3, \ldots, \widehat{\mu}_k/\widehat{\sigma}_k)^\top$. By substituting each estimator of unknown value in (5.1), we obtain the estimator of misclassification probability as follows:

$$\widehat{e_1} \;=\; 1 - F\left(\widehat{\boldsymbol{r}}, \widehat{\mathbf{R}}\right). \tag{5.2}$$

5.4 Numerical experiment

We compare the approximation accuracy of the method based on (5.1) and the method proposed by previous research, and compare the MSE of the plug-in estimator $\widehat{e_1}$ and the estimator based on the leave-one-out CV method. For simplicity, we treat the discrimination problem among three groups.

5.4.1 Accuracy of normal approximations

We investigate the accuracy of the asymptotic approximations

$$\text{(MI)} \;:\; e_1 \approx \tilde{e}_1, \quad \text{(MII)} \;:\; e_1 \approx \sum_{j=2}^{3} \Phi\left(-\frac{\|\boldsymbol{\delta}_{1j}\|^2}{\delta_j}\right),$$

where the approximation (MI) represents our proposed method based on (5.1), and the approximation (MII) represents the method proposed by Aoshima and Yata [3]. (MI) is derived by using the asymptotic multivariate normality for $(W_{21}, W_{31}, \ldots, W_{k1})^\top$ which is obtained Theorem 5.2.1. (MII) is the approximation of the upper bound of e_1 by combining the asymptotic normality of W_{ij} and Boole's inequality. This approximation is valid under some regularity conditions. Note that (MI) approximates e_1 directly, whereas (MII) approximates the upper bound of e_1.

The misclassification probability e_1 is calculated via Monte Carlo simulation with 100,000 replications.
For the distribution of $\mathbf{z}_{it} = (z_{i\ell t})$, we set the following two distributions:

$$\begin{aligned}
&\text{(D1)} \quad z_{i\ell t} \sim \mathcal{N}(0,1),\\
&\text{(D2)} \quad z_{i\ell t} = u_{i\ell t}/\sqrt{5/4} \text{ for } u_{i\ell t} \sim t_{10}.
\end{aligned}$$

Note that (D1) and (D2) satisfy our moment condition. The structure of the covariance matrix is set with the following:

$$\boldsymbol{\Sigma}_1 = \left(0.3^{|i-j|}\right), \; \boldsymbol{\Sigma}_2 = 1.2\left(0.3^{|i-j|}\right), \; \boldsymbol{\Sigma}_3 = 2.4\left(0.3^{|i-j|}\right).$$

We set the mean vectors as following two cases:

$$\begin{aligned}
&\text{(M1)} \quad \boldsymbol{\mu}_1 = \mathbf{0}, \; \boldsymbol{\mu}_2 = \left(\sqrt{30/p}, \sqrt{30/p}, \ldots, \sqrt{30/p}\right)^\top, \; \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2,\\
&\text{(M2)} \quad \boldsymbol{\mu}_1 = \mathbf{0}, \; \boldsymbol{\mu}_2 = (-1, 1, -1, 1, \ldots, -1, 1, 0, \ldots, 0)^\top, \; \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2.
\end{aligned}$$

Here, in (M2), the number of non-zero elements in $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ is $\lceil \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\}^{1/2}/2 \rceil$. The dimensions and sample sizes are chosen as follows:

$$p = 100, 250, 500, 1000;\ (n_1, n_2, n_3) = (20, 40, 60),\ (40, 80, 120),\ (60, 120, 180).$$

Then, we compare the true value e_1, the approximation (MI) and the approximation (MII) on these settings. By comparing the approximations in Table 5.1, it is seen that approximation (MI) is closer to the true value e_1 than (MII) is for all cases. In Table 5.2, (MI) and (MII) are close to true value e_1 when the sample size is relatively small, and (MI) is closer to the true value than (MII) when the sample size is relatively large. In situations where the dimension p is large and sample size n is small, (MII) is close to the true value e_1 and is a conservative approximation.

5.4.2 Accuracy of estimators

We investigate the MSE of the estimator $\widehat{e_1}$ on the same settings. For comparison, we consider the leave-one-out cross-validation method CV, which is a popular method for estimating prediction errors for small samples. The MSEs of the estimators CV and $\widehat{e_1}$ are given in Tables 5.3-5.6. These tables show that $\widehat{e_1}$ has smaller MSEs than CV does for all cases. Moreover, it can be confirmed that the MSE of our estimator is not influenced even if the distribution of $z_{i\ell t}$ is a t distribution.

Table 5.1: Comparison of approximations when (M1)

			(n_1, n_2, n_3)		
p			(20,40,60)	(40,80,120)	(60,120,180)
100	e_1	(D1)	0.0663	0.0552	0.0508
		(D2)	0.0662	0.0543	0.0515
	approx	(MI)	0.0683	0.0557	0.0516
		(MII)	0.0000	0.0000	0.0000
250	e_1	(D1)	0.0998	0.0702	0.0627
		(D2)	0.0984	0.0721	0.0627
	approx	(MI)	0.1007	0.0718	0.0623
		(MII)	0.0034	0.0000	0.0000
500	e_1	(D1)	0.1489	0.0979	0.0795
		(D2)	0.1469	0.0977	0.0796
	approx	(MI)	0.1499	0.0983	0.0799
		(MII)	0.0377	0.0032	0.0003
1000	e_1	(D1)	0.2217	0.1474	0.1134
		(D2)	0.2217	0.1464	0.1134
	approx	(MI)	0.2229	0.1472	0.1146
		(MII)	0.1414	0.0367	0.0104

Table 5.2: Comparison of approximations when (M2)

			(n_1, n_2, n_3)		
p			(20,40,60)	(40,80,120)	(60,120,180)
100	e_1	(D1)	0.3598	0.2854	0.2482
		(D2)	0.3573	0.2879	0.2459
	approx	(MI)	0.3642	0.2909	0.2500
		(MII)	0.3517	0.1855	0.1043
250	e_1	(D1)	0.3433	0.2629	0.2125
		(D2)	0.3460	0.2604	0.2150
	approx	(MI)	0.3485	0.2626	0.2135
		(MII)	0.3772	0.2093	0.1235
500	e_1	(D1)	0.3225	0.2313	0.1750
		(D2)	0.3232	0.2303	0.1753
	approx	(MI)	0.3258	0.2307	0.1770
		(MII)	0.3672	0.1998	0.1158
1000	e_1	(D1)	0.3183	0.2175	0.1613
		(D2)	0.3171	0.2169	0.1633
	approx	(MI)	0.3201	0.2190	0.1617
		(MII)	0.3773	0.2094	0.1236

Table 5.3: Comparison of MSEs when (M1) and (D1)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0016	0.0007	0.0004
	CV	0.0032	0.0013	0.0008
250	$\widehat{e}_1$	0.0022	0.0008	0.0005
	CV	0.0044	0.0017	0.0010
500	$\widehat{e}_1$	0.0030	0.0011	0.0006
	CV	0.0060	0.0022	0.0012
1000	$\widehat{e}_1$	0.0040	0.0015	0.0008
	CV	0.0078	0.0030	0.0016

Table 5.4: Comparison of MSEs when (M1) and (D2)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0016	0.0007	0.0004
	CV	0.0031	0.0013	0.0008
250	$\widehat{e}_1$	0.0022	0.0008	0.0005
	CV	0.0044	0.0016	0.0010
500	$\widehat{e}_1$	0.0030	0.0011	0.0006
	CV	0.0060	0.0021	0.0012
1000	$\widehat{e}_1$	0.0041	0.0015	0.0008
	CV	0.0078	0.0030	0.0016

Table 5.5: Comparison of MSEs when (M2) and (D1)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0042	0.0022	0.0014
	CV	0.0091	0.0041	0.0026
250	$\widehat{e}_1$	0.0042	0.0021	0.0012
	CV	0.0090	0.0038	0.0023
500	$\widehat{e}_1$	0.0041	0.0019	0.0011
	CV	0.0087	0.0036	0.0020
1000	$\widehat{e}_1$	0.0042	0.0017	0.0010
	CV	0.0088	0.0035	0.0019

Table 5.6: Comparison of MSEs when (M2) and (D2)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0042	0.0022	0.0014
	CV	0.0091	0.0041	0.0026
250	$\widehat{e}_1$	0.0042	0.0021	0.0013
	CV	0.0090	0.0039	0.0023
500	$\widehat{e}_1$	0.0041	0.0019	0.0011
	CV	0.0088	0.0036	0.0020
1000	$\widehat{e}_1$	0.0042	0.0017	0.0010
	CV	0.0088	0.0035	0.0019

5.5 Conclusion

We showed the asymptotic multivariate normality for several Euclidean distance-based discriminant functions under high-dimensional settings. Our theoretical results have been established under variance heterogeneity and nonnormality. Further, using asymptotic multivariate normality, we proposed a new estimator of misclassification probability of Euclidean distance-based discriminant rule. We confirmed that proposed estimators have good performances in high-dimensional situations through numerical simulations.

Chapter 6

Simultaneous confidence interval for paired mean vectors

In Chapter 4-5, we considered the Euclidean distance-based classifier in high-dimensional data. The accuracy of its classifier depends on the Euclidean distance between the mean vectors of the populations. As a research related to Euclidean distance in recent years, Bai and Saranadasa [6] considered the two-sample test with equal covariance matrices, and proposed an estimator based on the squared Euclidean norm instead of T^2 statistics. Nishiyama et al. [29] proposed a test procedure for linear hypotheses of a set of mean vectors from $k \geq 2$ normal populations. Chen and Qin [10] also derived a test for the two-sample problem without assuming normality. Besides, various alternative approaches for the two-sample test have been proposed; see, e.g., [8, 16, 32, 40, 41]. In addition, k-sample significance tests for high-dimensional mean vectors have been studied in [21, 45].

Previous studies focused on testing whether the gth population mean vector $\boldsymbol{\mu}_g$ and the hth population mean vector $\boldsymbol{\mu}_h$ are the same or not. However, interval estimation for $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$ has not been widely studied. In discriminant analysis based on Euclidean distance, it is important in terms of discrimination accuracy to investigate Euclidean distance between the mean vectors. Accordingly, in this Chapter, we consider simultaneous confidence interval for paired mean vectors in high-dimensional data. To derive it, we show the asymptotic distribution of the quadratic form in a set of sample mean vectors, and get an approximate confidence interval for $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$.

6.1 Statistical model

To give an approximate interval estimator, asymptotic normality is established under some conditions. For $g \in \{1, 2, \ldots, k\}$, let $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ be the mean vector and covariance matrix, respectively, of the gth population. Let $\boldsymbol{\mu} = \mathrm{vec}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k)$ be the unknown $pk \times 1$ vector of $k \geq 2$ mean vectors corresponding to the k group. Additionally, let $\mathbf{A}$ be a known $p \times p$ symmetric semi-positive definite matrix, and let $\mathbf{V} = (v_{ij})$ be a known $k \times k$ symmetric semi-positive definite matrix. Let $\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gn_g}$ be random observations

from the gth population, and let

$$\overline{\mathbf{x}} = (\overline{\mathbf{x}}_1^\top, \overline{\mathbf{x}}_2^\top, \ldots, \overline{\mathbf{x}}_k^\top)^\top, \quad \overline{\mathbf{x}}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{x}_{gi}, \quad \mathbf{S}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)(\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)^\top.$$

We consider the random variable

$$T = \overline{\mathbf{x}}^\top (\mathbf{V} \otimes \mathbf{A})\overline{\mathbf{x}} - \sum_{g=1}^{k} \frac{v_{gg}}{n_g} \mathrm{tr}(\mathbf{A}\mathbf{S}_g),$$

where "$\mathbf{A} \otimes \mathbf{B}$" denotes Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$.

Remark 6.1.1. The random variable T can be adapted to various hypothesis testing problems, as follows.

(a) If we set $k = 2$, $v_{11} = v_{22} = 1, v_{12} = v_{21} = -1$ and $\mathbf{A} = \mathbf{I}$, then T can be used to test $\mathcal{H} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $\mathcal{A} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. This statistic is used in Chen and Qin [10].

(b) If we set $v_{gg} = n_g(1 - n_g/n)$ for $g \in \{1, 2, \ldots, k\}$, $v_{gh} = -n_g n_h/n$ for $g \neq h, g, h \in \{1, 2, \ldots, k\}$, and $\mathbf{A} = \mathbf{I}$, then T can be used to test $\mathcal{H} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ vs. $\mathcal{A} : \neg\mathcal{H}$. Here, $n = n_1 + n_2 + \cdots + n_k$. This statistic is used by Yamada and Himeno [45].

(c) If we set $\mathbf{V} = (\beta_i\beta_j)$ and $\mathbf{A} = \mathbf{I}$, then T is used for linear hypotheses of a set of mean vectors $\mathcal{H} : \sum_{g=1}^{k} \beta_g \boldsymbol{\mu}_g = \mathbf{0}$ vs. $\mathcal{A} : \sum_{g=1}^{k} \beta_g \boldsymbol{\mu}_g \neq \mathbf{0}$, where $\beta_1, \beta_2, \ldots, \beta_k$ are some real constants.

Let $n_1, n_2, \ldots, n_k$, $\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}$ and $\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)$ for $g, h \in \{1, 2, \ldots, k\}$ be functions of p. We derive the asymptotic distribution of T under the assumptions listed below.

(E1) The random vector $\mathbf{x}_{gi}$ follows the model defined, for all $i \in \{1, 2, \ldots, n_g\}$ and $g \in \{1, 2, \ldots, k\}$, by

$$\mathbf{x}_{gi} = R_{gi}\boldsymbol{\Sigma}_g^{1/2}\mathbf{z}_{gi} + \boldsymbol{\mu}_g, \tag{6.1}$$

where R_{gi} is a non-negative random variable, and $\mathbf{z}_{gi}$ is a p-dimensional random vector. In addition, R_{gi} and $\mathbf{z}_{gi}$ satisfy the following conditions:

(i) For all $g \in \{1, 2, \ldots, k\}$, $R_{g1}, R_{g2}, \ldots, R_{gn_g}$ are identically distributed.

(ii) For all $g \in \{1, 2, \ldots, k\}$, $\mathbf{z}_{g1}, \mathbf{z}_{g2}, \ldots, \mathbf{z}_{gn_g}$ are identically distributed.

(iii) $R_{11}, R_{12}, \ldots, R_{1n_1}, \ldots, R_{k1}, R_{k2}, \ldots, R_{kn_k}$, $\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{1n_1}, \ldots, \mathbf{z}_{k1}, \mathbf{z}_{k2}, \ldots, \mathbf{z}_{kn_k}$ are mutually independent.

(iv) The jth element of $\mathbf{z}_{gi}$ (denoted as z_{gij}) has $\mathbf{E}(z_{gij}) = 0$, $\mathbf{E}(z_{gij}^2) = 1$, and $\mathbf{E}(z_{gij}^4)$ is uniformly bounded with respect to p.

(v) $\mathbf{E}(R_{gi}^2) = 1$, and $\mathbf{E}(R_{gi}^4)$ is uniformly bounded with respect to p.

(vi) For $r \in \{2,3,4\}$, define

$$M_g(j_1, j_2, \ldots, j_r \ ; \ \alpha_1, \alpha_2, \ldots, \alpha_r) = \mathrm{E}\left(\prod_{\ell=1}^{r} z_{gij_\ell}^{\alpha_\ell}\right)$$

where $\alpha_\ell \in \{1,2,3\}$, $\alpha_1+\alpha_2+\cdots+\alpha_r \leq 4$ whenever $j_1, j_2, \ldots, j_r$ are distinct indices. Then $M_g(j_1, j_2, \ldots, j_r \ ; \ \alpha_1, \alpha_2, \ldots, \alpha_r) = 0$ except $M_g(\cdot, \cdot \ ; \ 2,2)$. Also, $M_g(\cdot, \cdot \ ; \ 2,2)$ is uniformly bounded with respect to p.

(E2) $\min(p, n_1, n_2, \ldots, n_k) \to \infty$ as $p \to \infty$.

(E3) For all $g, h \in \{1, 2, \ldots, k\}$, $\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}/\{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)\}^2 = o(1)$ as $p \to \infty$.

(E3') For all $g \in \{1, 2, \ldots, k\}$, $\mathrm{tr}(\boldsymbol{\Sigma}_g^4)/\{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\}^2 = o(1)$ as $p \to \infty$.

From the following remark, it can be understood that the moment condition (E1) includes some of the special cases used in previous studies. We also note that condition (E1) covers the class of elliptical distributions.

Remark 6.1.2. The moment condition (E1) includes the following special cases.

(a) If we set $R_{gi} = 1$ a.s. and $M_g(\cdot, \cdot \ ; \ 2,2) = 1$ in (E1), we get the moment condition used by Chen and Qin [10].

(b) Let $\tilde{R}_{gi}$ be a non-negative random variable and let $\mathbf{u}_{gi}$ be a random vector distributed uniformly on the surface of a unit sphere in $\mathbb{R}^p$. Here, $\tilde{R}_{gi}$ and $\mathbf{u}_{gi}$ are independent. Then $\mathbf{x}_{gi} = \boldsymbol{\mu}_g + \tilde{R}_{gi}\boldsymbol{\Sigma}_g^{1/2}\mathbf{u}_{gi}$ has an elliptical distribution. We set $\mathbf{z}_{gi} = \sqrt{p}\mathbf{u}_{gi}$ and $R_{gi} = \tilde{R}_{gi}/\{\mathrm{E}(\tilde{R}_{gi}^2)\}^{1/2}$ in (E1). Then

$$\begin{aligned}
&\mathbf{E}(z_{gij}) = 0, \quad \mathbf{E}(z_{gij}^2) = 1, \\
&\mathbf{E}(z_{gij_1}z_{gij_2}z_{gij_3}z_{gij_4}) = \frac{p^2}{p(p+2)}([j_1 = j_2][j_3 = j_4] + [j_1 = j_3][j_2 = j_4] \\
&\qquad\qquad + [j_1 = j_4][j_2 = j_3]) \\
&\qquad < [j_1 = j_2][j_3 = j_4] + [j_1 = j_3][j_2 = j_4] + [j_1 = j_4][j_2 = j_3].
\end{aligned}$$

Here, $[\cdot]$ denotes the Iverson bracket. If $\mathbf{E}(\tilde{R}_{gi}^4)/\{\mathbf{E}(\tilde{R}_{gi}^2)\}^2 = O(1)$, then (E1) holds. Thus (E1) includes elliptical distributions. In the case of a multivariate normal distribution, $\mathbf{E}(\tilde{R}_{gi}^2) = p$, $\mathbf{E}(\tilde{R}_{gi}^4) = p(p+2)$.

The asymptotic frameworks $p \asymp n_g$, $\ln(p) = o(n_g^{1/3})$, and $n_g = O(p^\eta), \eta > 1/2$ are used by Bai and Saranadasa [6], Cai et al. [8], and Srivastava et al. [40], respectively. However, under these assumptions, the relationship between p and n_g is restricted. Condition (E2) is more flexible than these assumptions. Condition (E3) is used as an assumption to derive general results, and condition (E3') is used for interval estimation. In the following remarks, we introduce the some covariance matrices satisfying condition (E3').

Remark 6.1.3. The covariance matrices in (a), (b), and (c) satisfy condition (E3').

(a) The covariance matrix $\mathbf{\Sigma}_g$ is a symmetric matrix whose non-zero elements are arranged uniformly near the diagonal such that $\sigma_{g,ij} = 0$ if $|i-j| > q$. Then $\mathrm{tr}(\mathbf{\Sigma}_g^2) = O(pq)$ and $\mathrm{tr}(\mathbf{\Sigma}_g^4) = O(pq^3)$. If the $\sigma_{g,ij}^2$s are uniformly bounded away from infinity and zero and the half-band width is $q = O(p^\gamma)$ for some $\gamma \in [0, 1)$, then (E3') holds.

(b) The covariance matrix $\mathbf{\Sigma}_g = (\sigma_{g,i}\sigma_{g,j}\rho_g^{|i-j|})$ where $\sigma_{g,\ell}^2 = \mathrm{var}(x_{gi\ell})$ is the marginal variance for $\ell \in \{1, 2, \ldots, p\}$. If the $\sigma_{g,\ell}^2$s are uniformly bounded away from infinity and zero, (E3') is satisfied.

(c) Let $\lambda_{g,i} = O(p^\eta)$ for all $i \in \{1, 2, \ldots, \iota\}$ and $\lambda_{g,i} = O(1)$ for all $i \in \{\iota+1, \iota+2, \ldots, p\}$ be the eigenvalues of $\mathbf{\Sigma}_g$. Here, ι is an unknown and positive fixed integer. If $\eta < 1/2$, then (E3') is satisfied.

In any case, it is desirable that the covariance matrix $\mathbf{\Sigma}_g$ should be sparse. If it is not appropriate to assume sparsity in the covariance matrix $\mathbf{\Sigma}_g$, then please check two-sample tests of Aoshima and Yata [4].

Under (E1), the expectation and variance of T are obtained as follows:

$$\begin{aligned}
&\mathbf{E}(T) = \boldsymbol{\mu}^\top(\mathbf{V}\otimes\mathbf{A})\boldsymbol{\mu},\\
&\sigma^2 = \mathrm{var}(T) = \sum_{g=1}^{k}\frac{2v_{gg}^2\mathrm{tr}\{(\mathbf{A}\mathbf{\Sigma}_g)^2\}}{n_g(n_g-1)} + \sum_{g=2}^{k}\sum_{h=1}^{g-1}\frac{4v_{gh}^2\mathrm{tr}(\mathbf{A}\mathbf{\Sigma}_g\mathbf{A}\mathbf{\Sigma}_h)}{n_g n_h}\\
&\qquad\qquad + 4\boldsymbol{\mu}^\top(\mathbf{V}\otimes\mathbf{A})\left\{\sum_{g=1}^{k}\frac{1}{n_g}(\boldsymbol{e}_g\boldsymbol{e}_g^\top)\otimes\mathbf{\Sigma}_g\right\}(\mathbf{V}\otimes\mathbf{A})\boldsymbol{\mu}.
\end{aligned}$$

We state below that the asymptotic normality of the Chen-Qin-type test statistic T holds under (E1)–(E3) in a form that is weaker than the conditions considered previously.

Theorem 6.1.1 (Hyodo, Watanabe, and Seo [22])**.** *Under Assumptions (E1)–(E3),* $\{T - \boldsymbol{\mu}^\top(\mathbf{V}\otimes\mathbf{A})\boldsymbol{\mu}\}/\sigma \rightsquigarrow \mathcal{N}(0,1)$ *as* $p \to \infty$.

Proof. For all $i \in \{1, 2, \ldots, n_g\}$ and $g \in \{1, 2, \ldots, k\}$, let $\mathbf{y}_{gi} = \mathbf{x}_{gi} - \boldsymbol{\mu}_g$, $\overline{\mathbf{y}}_g = \sum_{i=1}^{n_g}\mathbf{y}_{gi}/n_g$, $n(0) = 0$, and $n(g) = n_1 + n_2 + \cdots + n_g$. For each $i \in \{n(g-1)+1, n(g-1)+2, \ldots, n(g)\}$, define

$$\varepsilon_i = \frac{2}{\sigma n_g(n_g-1)}\,\mathbf{y}_{gi-n(g-1)}^\top\mathbf{A}\boldsymbol{a}_{gi-n(g-1)},$$

where

$$\boldsymbol{a}_{gi} = v_{gg}\sum_{j=1}^{i-1}\mathbf{y}_{gj} + \mathbf{1}(g-1)(n_g-1)\sum_{h=1}^{g-1}v_{gh}\overline{\mathbf{y}}_h + (n_g-1)\sum_{h=1}^{k}v_{gh}\boldsymbol{\mu}_h.$$

Here, $\mathbf{1}(x)$ denotes the indicator function with $\mathbf{1}(x) = 0$ if $x \le 0$ and $\mathbf{1}(x) = 1$ if $x > 0$. Then

$$\frac{T - \boldsymbol{\mu}^\top(\mathbf{V}\otimes\mathbf{A})\boldsymbol{\mu}}{\sigma} = \sum_{i=1}^{n(k)}\varepsilon_i.$$

Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and, for all $i \in \{n(g-1)+1, n(g-1)+2, \ldots, n(g)\}$, $\mathcal{F}_i = \sigma\{\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1}, \ldots, \mathbf{y}_{g1}, \mathbf{y}_{g2}, \ldots, \mathbf{y}_{gi-n(g-1)}\}$. Then $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty$ and $\mathrm{E}(\varepsilon_i|\mathcal{F}_{i-1}) = 0$. We show the asymptotic normality of $\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{n(k)}$ by adapting the martingale-difference Central Limit Theorem; see, e.g., Hall and Heyde [17]. To apply this theorem, it is necessary to check the following two conditions under (E1)–(E3):

$$\text{(I)}: \sum_{i=1}^{n(k)} \mathrm{E}(\varepsilon_i^2|\mathcal{F}_{i-1}) = 1 + o_p(1), \quad \text{(II)}: \sum_{i=1}^{n(k)} \mathrm{E}(\varepsilon_i^4) = o(1).$$

To check (I), we first write the sum of the conditional expectations as $B_1 + \cdots + B_7$, where

$$\begin{aligned}
B_1 &= \sum_{g=1}^{k} \frac{4v_{gg}^2}{\sigma^2 n_g^2 (n_g-1)^2} \sum_{i=1}^{n_g} (n_g - i)[\mathbf{y}_{gi}^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\mathbf{y}_{gi} - \mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^2\}], \\
B_2 &= \sum_{g=1}^{k} \frac{8v_{gg}^2}{\sigma^2 n_g^2 (n_g-1)^2} \sum_{i=2}^{n_g} \sum_{j=1}^{i-1} (n_g - i)\mathbf{y}_{gi}^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\mathbf{y}_{gj}, \\
B_3 &= \sum_{g=2}^{k} \sum_{h=1}^{g-1} \frac{8v_{gg}v_{gh}}{\sigma^2 n_g^2 (n_g-1)} \sum_{i=1}^{n_g} (n_g - i)\mathbf{y}_{gi}^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\overline{\mathbf{y}}_h, \\
B_4 &= \sum_{g=1}^{k} \frac{8v_{gg}}{\sigma^2 n_g^2 (n_g-1)} \sum_{i=1}^{n_g} (n_g - i)\mathbf{y}_{gi}^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A} \left(\sum_{h=1}^{k} v_{gh}\boldsymbol{\mu}_h \right), \\
B_5 &= \sum_{g=2}^{k} \sum_{h=1}^{g-1} \frac{4v_{gh}^2}{\sigma^2 n_g} \left\{ \overline{\mathbf{y}}_h^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\overline{\mathbf{y}}_h - \frac{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\boldsymbol{\Sigma}_h)}{n_h} \right\}, \\
B_6 &= \sum_{g=3}^{k} \sum_{h=2}^{g-1} \sum_{\ell=1}^{h-1} \frac{8v_{gh}v_{g\ell}}{\sigma^2 n_g} \overline{\mathbf{y}}_h^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A}\overline{\mathbf{y}}_\ell, \\
B_7 &= \sum_{g=2}^{k} \sum_{h=1}^{g-1} \frac{8v_{gh}}{\sigma^2 n_g} \overline{\mathbf{y}}_h^\top \mathbf{A}\boldsymbol{\Sigma}_g \mathbf{A} \left(\sum_{h=1}^{k} v_{gh}\boldsymbol{\mu}_h \right).
\end{aligned}$$

Using Hölder's inequality, we find

$$\mathrm{E}\left[\left\{ \sum_{i=1}^{n(k)} \mathrm{E}(\varepsilon_i^2|\mathcal{F}_{i-1}) - 1 \right\}^2 \right] \leq 7 \sum_{i=1}^{7} \mathbf{E}(B_i^2).$$

The proof of (I) is complete upon noting that under (E1)–(E3), we have

$$\mathbf{E}(B_1^2) = O\left(\sum_{g=1}^{k} n_g^{-1}\right), \quad \mathbf{E}(B_2^2) = O\left[\sum_{g=1}^{k} \frac{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^4\}}{[\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^2\}]^2}\right],$$

$$\mathbf{E}(B_3^2) = O\left[\sum_{g=2}^{k}\sum_{h=1}^{g-1} \frac{\sqrt{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^4\}\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}}}{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^2\}\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)}\right],$$

$$\mathbf{E}(B_4^2) = O\left[\sum_{g=1}^{k} \frac{\sqrt{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^4\}}}{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g)^2\}}\right],$$

$$\mathbf{E}(B_5^2) = O\left[\sum_{g=2}^{k}\sum_{h=1}^{g-1} \frac{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}}{\{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)\}^2}\right],$$

$$\mathbf{E}(B_6^2) = O\left[\sum_{g=3}^{k}\sum_{h=2}^{g-1}\sum_{\ell=1}^{h-1} \frac{\sqrt{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_\ell)^2\}}}{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_\ell)}\right],$$

$$\mathbf{E}(B_7^2) = O\left[\sum_{g=2}^{k}\sum_{h=1}^{g-1} \frac{\sqrt{\mathrm{tr}\{(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)^2\}}}{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_g\mathbf{A}\boldsymbol{\Sigma}_h)}\right].$$

To check (II), define

$$\varepsilon_i^{(1)} = \frac{2v_{gg}\mathbf{y}_{gi-n(g-1)}^\top \mathbf{A}\sum_{j=1}^{i-n(g-1)-1}\mathbf{y}_{gj}}{\sigma n_g(n_g-1)},\ \varepsilon_i^{(2)} = \frac{2\mathbf{1}(g-1)\mathbf{y}_{gi-n(g-1)}^\top \mathbf{A}\sum_{h=1}^{g-1} v_{gh}\overline{\mathbf{y}}_h}{\sigma n_g},$$

$$\varepsilon_i^{(3)} = \frac{2\mathbf{y}_{gi-n(g-1)}^\top \mathbf{A}\sum_{h=1}^{k} v_{gh}\boldsymbol{\mu}_h}{\sigma n_g}.$$

Using Hölder's inequality, we find

$$\sum_{i=1}^{n(k)} \mathrm{E}(\varepsilon_i^4) = \sum_{g=1}^{k}\sum_{i=n(g-1)+1}^{n(g)} \mathbf{E}\left\{\left(\sum_{j=1}^{3}\varepsilon_i^{(j)}\right)^4\right\} \le \sum_{g=1}^{k}\sum_{i=n(g-1)+1}^{n(g)} \mathbf{E}\left(3^3\sum_{j=1}^{3}\varepsilon_i^{(j)^4}\right)$$
$$= 3^3\sum_{g=1}^{k}\sum_{j=1}^{3}\sum_{i=n(g-1)+1}^{n(g)} \mathbf{E}\left(\varepsilon_i^{(j)^4}\right).$$

Under (E1)–(E3), $\sum_{i=n(g-1)+1}^{n(g)} \mathbf{E}(\varepsilon_i^{(j)^4}) = O(n_g^{-1})$ for $j \in \{1,2,3\}$. Thus, the proof of (II) is complete. □

We obtain the asymptotic result for interval estimation as a special case of Theorem 6.1.1. In fact, if $\mathbf{V} = (\mathbf{e}_g - \mathbf{e}_h)(\mathbf{e}_g - \mathbf{e}_h)^\top$ and $\mathbf{A} = \mathbf{I}$ in Theorem 6.1.1, we obtain the following corollary.

Corollary 6.1.1 (Hyodo, Watanabe, and Seo [22]). *Define* $\delta_{gh} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$, $\widehat{\delta}_{gh} = \|\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_h\|^2 - \mathrm{tr}(\mathbf{S}_g)/n_g - \mathrm{tr}(\mathbf{S}_h)/n_h$, *and*

$$\sigma_{gh} = 2\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{2n_g(n_g-1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_h^2)}{2n_h(n_h-1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}{n_g n_h} + \frac{\Delta_{gh}}{n_g} + \frac{\Delta_{hg}}{n_h}},$$

where $\Delta_{gh} = (\boldsymbol{\mu}_g - \boldsymbol{\mu}_h)^\top \boldsymbol{\Sigma}_g (\boldsymbol{\mu}_g - \boldsymbol{\mu}_h)$. *Then, under (E1), (E2) and (E3'),* $(\widehat{\delta}_{gh} - \delta_{gh})/\sigma_{gh} \rightsquigarrow \mathcal{N}(0,1)$ *as* $p \to \infty$.

Remark 6.1.4. Assuming that $\delta_{gh} = 0$ in Corollary 6.1.1 (i.e., $\boldsymbol{\mu}_g = \boldsymbol{\mu}_h$), we obtain the same result as Chen and Qin [10]. They also showed the asymptotic normality of $\widehat{\delta}_{gh}$ under $\Delta_{gh} = o[\mathrm{tr}\{(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_h)^2\}/(n_g + n_h)]$. Not only does our result not require this assumption but it also relaxes the assumptions about the moments of $\mathbf{z}_{gi}$ in (6.1). If we assume $\Delta_{gh} = o[\mathrm{tr}\{(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_h)^2\}/(n_g + n_h)]$ as reported by Chen and Qin [10], the term Δ_{gh} in σ_{gh} can be ignored. However, this assumption implies a local alternative hypothesis for testing $\mathcal{H} : \boldsymbol{\mu}_g = \boldsymbol{\mu}_h$ vs. $\mathcal{A} : \boldsymbol{\mu}_g \neq \boldsymbol{\mu}_h$ in their paper, and it is not appropriate to use it in interval estimation. Therefore, in our results, no assumptions are made regarding $\boldsymbol{\mu}_g$ and $\boldsymbol{\mu}_h$.

6.2 Confidence interval for $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$

In this section, we construct an approximate confidence interval for δ_{gh} based on the asymptotic results in Theorem 6.1.1. To derive a confidence interval for δ_{gh}, it is necessary to prepare some estimators. We introduce the unbiased estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_g^2)$ as follows:

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)} = \mathrm{tr}(\mathbf{S}_g \mathbf{S}_h),$$
$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)} = \frac{n_g - 1}{n_g(n_g-2)(n_g-3)} \left[(n_g-1)(n_g-2)\mathrm{tr}(\mathbf{S}_g^2) + \{\mathrm{tr}(\mathbf{S}_g)\}^2 - n_g K_g\right],$$

where

$$K_g = \frac{1}{n_g - 1} \sum_{j=1}^{n_g} \|\mathbf{x}_{gj} - \overline{\mathbf{x}}_g\|^4.$$

These estimates are used for testing $\mathcal{H} : \boldsymbol{\mu}_g = \boldsymbol{\mu}_h$ vs $\mathcal{A} : \boldsymbol{\mu}_g \neq \boldsymbol{\mu}_h$. The unbiased estimator $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}$ was proposed by Himeno and Yamada [20]. In addition to these estimators, it is necessary to have an estimator of Δ_{gh} for the interval estimation of $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$. To this end, we propose the following unbiased estimator of Δ_{gh}:

$$\widehat{\Delta}_{gh} = \frac{(n_g-2)V_{gh} - 2U_{gh}}{(n_g-1)(n_g-2)} - \frac{\mathrm{tr}(\mathbf{S}_g \mathbf{S}_h)}{n_h} + \frac{2n_g K_g - (n_g-1)\{\mathrm{tr}(\mathbf{S}_g)\}^2 - (n_g-1)^2 \mathrm{tr}(\mathbf{S}_g^2)}{n_g(n_g-2)(n_g-3)},$$

where

$$V_{gh} = \sum_{j=1}^{n_g} \{(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_h)^\top (\mathbf{x}_{gj} - \overline{\mathbf{x}}_g)\}^2, \; U_{gh} = \sum_{j=1}^{n_g} (\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_h)^\top (\mathbf{x}_{gj} - \overline{\mathbf{x}}_g)(\mathbf{x}_{gj} - \overline{\mathbf{x}}_g)^\top (\mathbf{x}_{gj} - \overline{\mathbf{x}}_g).$$

We also investigate the order in probability of these estimators under (E1), (E2), and (E3').

Theorem 6.2.1 (Hyodo, Watanabe, and Seo [22])**.** *Under (E1) and (E2), we have* $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)} = \mathrm{tr}(\boldsymbol{\Sigma}_g^2) + o_p(n_g^2 \sigma_{gh}^2)$ *and* $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)} = \mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h) + o_p(n_g n_h \sigma_{gh}^2)$ *as* $p \to \infty$. *If (E3') also holds, then* $\widehat{\Delta}_{gh} = \Delta_{gh} + o_p(n_g \sigma_{gh}^2)$ *as* $p \to \infty$.

Proof. We rewrite $\widehat{\mathrm{tr}(\mathbf{\Sigma}_g^2)} = C_1 + C_2 + C_3$, where

$$
\begin{aligned}
C_1 &= \frac{1}{n_g(n_g-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{n_g} (\mathbf{y}_{gi}^\top \mathbf{y}_{gj})^2, \\
C_2 &= -\frac{2}{n_g(n_g-1)(n_g-2)} \sum_{\substack{i,j,k=1 \\ i\neq j, j\neq k, k\neq i}}^{n_g} \mathbf{y}_{gi}^\top \mathbf{y}_{gj} \mathbf{y}_{gi}^\top \mathbf{y}_{gk}, \\
C_3 &= \frac{1}{n_g(n_g-1)(n_g-2)(n_g-3)} \sum_{\substack{i,j,k,\ell=1 \\ i\neq j\neq k\neq \ell, k\neq i\neq \ell\neq j}}^{n_g} \mathbf{y}_{gi}^\top \mathbf{y}_{gj} \mathbf{y}_{gk}^\top \mathbf{y}_{g\ell}.
\end{aligned}
$$

Under (E1) and (E2), the variances of C_i for $i = \{1, 2, 3\}$ are evaluated as follows:

$$
\begin{aligned}
\mathrm{var}(C_1) &= O\left[\frac{\{\mathrm{tr}(\mathbf{\Sigma}_g^2)\}^2}{n_g}\right] = O(n_g^3 \sigma_{gh}^4) = o(n_g^4 \sigma_{gh}^4), \\
\mathrm{var}(C_2) &= O\left[\frac{\{\mathrm{tr}(\mathbf{\Sigma}_g^2)\}^2}{n_g^2}\right] = O(n_g^2 \sigma_{gh}^4) = o(n_g^4 \sigma_{gh}^4), \\
\mathrm{var}(C_3) &= O\left[\frac{\{\mathrm{tr}(\mathbf{\Sigma}_g^2)\}^2}{n_g^4}\right] = O(\sigma_{gh}^4) = o(n_g^4 \sigma_{gh}^4).
\end{aligned}
$$

Thus we get $\widehat{\mathrm{tr}(\mathbf{\Sigma}_g^2)} = \mathrm{tr}(\mathbf{\Sigma}_g^2) + o_p(n_g^2 \sigma_{gh}^2)$ under (E1) and (E2).

Furthermore, we rewrite $\widehat{\mathrm{tr}(\mathbf{\Sigma}_g \mathbf{\Sigma}_h)} = D_1 + D_2 + D_3 + D_4$, where

$$
\begin{aligned}
D_1 &= \frac{1}{n_g n_h} \sum_{j=1}^{n_g} \sum_{k=1}^{n_h} (\mathbf{y}_{gj}^\top \mathbf{y}_{hk})^2, \quad D_2 = -\frac{1}{n_g n_h (n_h-1)} \sum_{j=1}^{n_g} \sum_{\substack{k,\ell=1 \\ k\neq \ell}}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{hk} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell}, \\
D_3 &= -\frac{1}{n_g n_h (n_g-1)} \sum_{j=1}^{n_h} \sum_{\substack{k,\ell=1 \\ k\neq \ell}}^{n_g} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell} \mathbf{y}_{gk}^\top \mathbf{y}_{h\ell}, \\
D_4 &= \frac{1}{n_g(n_g-1) n_h(n_h-1)} \sum_{\substack{j,k=1 \\ j\neq k}}^{n_g} \sum_{\substack{\ell,m=1 \\ \ell\neq m}}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell} \mathbf{y}_{gk}^\top \mathbf{y}_{hm}.
\end{aligned}
$$

Under (E1) and (E2), we find

$$\begin{aligned}
\mathrm{var}(D_1) &= O\left[\frac{n_g+n_h}{n_g n_h}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2\right] = o(n_g^2 n_h^2 \sigma_{gh}^4),\\
\mathrm{var}(D_2) &= O\left[\frac{1}{n_h^2}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2\right] = o(n_g^2 n_h^2 \sigma_{gh}^4),\\
\mathrm{var}(D_3) &= O\left[\frac{1}{n_g^2}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2\right] = o(n_g^2 n_h^2 \sigma_{gh}^4),\\
\mathrm{var}(D_4) &= O\left[\frac{1}{n_g^2 n_h^2}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2\right] = o(n_g^2 n_h^2 \sigma_{gh}^4).
\end{aligned}$$

Thus we get $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h}) = \mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h) + o_p(n_g n_h \sigma_{gh}^2)$ under (E1) and (E2).

Finally, we rewrite $\widehat{\Delta}_{gh} = E_1 + E_2 + \cdots + E_{12}$, where

$$\begin{aligned}
E_1 &= \frac{1}{n_g(n_g-1)(n_g-2)} \sum_{\substack{j,k,\ell=1\\ j\neq k, k\neq \ell, \ell\neq j}}^{n_g} \mathbf{y}_{gj}^\top \mathbf{y}_{gk} \mathbf{y}_{gj}^\top \mathbf{y}_{g\ell},\\
E_2 &= -\frac{1}{n_g(n_g-1)(n_g-2)(n_g-3)} \sum_{\substack{j,k,\ell,m=1\\ j\neq k\neq \ell\neq m\\ \ell\neq j\neq m\neq k}}^{n_g} \mathbf{y}_{gj}^\top \mathbf{y}_{gk} \mathbf{y}_{g\ell}^\top \mathbf{y}_{gm},\\
E_3 &= -\frac{2}{n_g n_h(n_g-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n_g} \sum_{\ell=1}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{gk} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell},\\
E_4 &= \frac{2}{n_g n_h(n_g-1)(n_g-2)} \sum_{\substack{j,k,\ell=1\\ j\neq k, k\neq \ell, \ell\neq j}}^{n_g} \sum_{m=1}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{gk} \mathbf{y}_{g\ell}^\top \mathbf{y}_{hm},\\
E_5 &= \frac{1}{n_g n_h(n_h-1)} \sum_{j=1}^{n_g} \sum_{\substack{k,\ell=1\\ k\neq \ell}}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{hj} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell},\\
E_6 &= -\frac{2}{n_g(n_g-1)n_h(n_h-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n_g} \sum_{\substack{\ell,m=1\\ \ell\neq m}}^{n_h} \mathbf{y}_{gj}^\top \mathbf{y}_{h\ell} \mathbf{y}_{gk}^\top \mathbf{y}_{hm},\\
E_7 &= \frac{2}{n_g(n_g-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n_g} \boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj} \mathbf{y}_{gj}^\top \mathbf{y}_{gk},\\
E_8 &= -\frac{2}{n_g(n_g-1)(n_g-2)} \sum_{\substack{j,k,\ell=1\\ j\neq k, k\neq \ell, \ell\neq j}}^{n_g} \boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj} \mathbf{y}_{gk}^\top \mathbf{y}_{g\ell},
\end{aligned}$$

$$E_9 = -\frac{2}{n_g n_h}\sum_{j=1}^{n_g}\sum_{k=1}^{n_h}\boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj}\mathbf{y}_{gj}^\top \mathbf{y}_{hj}, \quad E_{10} = \frac{2}{n_g n_h(n_g-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{n_g}\sum_{\ell=1}^{n_h}\boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj}\mathbf{y}_{gk}^\top \mathbf{y}_{h\ell},$$

$$E_{11} = \frac{1}{n_g}\sum_{j=1}^{n_g}(\boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj}\mathbf{y}_{gj}^\top \boldsymbol{\delta}_{gh} - \boldsymbol{\delta}_{gh}^\top \Sigma_g \boldsymbol{\delta}_{gh}), \quad E_{12} = -\frac{1}{n_g(n_g-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{n_g}\boldsymbol{\delta}_{gh}^\top \mathbf{y}_{gj}\mathbf{y}_{gk}^\top \boldsymbol{\delta}_{gh}.$$

Under (E1), (E2) and (E3'), we find

$$\mathrm{var}(E_1) = O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\}^2}{n_g^3} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)}{n_g^2}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_2) = O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\}^2}{n_g^4}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_3) = O\left[\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}{n_g^2 n_h} + \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)\}^{1/2}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)^2\}^{1/2}}{n_g n_h}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_4) = O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}{n_g^3 n_h}\right\} = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_5) = O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2}{n_g n_h^2} + \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)^2\}}{n_h^2}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_6) = O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)\}^2}{n_g^2 n_h^2}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_7) = O\left[\frac{\Delta_{gh}\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{n_g^2} + \frac{\Delta_{gh}\{\mathrm{tr}(\boldsymbol{\Sigma}_g^4)\}^{1/2}}{n_g}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_8) = O\left\{\frac{\Delta_{gh}\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{n_g^3}\right\} = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_9) = O\left[\frac{\Delta_{gh}\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}{n_g n_h} + \frac{\Delta_{gh}\{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)^2\}^{1/2}}{n_h}\right] = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_{10}) = O\left\{\frac{\Delta_{gh}\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}{n_g^2 n_h}\right\} = o(n_g^2\sigma_{gh}^4), \quad \mathrm{var}(E_{11}) = O\left(\frac{\Delta_{gh}^2}{n_g}\right) = o(n_g^2\sigma_{gh}^4),$$

$$\mathrm{var}(E_{12}) = O\left(\frac{\Delta_{gh}^2}{n_g^2}\right) = o(n_g^2\sigma_{gh}^4).$$

With the above result, the proof is complete. □

First, we estimate the asymptotic standard deviation σ_{gh}. Before deriving the estimator of σ_{gh}, we describe the risk of using the estimator cq_{gh} obtained by Chen and Qin [10] to estimate σ_{gh} directly. The estimator cq_{gh} was introduced by Chen and Qin [10] to test for the equality of the population mean vectors. We have

$$cq_{gh} = 2\sqrt{\frac{\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}}{2n_g(n_g-1)} + \frac{\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_h^2)}}{2n_h(n_h-1)} + \frac{\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_h)}}{n_g n_h}},$$

where, for each $\ell \in \{g, h\}$,

$$\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_\ell^2)} = \frac{1}{n_\ell(n_\ell - 1)} \mathrm{tr}\left\{ \sum_{i \neq j}^{n_\ell} (\mathbf{x}_{\ell i} - \overline{\mathbf{x}}_{\ell(i,j)}) \mathbf{x}_{\ell i}^\top (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_{\ell(i,j)}) \mathbf{x}_{\ell j}^\top \right\},$$

$$\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)} = \frac{1}{n_g n_h} \mathrm{tr}\left\{ \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} (\mathbf{x}_{gi} - \overline{\mathbf{x}}_{g(i)}) \mathbf{x}_{gi}^\top (\mathbf{x}_{hj} - \overline{\mathbf{x}}_{h(j)}) \mathbf{x}_{hj}^\top \right\}.$$

Here, $\overline{\mathbf{x}}_{\ell(i,j)}$ is ℓth sample mean after excluding $\mathbf{x}_{\ell i}$ and $\mathbf{x}_{\ell j}$, and $\overline{\mathbf{x}}_{\ell(i)}$ is the ℓth sample mean without $\mathbf{x}_{\ell i}$. Note that cq_{gh} is a ratio-consistent estimator of σ_{gh} under $\boldsymbol{\mu}_g = \boldsymbol{\mu}_h$, (E1), (E2) and (E3'). However, this estimator is not suitable for the following reasons. Let $T_{cq} = (\widehat{\delta}_{gh} - \delta_{gh})/cq_{gh}$ and $z_{\alpha/2}$ denote the upper $\alpha/2$ quantile of the standard normal distribution. Under assumptions (E1), (E2) and (E3'),

$$P = \Pr(-z_{\alpha/2} \leq T_{cq} \leq z_{\alpha/2}) = 1 - 2\Phi(-\widetilde{\sigma}_{gh} z_{\alpha/2}/\sigma_{gh}) + o(1) \tag{6.2}$$

as $p \to \infty$. Here,

$$\widetilde{\sigma}_{gh} = 2\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}{2n_g(n_g - 1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_h^2)}{2n_h(n_h - 1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)}{n_g n_h}}.$$

However, it is not suitable to use cq_{gh} because P generally does not converge to $1 - \alpha$. To clarify this point, we consider the following four cases:

$$\begin{aligned}
&\text{(a1):} \quad n_g \Delta_{gh}/\mathrm{tr}(\boldsymbol{\Sigma}_g^2) = o(1),\ n_h \Delta_{hg}/\mathrm{tr}(\boldsymbol{\Sigma}_h^2) = o(1);\\
&\text{(a2):} \quad n_g \Delta_{gh}/\mathrm{tr}(\boldsymbol{\Sigma}_g^2) = O(1),\ n_h \Delta_{hg}/\mathrm{tr}(\boldsymbol{\Sigma}_h^2) = o(1);\\
&\text{(a3):} \quad n_g \Delta_{gh}/\mathrm{tr}(\boldsymbol{\Sigma}_g^2) = o(1),\ n_h \Delta_{hg}/\mathrm{tr}(\boldsymbol{\Sigma}_h^2) = O(1);\\
&\text{(a4):} \quad n_g \Delta_{gh}/\mathrm{tr}(\boldsymbol{\Sigma}_g^2) \to \infty \ \text{ or } \ n_h \Delta_{hg}/\mathrm{tr}(\boldsymbol{\Sigma}_h^2) \to \infty.
\end{aligned}$$

We assume (E1), (E2) and (E3'). Then,

$$P = \begin{cases} 1 - \alpha + o(1) & \text{as } p \to \infty \ \text{ under } \ \text{(a1)}, \\ 1 - \alpha + O(1) & \text{as } p \to \infty \ \text{ under } \ \text{(a2) or (a3)}, \\ o(1) & \text{as } p \to \infty \ \text{ under } \ \text{(a4)}, \end{cases}$$

i.e., P has a bias except for situation (a1) close to the null hypothesis $\boldsymbol{\mu}_g = \boldsymbol{\mu}_h$. Based on the key estimators that we introduced in Theorem 6.2.1, we propose the following estimator for σ_{gh}:

$$\widehat{\sigma}_{gh} = 2\sqrt{\frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g^2)}}{2n_g(n_g - 1)} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_h^2)}}{2n_h(n_h - 1)} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h)}}{n_g n_h} + \max\left(0, \frac{\widehat{\Delta}_{gh}}{n_g} + \frac{\widehat{\Delta}_{hg}}{n_h}\right)}.$$

Using Theorem 6.2.1, under (E1), (E2) and (E3'), we get the following ratio consistency:

$$\widehat{\sigma}_{gh} = \sigma_{gh}\{1 + o_p(1)\}. \tag{6.3}$$

Next, we establish the asymptotic normality of the Studentized statistic defined by

$$T_s = \frac{\widehat{\delta}_{gh} - \delta_{gh}}{\widehat{\sigma}_{gh}}.$$

From Slutsky's theorem, (6.3) and Corollary 6.1.1, we obtain the asymptotic distribution of the Studentized statistic T_s in the following corollary.

Corollary 6.2.1 (Hyodo, Watanabe, and Seo [22]). *Under assumptions (E1), (E2), and (E3'), $T_s \rightsquigarrow \mathcal{N}(0,1)$ as $p \to \infty$.*

From Corollary 6.2.1, under assumptions (E1), (E2), and (E3'), we have

$$\Pr\left(-z_{\alpha/2} \leq T_s \leq z_{\alpha/2}\right) = 1 - \alpha + o(1) \tag{6.4}$$

as $p \to \infty$. Comparing (6.2) and (6.4), it is clear that T_s should be used for confidence intervals rather than T_{cq}.

Finally, after applying the Bonferroni inequality and Corollary 6.2.1, we propose the approximate simultaneous confidence intervals:

$$\forall_{g,h\in\{1,\ldots,k\}} \text{ with } g < h \quad CI_{gh} = (\max(\widehat{\delta}_{gh} - L_{gh}, 0),\ \widehat{\delta}_{gh} + L_{gh}) \tag{6.5}$$

where $L_{gh} = \widehat{\sigma}_{gh} z_{\alpha/\{k(k-1)\}}$. The asymptotic coverage probability is given in the following theorem.

Theorem 6.2.2 (Hyodo, Watanabe, and Seo [22]). *Under assumptions (E1),(E2), and (E3'), as $p \to \infty$, $\Pr(\forall_{g<h\in\{1,\ldots,k\}}\ \delta_{gh} \in CI_{gh}) \geq 1 - \alpha$.*

Proof. Combining the Bonferroni inequality and Corollary 6.2.1, we get

$$\begin{aligned}\Pr(\forall_{g<h\in\{1,\ldots,k\}}\ \delta_{gh} \in CI_{gh}) &\geq 1 - \sum_{g<h}^{k} \left\{1 - \Pr\left(-z_{\alpha/\{k(k-1)\}} \leq T_s \leq z_{\alpha/\{k(k-1)\}}\right)\right\} \\ &= 1 - \alpha + o(1).\end{aligned}$$

□

6.3 Numerical experiment

In this section, we investigate the performance of the proposed approximate confidence interval (6.5). To concentrate on confirming the approximate accuracy of the asymptotic distribution derived in Corollary 6.2.1, we conduct a numerical experiment with $k = 2$.

6.3.1 Empirical coverage probability

In this subsection, we investigate the empirical coverage probability of our confidence interval under various distributions. The data were generated from the model defined, for all $i \in \{1, \ldots, n_g\}$ and $g \in \{1, 2\}$, by

$$\mathbf{x}_{gi} = \boldsymbol{\Sigma}_g^{1/2} \mathbf{z}_{gi} + \boldsymbol{\mu}_g, \tag{6.6}$$

where

$$\boldsymbol{\mu}_1 = (n_1+n_2)p^{-1/2}(10,\ldots,10)^\top, \quad \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \quad \boldsymbol{\Sigma}_1 = \mathbf{B}(0.3^{|i-j|})\mathbf{B}, \quad \boldsymbol{\Sigma}_2 = 1.2\boldsymbol{\Sigma}_1.$$

Here,

$$B = \mathrm{diag}\left(\left(0.5+\frac{1}{p+1}\right)^{1/2}, \left(0.5+\frac{2}{p+1}\right)^{1/2}, \ldots, \left(0.5+\frac{p}{p+1}\right)^{1/2}\right).$$

The random vector $\mathbf{z}_{gj} = (z_{gij})$ in (6.6) is generated from either one of the following distributions:

(P1) $\mathbf{z}_{gi} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$;

(P2) $z_{gij} = (u_{gij} - 10)/\sqrt{20}$ for $u_{gij} \sim \chi^2_{10}$;

(P3) $\mathbf{z}_{gi} = \sqrt{4/5}\mathbf{u}_{gi}$ for $\mathbf{u}_{gi} \sim \mathcal{T}_p(10, \mathbf{0}, \mathbf{I}_p)$;

(P4) $\mathbf{z}_{gi} = \left\{\mathbf{I}_p - \frac{2}{\pi(1+p)}\mathbf{1}\mathbf{1}^\top\right\}^{-1/2}\left\{\mathbf{u}_{gi} - \sqrt{\frac{2}{\pi(p+1)}}\mathbf{1}\right\}$ for $\mathbf{u}_{gi} \sim \mathcal{SN}_p(\mathbf{0}, \mathbf{I}_p, \mathbf{1})$.

In the setting (P3), we use the standardized multivariate t variable for $\mathbf{z}_{gi}$. In the setting (P4), we set the standardized multivariate skew normal variable for $\mathbf{z}_{gi}$. Note that settings (P1)–(P4) all satisfy $\mathrm{E}(\mathbf{z}_{gi}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{z}_{gi}) = \mathbf{I}_p$, and that settings (P1)–(P3) satisfy the moment condition (E1). The values of n_1, n_2, and p are chosen as follows:

Dimension: $p \in \{32,\ 64,\ 128,\ 256,\ 512\}$

Balanced sample: $(n_1, n_2) \in \{(16,16),\ (32,32),\ (64,64),\ (128,128),\ (256,256)\}$

Unbalanced sample: $(n_1, n_2) \in \{(20,12),\ (40,24),\ (80,48),\ (160,96),\ (320,192)\}$

Unbalanced sample: $(n_1, n_2) \in \{(12,20),\ (24,40),\ (48,80),\ (96,160),\ (192,320)\}$

In each case, we computed the empirical coverage probability based on 10^5 replications. The results are listed in Tables 6.1–6.12.

As it can be seen from Tables 6.1–6.12, when the sample sizes n_1 and n_2 or dimension p are increased, the empirical coverage probability becomes very close to the nominal confidence level. The coverage probability is smaller than the nominal confidence level when p, n_1, and n_2 are small. However, the coverage probability approaches the nominal confidence level if p is small, but n_1 and n_2 are large. The setting (P4) does not satisfy (E1), but we see that for setting (P4) the approximation accuracy shows no sudden drops.

Table 6.1: The empirical coverage probability when (P1)

$(n_1, n_2)/p$		32	64	128	256	512
$(16, 16)$	0.90	0.890	0.889	0.891	0.890	0.892
	0.95	0.941	0.941	0.942	0.941	0.942
	0.99	0.985	0.985	0.985	0.985	0.985
$(32, 32)$	0.90	0.895	0.893	0.894	0.895	0.896
	0.95	0.945	0.945	0.945	0.945	0.946
	0.99	0.987	0.988	0.988	0.987	0.988
$(64, 64)$	0.90	0.897	0.898	0.898	0.898	0.897
	0.95	0.947	0.948	0.948	0.948	0.947
	0.99	0.988	0.989	0.990	0.989	0.988
$(128, 128)$	0.90	0.900	0.900	0.897	0.899	0.899
	0.95	0.949	0.950	0.948	0.949	0.949
	0.99	0.990	0.990	0.989	0.989	0.990
$(256, 256)$	0.90	0.901	0.900	0.900	0.901	0.901
	0.95	0.951	0.949	0.950	0.950	0.950
	0.99	0.990	0.990	0.989	0.990	0.990

Table 6.2: The empirical coverage probability when (P2)

$(n_1, n_2)/p$		32	64	128	256	512
$(16, 16)$	0.90	0.889	0.889	0.890	0.889	0.891
	0.95	0.940	0.940	0.941	0.941	0.941
	0.99	0.985	0.984	0.985	0.985	0.984
$(32, 32)$	0.90	0.895	0.894	0.896	0.895	0.896
	0.95	0.946	0.946	0.946	0.946	0.947
	0.99	0.987	0.988	0.988	0.988	0.988
$(64, 64)$	0.90	0.900	0.897	0.898	0.897	0.899
	0.95	0.949	0.948	0.949	0.948	0.949
	0.99	0.989	0.989	0.989	0.989	0.989
$(128, 128)$	0.90	0.900	0.898	0.898	0.898	0.902
	0.95	0.950	0.949	0.949	0.948	0.950
	0.99	0.990	0.989	0.990	0.990	0.990
$(256, 256)$	0.90	0.900	0.899	0.901	0.903	0.901
	0.95	0.950	0.950	0.950	0.951	0.950
	0.99	0.989	0.990	0.990	0.990	0.990

Table 6.3: The empirical coverage probability when (P3)

$(n_1, n_2)/p$		32	64	128	256	512
$(16, 16)$	0.90	0.890	0.890	0.891	0.890	0.890
	0.95	0.941	0.941	0.942	0.940	0.942
	0.99	0.986	0.986	0.985	0.986	0.986
$(32, 32)$	0.90	0.894	0.895	0.894	0.895	0.894
	0.95	0.946	0.946	0.945	0.946	0.944
	0.99	0.988	0.988	0.988	0.988	0.987
$(64, 64)$	0.90	0.899	0.897	0.898	0.897	0.897
	0.95	0.949	0.947	0.948	0.948	0.948
	0.99	0.990	0.989	0.989	0.989	0.989
$(128, 128)$	0.90	0.898	0.898	0.898	0.900	0.900
	0.95	0.949	0.948	0.948	0.949	0.951
	0.99	0.990	0.990	0.989	0.989	0.990
$(256, 256)$	0.90	0.899	0.899	0.900	0.899	0.900
	0.95	0.950	0.949	0.950	0.951	0.950
	0.99	0.990	0.990	0.990	0.990	0.990

Table 6.4: The empirical coverage probability when (P4)

$(n_1, n_2)/p$		32	64	128	256	512
$(16, 16)$	0.90	0.889	0.890	0.890	0.887	0.890
	0.95	0.940	0.940	0.941	0.940	0.941
	0.99	0.985	0.984	0.985	0.985	0.985
$(32, 32)$	0.90	0.895	0.894	0.895	0.896	0.895
	0.95	0.946	0.945	0.945	0.946	0.946
	0.99	0.988	0.989	0.987	0.987	0.988
$(64, 64)$	0.90	0.896	0.897	0.896	0.897	0.899
	0.95	0.947	0.948	0.947	0.948	0.949
	0.99	0.989	0.988	0.989	0.989	0.989
$(128, 128)$	0.90	0.898	0.896	0.897	0.899	0.898
	0.95	0.949	0.947	0.948	0.950	0.949
	0.99	0.989	0.989	0.990	0.990	0.989
$(256, 256)$	0.90	0.899	0.897	0.900	0.897	0.900
	0.95	0.949	0.948	0.950	0.949	0.949
	0.99	0.989	0.989	0.990	0.990	0.990

Table 6.5: The empirical coverage probability when (P1)

$(n_1, n_2)/p$		32	64	128	256	512
$(20, 12)$	0.90	0.884	0.885	0.886	0.888	0.896
	0.95	0.936	0.937	0.937	0.938	0.937
	0.99	0.983	0.983	0.983	0.983	0.983
$(40, 24)$	0.90	0.894	0.893	0.894	0.893	0.893
	0.95	0.944	0.944	0.944	0.944	0.943
	0.99	0.986	0.987	0.989	0.986	0.987
$(80, 48)$	0.90	0.897	0.895	0.896	0.896	0.899
	0.95	0.947	0.946	0.947	0.947	0.948
	0.99	0.988	0.988	0.989	0.989	0.989
$(160, 96)$	0.90	0.897	0.899	0.897	0.898	0.897
	0.95	0.948	0.949	0.948	0.948	0.948
	0.99	0.990	0.989	0.989	0.989	0.990
$(320, 192)$	0.90	0.898	0.899	0.900	0.900	0.899
	0.95	0.948	0.949	0.949	0.949	0.949
	0.99	0.989	0.990	0.989	0.990	0.989

Table 6.6: The empirical coverage probability when (P2)

$(n_1, n_2)/p$		32	64	128	256	512
$(20, 12)$	0.90	0.887	0.887	0.884	0.885	0.884
	0.95	0.938	0.938	0.937	0.937	0.936
	0.99	0.984	0.983	0.982	0.982	0.983
$(40, 24)$	0.90	0.893	0.893	0.893	0.894	0.893
	0.95	0.944	0.945	0.944	0.945	0.944
	0.99	0.987	0.987	0.987	0.987	0.986
$(80, 48)$	0.90	0.896	0.897	0.897	0.896	0.897
	0.95	0.947	0.947	0.948	0.947	0.947
	0.99	0.989	0.988	0.989	0.989	0.989
$(160, 96)$	0.90	0.899	0.897	0.898	0.899	0.898
	0.95	0.948	0.948	0.949	0.950	0.949
	0.99	0.989	0.989	0.989	0.989	0.989
$(320, 192)$	0.90	0.900	0.900	0.899	0.897	0.899
	0.95	0.950	0.949	0.950	0.949	0.950
	0.99	0.989	0.989	0.990	0.989	0.990

Table 6.7: The empirical coverage probability when (P3)

$(n_1, n_2)/p$		32	64	128	256	512
$(20, 12)$	0.90	0.887	0.887	0.887	0.887	0.887
	0.95	0.939	0.939	0.938	0.939	0.939
	0.99	0.984	0.984	0.984	0.984	0.984
$(40, 24)$	0.90	0.892	0.893	0.893	0.894	0.893
	0.95	0.943	0.944	0.944	0.944	0.944
	0.99	0.987	0.987	0.987	0.987	0.987
$(80, 48)$	0.90	0.897	0.898	0.896	0.896	0.896
	0.95	0.948	0.947	0.946	0.947	0.947
	0.99	0.989	0.989	0.988	0.988	0.988
$(160, 96)$	0.90	0.897	0.899	0.897	0.897	0.897
	0.95	0.948	0.949	0.948	0.947	0.946
	0.99	0.989	0.990	0.990	0.989	0.988
$(320, 192)$	0.90	0.900	0.900	0.899	0.900	0.899
	0.95	0.950	0.950	0.949	0.949	0.950
	0.99	0.990	0.990	0.990	0.990	0.989

Table 6.8: The empirical coverage probability when (P4)

$(n_1, n_2)/p$		32	64	128	256	512
$(20, 12)$	0.90	0.886	0.886	0.887	0.886	0.885
	0.95	0.938	0.937	0.938	0.938	0.936
	0.99	0.983	0.983	0.983	0.983	0.983
$(40, 24)$	0.90	0.893	0.892	0.894	0.894	0.893
	0.95	0.944	0.943	0.944	0.945	0.944
	0.99	0.987	0.986	0.987	0.987	0.987
$(80, 48)$	0.90	0.898	0.896	0.898	0.898	0.899
	0.95	0.948	0.946	0.948	0.948	0.949
	0.99	0.989	0.988	0.989	0.989	0.989
$(160, 96)$	0.90	0.897	0.899	0.899	0.899	0.898
	0.95	0.948	0.949	0.949	0.949	0.949
	0.99	0.989	0.989	0.989	0.990	0.989
$(320, 192)$	0.90	0.900	0.899	0.899	0.899	0.898
	0.95	0.950	0.949	0.949	0.949	0.948
	0.99	0.990	0.989	0.990	0.990	0.989

Table 6.9: The empirical coverage probability when (P1)

$(n_1, n_2)/p$		32	64	128	256	512
$(12, 20)$	0.90	0.888	0.888	0.889	0.888	0.888
	0.95	0.939	0.938	0.940	0.939	0.939
	0.99	0.983	0.984	0.984	0.983	0.984
$(24, 40)$	0.90	0.894	0.893	0.894	0.894	0.893
	0.95	0.945	0.945	0.945	0.945	0.944
	0.99	0.987	0.987	0.987	0.987	0.987
$(48, 80)$	0.90	0.896	0.896	0.898	0.896	0.898
	0.95	0.947	0.947	0.948	0.946	0.948
	0.99	0.988	0.988	0.990	0.988	0.989
$(96, 160)$	0.90	0.899	0.898	0.898	0.899	0.899
	0.95	0.949	0.949	0.948	0.949	0.950
	0.99	0.989	0.989	0.989	0.990	0.990
$(192, 320)$	0.90	0.900	0.899	0.899	0.900	0.900
	0.95	0.949	0.950	0.949	0.949	0.950
	0.99	0.990	0.989	0.990	0.990	0.990

Table 6.10: The empirical coverage probability when (P2)

$(n_1, n_2)/p$		32	64	128	256	512
$(12, 20)$	0.90	0.887	0.889	0.889	0.889	0.887
	0.95	0.938	0.941	0.940	0.939	0.938
	0.99	0.984	0.984	0.985	0.984	0.983
$(24, 40)$	0.90	0.894	0.896	0.894	0.894	0.894
	0.95	0.944	0.946	0.944	0.945	0.944
	0.99	0.987	0.988	0.987	0.987	0.987
$(48, 80)$	0.90	0.897	0.898	0.897	0.896	0.898
	0.95	0.947	0.948	0.947	0.947	0.948
	0.99	0.988	0.988	0.988	0.989	0.989
$(96, 160)$	0.90	0.899	0.898	0.900	0.899	0.898
	0.95	0.950	0.949	0.949	0.950	0.949
	0.99	0.989	0.990	0.989	0.990	0.990
$(192, 320)$	0.90	0.899	0.899	0.900	0.899	0.900
	0.95	0.948	0.949	0.950	0.950	0.950
	0.99	0.989	0.990	0.990	0.990	0.990

Table 6.11: The empirical coverage probability when (P3)

$(n_1, n_2)/p$		32	64	128	256	512
$(12, 20)$	0.90	0.890	0.888	0.889	0.889	0.889
	0.95	0.940	0.940	0.940	0.941	0.940
	0.99	0.985	0.985	0.985	0.985	0.984
$(24, 40)$	0.90	0.894	0.894	0.896	0.894	0.895
	0.95	0.945	0.945	0.947	0.944	0.945
	0.99	0.987	0.987	0.989	0.988	0.989
$(48, 80)$	0.90	0.896	0.897	0.897	0.898	0.899
	0.95	0.948	0.948	0.947	0.949	0.950
	0.99	0.989	0.989	0.988	0.989	0.990
$(96, 160)$	0.90	0.898	0.898	0.899	0.898	0.899
	0.95	0.948	0.948	0.949	0.948	0.949
	0.99	0.989	0.989	0.990	0.989	0.990
$(192, 320)$	0.90	0.899	0.898	0.900	0.899	0.900
	0.95	0.949	0.949	0.950	0.950	0.950
	0.99	0.990	0.989	0.990	0.990	0.990

Table 6.12: The empirical coverage probability when (P4)

$(n_1, n_2)/p$		32	64	128	256	512
$(12, 20)$	0.90	0.888	0.888	0.887	0.887	0.888
	0.95	0.939	0.939	0.939	0.938	0.939
	0.99	0.984	0.984	0.984	0.985	0.984
$(24, 40)$	0.90	0.894	0.894	0.895	0.893	0.894
	0.95	0.945	0.945	0.945	0.945	0.944
	0.99	0.987	0.987	0.988	0.987	0.987
$(48, 80)$	0.90	0.899	0.896	0.897	0.897	0.897
	0.95	0.948	0.947	0.948	0.947	0.947
	0.99	0.989	0.988	0.989	0.988	0.989
$(96, 160)$	0.90	0.899	0.897	0.898	0.898	0.900
	0.95	0.949	0.948	0.949	0.948	0.950
	0.99	0.989	0.989	0.990	0.989	0.989
$(192, 320)$	0.90	0.897	0.898	0.898	0.898	0.900
	0.95	0.948	0.949	0.948	0.948	0.949
	0.99	0.990	0.989	0.989	0.990	0.990

6.3.2 Compare the empirical coverage probability

In this subsection, We compare the empirical coverage probability of the confidence intervals described below under various situations and three different confidence levels $\alpha \in \{0.01, 0.05, 0.1\}$:

- $CI1 = (\max(\widehat{\delta}_{12} - cq_{12}z_{\alpha/2}, 0), \widehat{\delta}_{12} + cq_{12}z_{\alpha/2})$.
- $CI2 = (\max(\widehat{\delta}_{12} - \widehat{\sigma}_{12}z_{\alpha/2}, 0), \widehat{\delta}_{12} + \widehat{\sigma}_{12}z_{\alpha/2})$.

The confidence interval $CI1$ is based on test statistics for a two-sample test given by Chen and Qin [10]. The confidence interval $CI2$ is our proposed method (6.5) with $k = 2$. These asymptotic coverage probabilities are obtained by (6.2) and (6.4). For these simulations, we set $n_1 = n_2 = p$ for $p \in \{32,\ 64,\ 128,\ 256,\ 512\}$. The data were generated from the following model:

$$\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1} \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),\ \mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2} \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

where

$$\boldsymbol{\mu}_1 = p^{-\eta}(\sqrt{5}, \ldots, \sqrt{5})^\top,\ \boldsymbol{\mu}_2 = \mathbf{0},\ \boldsymbol{\Sigma}_1 = \mathbf{B}(0.3^{|i-j|})\mathbf{B},\ \boldsymbol{\Sigma}_2 = 1.2\boldsymbol{\Sigma}_1.$$

Here, $\eta \in \{0.1, 0.2, \ldots, 1.0\}$. To reflect the difference in variance for each component $\mathbf{x}_{gi}$, we choose the covariance structure $\boldsymbol{\Sigma}_g$ that is a first-order autoregressive structure with heterogeneous variances. The correlation between any two elements is equal to 0.3 for adjacent elements, 0.3^2 for two elements separated by a third, and so on. The empirical coverage probabilities were calculated with 10^5 replications. The resulting coverage probabilities for $CI1$ and $CI2$ intervals are presented in Figs. 1–5. These figures are scatter plots of the coverage probability [vertical axis] versus each η [horizontal axis]. In each graph, the circle (∘), black circle (•), square (□), black square (■), triangle (△), and black triangle (▲) marks denote the empirical coverage probabilities of the 0.90 confidence interval $CI1$, the 0.90 confidence interval $CI2$, the 0.95 confidence interval $CI1$, the 0.95 confidence interval $CI2$, the 0.99 confidence interval $CI1$, and the 0.99 confidence interval $CI2$, respectively.

Note that a large η corresponds to a small $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ because $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = 5p^{1-2\eta}$. When $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ is large, the empirical coverage probability of $CI1$ is much smaller than the nominal confidence level, while that of $CI2$ is close to the nominal confidence level. Overall, the empirical coverage probability of $CI2$ is close to the nominal confidence, but it is somewhat conservative when $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ is relatively small.

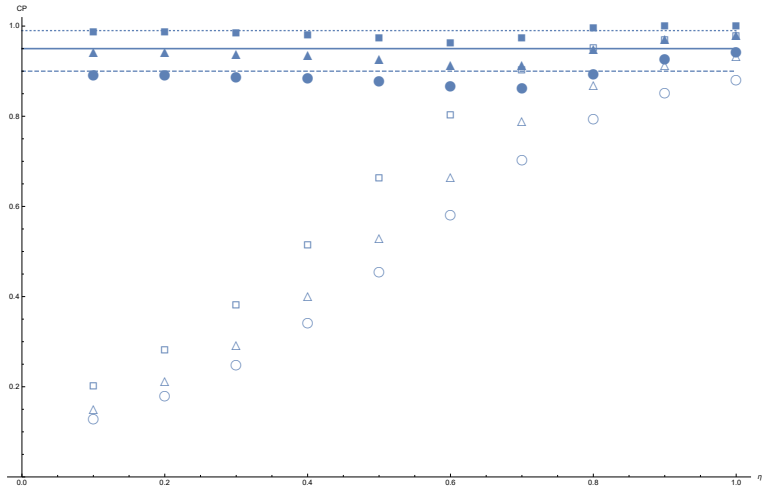

Figure 6.1: The empirical coverage probabilities when $p = 32$

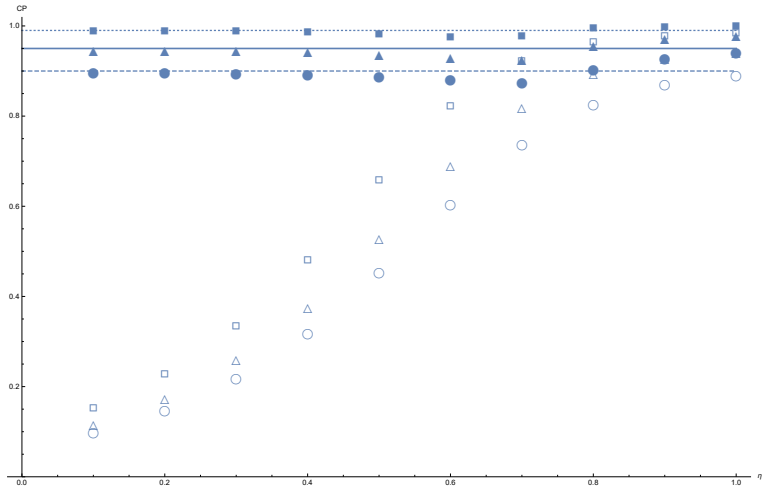

Figure 6.2: The empirical coverage probabilities when $p = 64$

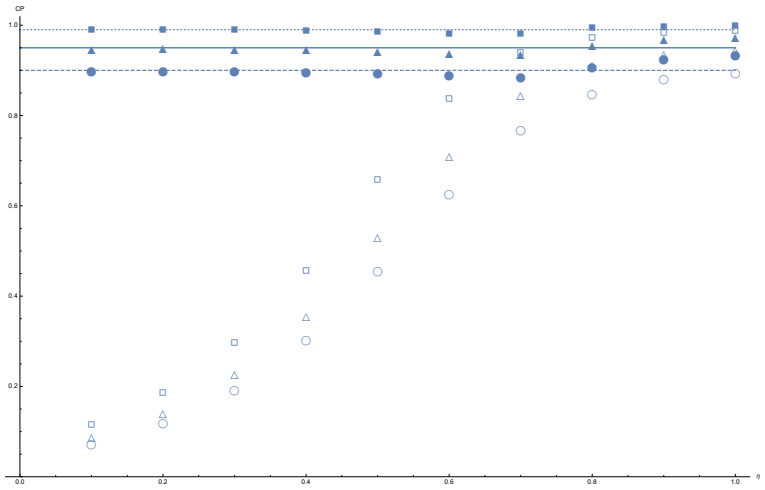

Figure 6.3: The empirical coverage probabilities when $p = 128$

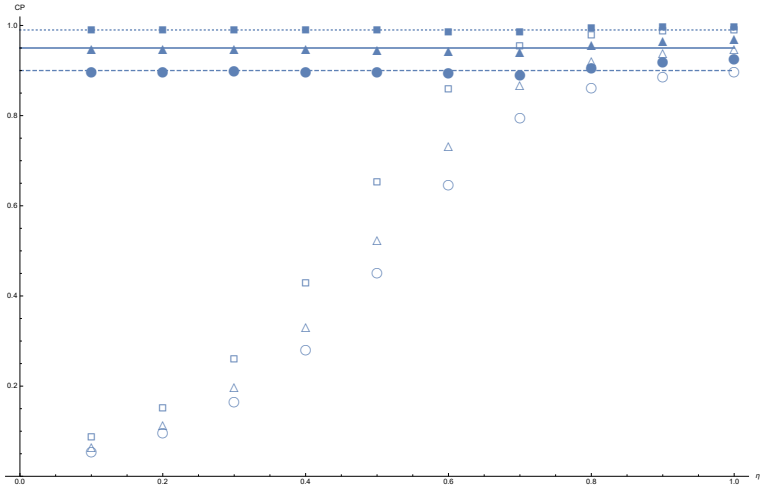

Figure 6.4: The empirical coverage probabilities when $p = 256$

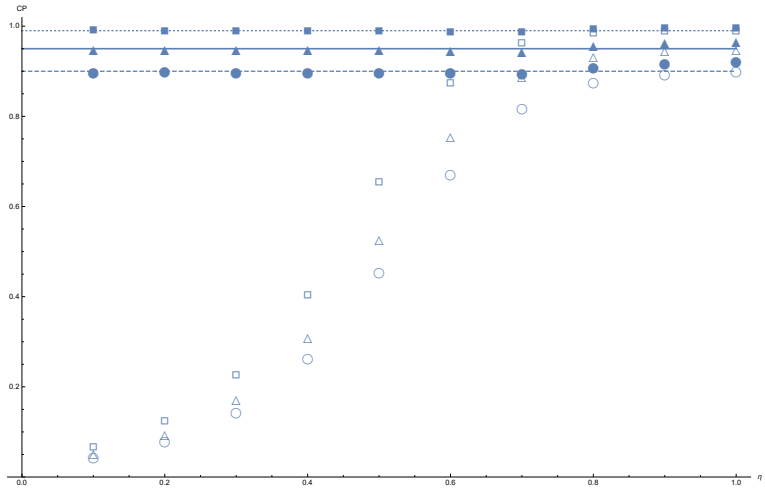

Figure 6.5: The empirical coverage probabilities when $p = 512$

6.3.3 Real data analysis

In this subsection, we use training data to compare the population means of the expression of four types of genes. Khan et al. [25] studied the expression of the following four types of genes of childhood small round blue cell tumors (SRBCTs): (i) the Ewing family of tumors (EWS, 23 cases); (ii) Burkitt lymphoma, a subset of non-Hodgkin lymphoma (BL, 8 cases); (iii) neuroblastoma (NB, 12 cases); and (iv) rhabdomyosarcoma (RMS, 21 cases). Contained in this dataset are gene expression profiles from both tumor biopsy and cell line samples. The dataset contains the filtered dataset of 2308 gene expression profiles, as described by Khan et al. [25]; this dataset is available from http://bioinf.ucd.ie/people/aedin/R/.

As a method of classifying microarray data, Aoshima and Yata [3] proposed the Euclidean distance classifier. The accuracy of this classifier depends on the Euclidean distance of all pair-wise differences between the mean vectors of the populations. We construct a simultaneous confidence interval for all differences of population mean vectors $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$ using our proposed method. The approximate simultaneous confidence interval of 95% confidence was constructed as follows:

$$
\begin{aligned}
&p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{EWS}} - \boldsymbol{\mu}_{\mathrm{BL}}\|^2 : (4.2, 7.2), \quad p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{EWS}} - \boldsymbol{\mu}_{\mathrm{NB}}\|^2 : (2.4, 6.0),\\
&p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{EWS}} - \boldsymbol{\mu}_{\mathrm{RMS}}\|^2 : (2.2, 5.2), \quad p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{BL}} - \boldsymbol{\mu}_{\mathrm{NB}}\|^2 : (4.2, 5.2),\\
&p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{BL}} - \boldsymbol{\mu}_{\mathrm{RMS}}\|^2 : (5.2, 9.7), \quad p^{-0.6}\|\boldsymbol{\mu}_{\mathrm{NB}} - \boldsymbol{\mu}_{\mathrm{RMS}}\|^2 : (2.1, 6.8).
\end{aligned}
$$

From this result, we can see that the Euclidean norm of the difference vector of the two mean vectors has a large value of about $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 = O(p^{0.6})$.

6.4 Conclusion

We discussed the construction of confidence intervals for $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$ in high-dimensional settings. Under the non-normal assumption and high-dimensional settings, we derived

the asymptotic distribution of a quadratic form in the set of sample mean vectors, and an unbiased estimator Δ_{gh}. Using these methods, we obtained an approximate confidence interval for $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$. The performance of this approximate confidence interval was evaluated via simulation. It was confirmed that the coverage probability of the proposed confidence interval is close to nominal even in some non-normal settings. Thus, the proposed confidence interval is robust against departures from normality without impairing the approximation accuracy. Furthermore, we applied the proposed method to a microarray data set and evaluated $\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2$.

Bibliography

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[2] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis. Design Methods & Applications*, **30**, 356–399.

[3] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, **66**, 983–1010.

[4] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, **28**, 43–62.

[5] Aoshima, M. and Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, **71**, 473–503.

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

[7] 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. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, **10**, 989–1010.

[8] Cai, T.T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **76**, 349–372.

[9] Chan, Y.-B. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika.* **96** , 175–186.

[10] Chen, S.X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, **38**, 808–835.

[11] Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *The Journal of the London Mathematical Society*, s1-**11**, 290–294.

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

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

[14] Fujikoshi, Y. (2000). Error bounds for asymptotic approximations of the linear discriminant function when the sample size and dimensionality are large. *Journal of Multivariate Analysis*, **73**, 1–17.

[15] Fujikoshi, Y. and Seo, T. (1998). Asymptotic approximations for EPMC's of the linear and the quadratic discriminant function when the sample size and the dimension are large. *Random Operators and Stochastic Equations*, **6**, 269–280.

[16] Gregory, K.B., Carroll, R.J., Baladandayuthapani, V., and Lahiri, S.N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, **110**, 837–849.

[17] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*, Academic Press, New York.

[18] Hall, P., Marron, J.S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of Royal Statistical Society. Series B. Statistical Methodology*, **67**, 427–444.

[19] Hall, P., Pittelknow Y., and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of Royal Statistical Society. Series B. Statistical Methodology*, **70**, 159–173.

[20] Himeno, T. and Yamada, T. (2014). Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. *Journal of Multivariate Analysis*, **130**, 27–44.

[21] Hu, J., Bai, Z. Wang, C., and Wang, W. (2017). On testing the equality of high dimensional mean vectors with unequal covariance matrices. *Annals of the Institude of Statistical Mathematics*, *69*, 365–387.

[22] Hyodo, M., Watanabe, H and Seo, T. (2018). On simultaneous confidence interval estimation for the difference of paired mean vectors in high-dimensional settings. *Journal of Multivariate Analysis*, **168**, 160–173.

[23] Hyodo, M., Yamada, T., Himeno, T. and Seo, T. (2012). A modified linear discriminant analysis for high-dimensional data. *Hiroshima Mathematical Journal*, **42**, 209–231.

[24] Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis, 2nd ed,* Prentice Hall.

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

[26] Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics. Journal of the Biometric Society*, **24**, 823–834.

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

[28] Ng, K. and Liu, H. (2000). Customer retention via data mining. *AI Review*, **14**, 569–590.

[29] Nishiyama, T., Hyodo, M., Seo, T., and Pavlenko, T. (2013). Testing linear hypotheses of mean vectors for high-dimensional data with unequal covariance matrices. *Journal of Statistical Planning and Inference*, **143**, 1898–1911.

[30] Okamoto, M. (1963). An asymptotic expansion of the distribution of the linear discriminant function. *The Annals of Mathematical Statistics*, **34**, 1286–1301.

[31] Okamoto, M. (1968). Correction to "An asymptotic expansion of the distribution of the linear discriminant function." *The Annals of Mathematical Statistics*, **39**, 1358–1359.

[32] Park, J. and Ayyala, D.N. (2013) . A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference*, **143**, 929–943.

[33] Raudys, S. (1972). On the amount of a priori information in construction of a classification algorithm. *Engineering Cybernetics. English Edition of Tekhnicheskaya Kibernetika*, **4**, 711–718.

[34] Rui, Y., Huang, T. S. and Chang, S. (1999). Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, **10**, 39–62.

[35] Saranadasa, H. (1993). Asymptotic expansion of the misclassification probabilities of D- and A-criteria for distribution from two high-dimensional populations using the theory of large-dimensional random matrices. *Journal of Multivariate Analysis*, **46**, 154–174.

[36] Shiryaev, A.N. (1984). *Probability, Second Ed*, Springer-Verlag.

[37] Siotani, M. (1982). Large sample approximations and asymptotic expansions of classification statistic. *Handbook of Statistics* 2 (P. R. Krishnaiah and L. N. Kanal, Eds.), North-Holland Publishing Company, 61-100.

[38] Siotani, M. and Wang, R. H. (1977). Asymptotic expansions for error rates and comparison of the W-procedure and the Z-procedure in discriminant analysis. *In Multivariate Analysis (Vol. IV), P. M. Krishnaiah(Ed.). Amsterdam:North-Holland*, 523–545.

[39] Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society*, **37**, 53–86.

[40] Srivastava, M.S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, **114**, 349–358.

[41] Srivastava, R., Li, P., and Ruppert, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, **25**, 954–970.

[42] Watanabe, H., Hyodo, M., and Seo, T. (2019). An estimator of misclassification probability for multi-class Euclidean distance classifier in high-dimensional data. *SUT Journal of Mathematics*, to appear.

[43] Watanabe, H., Hyodo, M., Yamada, Y., and Seo, T. (2019). Estimation of misclassification probability for a distance-based classifier in high-dimensional data. *Hiroshima Mathematical Journal*, to appear.

[44] Xing, E., Jordan, M. and Karp, R. (2001). Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 601–608.

[45] Yamada, T. and Himeno, T. (2015). Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. *Journal of Multivariate Analysis*, **139**, 7–27.

[46] Yang, Y. and Pederson, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420.