Ann Inst Stat Math (2014) 66:983–1010
DOI 10.1007/s10463-013-0435-8

A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data

Makoto Aoshima · Kazuyoshi Yata

Received: 5 November 2012 / Revised: 20 August 2013 / Published online: 29 October 2013

Abstract In this paper, we consider a scale adjusted-type distance-based classifier for high-dimensional data. We first give such a classifier that can ensure high accuracy in misclassification rates for two-class classification. We show that the classifier is not only consistent but also asymptotically normal for high-dimensional data. We provide sample size determination so that misclassification rates are no more than a prespecified value. We propose a classification procedure called the *misclassification rate adjusted classifier*. We further develop the classifier to multiclass classification. We show that the classifier can still enjoy asymptotic properties and ensure high accuracy in misclassification rates for multiclass classification. Finally, we demonstrate the proposed classifier in actual data analyses by using a microarray data set.

Keywords Asymptotic normality · Distance-based classifier · HDLSS · Sample size determination · Two-stage procedure

1 Introduction

High-dimensional data situations occur in many areas of modern science, such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, etc. A

Research of M. Aoshima was partially supported by Grants-in-Aid for Scientific Research (B) and Challenging Exploratory Research, JSPS, under Contract Numbers 22300094 and 23650142. Research of K. Yata was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 23740066.

M. Aoshima (✉) · K. Yata
Institute of Mathematics, University of Tsukuba, 1-1-1 Tennodai,
Tsukuba, Ibaraki 305-8571, Japan
e-mail: aoshima@math.tsukuba.ac.jp

K. Yata
e-mail: yata@math.tsukuba.ac.jp

common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively small. This is the so-called "HDLSS" or "large p, small n" situation where $p/n \to \infty$; here p is the data dimension and n is the sample size. Substantial work had been done on the asymptotic behavior of eigenvalues of the sample covariance matrix in the limit as $p \to \infty$, see Johnstone (2001) and Paul (2007) under Gaussian assumptions and Baik and Silverstein (2006) under non-Gaussian but i.i.d. assumptions. Those literatures handled the cases when p and n increase at the same rate, i.e. $p/n \to c > 0$. The asymptotic behaviors of high-dimensional, low-sample-size (HDLSS) data were studied by Hall et al. (2005), Ahn et al. (2007) and Yata and Aoshima (2012a) when $p \to \infty$ while n is fixed. They explored conditions to give a geometric representation of HDLSS data. The HDLSS asymptotic study usually assumes either the normality for the population distribution or a ρ-mixing condition for the dependency of random variables in a sphered data matrix, see also Jung and Marron (2009). However, Yata and Aoshima (2009) succeeded in investigating consistency properties of both eigenvalues and eigenvectors of the sample covariance matrix in general settings such as including the case when all eigenvalues are in the range of sphericity. In addition, Yata and Aoshima (2010) created the *cross-data-matrix (CDM) methodology* that provides effective inference on the eigenspace for HDLSS data. Aoshima and Yata (2011a,b) developed a variety of inference, including two-class classification, for high-dimensional data based on geometric representations of HDLSS data and presented sample size determination to ensure prespecified accuracy. In this paper, we make a first attempt on multiclass classification for high-dimensional data, ensuring accuracy in misclassification rates. The key is a scale adjusted-type distance-based classifier for multiclass classification.

Suppose we have independent and p-variate populations, $\pi_i,\ i = 1, \ldots, k$, having unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i (\geq \boldsymbol{O})$ for each π_i. *We do not assume that* $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k$. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k)$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \geq \cdots \geq \lambda_{ip} \geq 0$, and $\boldsymbol{H}_i$ is an orthogonal matrix of the corresponding eigenvectors. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from each π_i, where $\boldsymbol{x}_{ij} = (x_{i1j}, \ldots, x_{ipj})^T,\ j = 1, \ldots, n_i$. We assume $n_i \geq 2,\ i = 1, \ldots, k$. Let $\boldsymbol{x}_{ij} = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2} z_{ij} + \boldsymbol{\mu}_i$, where z_{ij} is considered as a sphered data vector from a distribution with the zero mean vector and the identity covariance matrix.

In this paper, we assume the following model as necessary:

$$\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i \boldsymbol{y}_{ij} + \boldsymbol{\mu}_i \tag{1}$$

for $i = 1, \ldots, k;\ j = 1, \ldots, n_i$, where $\boldsymbol{\Gamma}_i$ is a $p \times r_i$ matrix for some $r_i > 0$ such that $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$, and $\boldsymbol{y}_{ij},\ j = 1, \ldots, n_i$, are i.i.d. random vectors having $E(\boldsymbol{y}_{ij}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{y}_{ij}) = \boldsymbol{I}_{r_i}$. Here, $\boldsymbol{I}_{r_i}$ denotes the identity matrix of dimension r_i. See Bai and Saranadasa (1996) and Chen and Qin (2010) for the model. Let $\boldsymbol{y}_{ij} = (y_{i1j}, \ldots, y_{ir_ij})^T$ for all i, j. As for $\boldsymbol{y}_{ij},\ i = 1, \ldots, k$, we assume that

(A-i) The fourth moments of each variable in $\boldsymbol{y}_{ij}$ are uniformly bounded, $E(y_{iqj}^2 y_{isj}^2) = 1$ and $E(y_{iqj} y_{isj} y_{itj} y_{iuj}) = 0$ for all $q \neq s, t, u$.

Note that (1) includes the case that $\boldsymbol{\Gamma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2}$ and $\boldsymbol{y}_{ij} = \boldsymbol{z}_{ij}$. We assume the following assumptions for $\boldsymbol{\Sigma}_i$s as necessary:

$$\text{(A-ii)} \quad \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i^4)}{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)^2} \to 0 \text{ and } \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_l)}{\operatorname{tr}(\boldsymbol{\Sigma}_j^2)} \in (0, \infty) \text{ as } p \to \infty \quad \text{for} \quad i, j, l = 1, \ldots, k.$$

Here, for a function, $f(\cdot)$, "$f(p) \in (0, \infty)$ as $p \to \infty$" implies $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$. Note that "$\operatorname{tr}(\boldsymbol{\Sigma}_i^4)/\operatorname{tr}(\boldsymbol{\Sigma}_i^2)^2 \to 0$ as $p \to \infty$" is equivalent to the condition that "$\lambda_{i1}/\operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/2} \to 0$ as $p \to \infty$".

Remark 1 If π_i is $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, (A-i) naturally follows. If all λ_{ij}s are bounded such as $\lambda_{ij} \in (0, \infty)$ as $p \to \infty$, (A-ii) trivially holds. For a spiked model such as $\lambda_{ij} = a_{ij} p^{\alpha_{ij}}$ $(j = 1, \ldots, t_i)$ and $\lambda_{ij} = c_{ij}$ $(j = t_i + 1, \ldots, p)$ with positive constants, a_{ij}s, c_{ij}s and α_{ij}s, and positive integers t_is, (A-ii) holds under the condition that $\alpha_{ij} < 1/2$ for $j = 1, \ldots, t_i (< \infty)$; $i = 1, \ldots, k$. See Yata and Aoshima (2010) for the details of a spiked model. As an interesting example, (A-ii) holds for $\boldsymbol{\Sigma}_{i'} = c_{i'}(\rho_{i'}^{|i-j|^{q_{i'}}})$, $i' = 1, \ldots, k$, where $c_{i'}$ and $q_{i'}$ are positive constants and $0 < \rho_{i'} < 1$.

Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the k populations. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^{\mathrm{T}}/(n_i - 1)$. When $k = 2$, a typical classification rule is that one classifies an individual into π_1 if

$$\begin{aligned}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^{\mathrm{T}} \boldsymbol{S}_{1n_1}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - \log\left\{\frac{\det(\boldsymbol{S}_{2n_2})}{\det(\boldsymbol{S}_{1n_1})}\right\} \\ < (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^{\mathrm{T}} \boldsymbol{S}_{2n_2}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}),\end{aligned} \tag{2}$$

and into π_2 otherwise. However, the inverse matrix of $\boldsymbol{S}_{in_i}$ does not exist in HDLSS context ($p > n_i$). When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Saranadasa (1993) considered substituting the identity matrix $\boldsymbol{I}_p$ for $\boldsymbol{S}_{in_i}$. Bickel and Levina (2004) considered the inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. Yata and Aoshima (2012a) considered using a ridge-type inverse covariance matrix derived by the *noise reduction methodology*. When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Dudoit et al. (2002) considered using the inverse matrix defined by only diagonal elements of $\boldsymbol{S}_{in_i}$. Aoshima and Yata (2011a) considered substituting $\{\operatorname{tr}(\boldsymbol{S}_{in_i})/p\}\boldsymbol{I}_p$ for $\boldsymbol{S}_{in_i}$ by using the difference of a geometric representation of HDLSS data from each π_i and showed that a quadratic classifier has misclassification rates which are no more than a prespecified value. On the other hand, Hall et al. (2005) and Marron et al. (2007) considered distance weighted classifiers. Hall et al. (2005, 2008) and Chan and Hall (2009) considered distance-based classifiers. The previous references mainly discussed two-class classification in high-dimensional, low sample size settings.

In this paper, we consider a scale adjusted-type distance-based classifier given by Chan and Hall (2009). In Sect. 2, we develop such a classifier that can ensure high accuracy in misclassification rates for two-class classification. We show that the classifier is not only consistent but also asymptotically normal for high-dimensional data. In Sect. 3, we provide sample size determination so that misclassification rates are no more than a prespecified value. We propose a classification procedure called the *misclassification rate adjusted classifier*. In Sect. 4, we further develop the classifier to multiclass classification when $k \geq 3$. This is the first attemp t on multiclass classification ensuring accuracy in misclassification rates. We show that the classifier can still enjoy asymptotic properties and ensure high accuracy in misclassification rates for multiclass classification. Finally, in Sect. 5, we demonstrate the proposed classifier in actual data analyses by using a microarray data set.

2 Two-class classification

Throughout this section, we assume $k = 2$. We consider a classification rule that is given by substituting the identity matrix $\boldsymbol{I}_p$ for $\boldsymbol{S}_{in_i}$ in (2) as follows: one classifies an individual into π_1 if

$$\left(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\right)^{\mathrm{T}} (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2} < 0 \tag{3}$$

and into π_2 otherwise. Here, $-\mathrm{tr}(\boldsymbol{S}_{1n_1})/(2n_1) + \mathrm{tr}(\boldsymbol{S}_{2n_2})/(2n_2)$ is a bias-correction term. On the other hand, Chan and Hall (2009) considered a scale-adjusted distance-based classifier as follows: One classifies an individual into π_1 if

$$\sum_{j=1}^{n_1} \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_{1j}||^2}{n_1} - \sum_{j=1}^{n_2} \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_{2j}||^2}{n_2} - \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \frac{||\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}||^2}{2n_1(n_1-1)}$$
$$+ \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} \frac{||\boldsymbol{x}_{2i} - \boldsymbol{x}_{2j}||^2}{2n_2(n_2-1)} < 0 \tag{4}$$

and into π_2 otherwise. We note the classifier given by (3) is equivalent to (4).

2.1 Consistency of the classifier

We denote the error of misclassifying an individual from π_1 (into π_2) or π_2 (into π_1) by $e(2|1)$ or $e(1|2)$, respectively. Let $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ and

$$w(\boldsymbol{x}_0|n_1, n_2) = \left(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\right)^{\mathrm{T}} (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2}.$$

We consider asymptotic properties of $w(\boldsymbol{x}_0|n_1, n_2)$ under the following assumptions:

(A-iii) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta^2} \to 0$ as $p \to \infty$ for $i = 1, 2$;

(A-iv) $\dfrac{\max_{j=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{n_i \Delta^2} \to 0$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$.

Theorem 1 *Assume* (A-iii) *and* (A-iv). *Then, we have as* $p \to \infty$ *that*

$$\frac{w(\boldsymbol{x}_0|n_1, n_2)}{\Delta} = \frac{(-1)^i}{2} + o_p(1) \quad \textit{when}\, \boldsymbol{x}_0 \in \pi_i$$

for $i = 1, 2$. *For the classifier given by* (3), *we have as* $p \to \infty$ *that*

$$e(2|1) \to 0 \quad \textit{and} \quad e(1|2) \to 0. \tag{5}$$

Remark 2 If one can assume that $\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)\}/\Delta^2 \to 0$ as $p \to \infty$, it follows naturally that (A-iii) and (A-iv) hold from the fact that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \le \Delta\lambda_{i1} \le \Delta \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. Then, one can claim Theorem 1 even when n_i is fixed for $i = 1, 2$.

Remark 3 Chan and Hall (2009) gave (5) for a different distance-based classifier under different assumptions.

Here, we consider a quadratic classifier given by Aoshima and Yata (2011a). By substituting $\{\mathrm{tr}(\boldsymbol{S}_{in_i})/p\}\boldsymbol{I}_p$ for $\boldsymbol{S}_{in_i}$ in (2), they proposed the following classification rule: One classifies an individual into π_1 if

$$\frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - p\log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\} - \frac{p}{n_1} + \frac{p}{n_2} < 0 \tag{6}$$

and into π_2 otherwise. Here, $-p/n_1 + p/n_2$ is a bias correction term. Let $\Delta_\star = \Delta + \mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2/\{2\max_{i=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_i)\}$. Then, from the results given by Aoshima and Yata (2011a), we have the following theorem.

Theorem 2 *Assume that* $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \in (0, \infty)$ *as* $p \to \infty$ *and* $\limsup_{p\to\infty} \{\Delta/\mathrm{tr}(\boldsymbol{\Sigma}_i)\} < \infty$ *for* $i = 1, 2$. *Assume also that*

(AY-i) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2\Delta_\star^2} \to 0$ *as* $p \to \infty$ *for* $i = 1, 2$;

(AY-ii) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta_\star^2} \to 0$ *as* $p \to \infty$ *for* $i = 1, 2$;

(AY-iii) $\dfrac{\max_{j=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{n_i\Delta_\star^2} \to 0$ *as* $p \to \infty$ *either when* n_i *is fixed or* $n_i \to \infty$ *for* $i = 1, 2$.

For the classifier given by (6), we have (5) as $p \to \infty$ *under (1) with (A-i).*

Remark 4 If we can assume that $\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)\}/\Delta_\star^2 \to 0$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \in (0, \infty)$ as $p \to \infty$, it follows that (AY-i) to (AY-iii) hold even when n_i is fixed for $i = 1, 2$. It should be noted that if one can assume that $\liminf_{p\to\infty} |\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2 \to 0$ as $p \to \infty$ for $i = 1, 2$, we can claim that $\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)\}/\Delta_\star^2 \to 0$ as $p \to \infty$ even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (i.e., $\Delta = 0$).

The conditions (AY-ii) and (AY-iii) are milder than (A-iii) and (A-iv) from the fact that $\Delta_\star \geq \Delta$. On the other hand, the classifier given by (3) holds (5) without assuming (1) with (A-i). Hence, we recommend that the experimenter should use the classifier given by (6) if $\Delta_\star$ is sufficiently larger than Δ when (1) with (A-i) is assumed. Otherwise, i.e. if $\Delta_\star$ is not sufficiently larger than Δ or if (1) with (A-i) (or (AY-i)) is not assumed, the experimenter is recommended to use the classifier given by (3).

Remark 5 Let $\widehat{\Delta} = ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i$ and $\widehat{\Delta}_\star = \widehat{\Delta} + \mathrm{tr}(\boldsymbol{S}_{1n_1} - \boldsymbol{S}_{2n_2})^2/\{2\max_{i=1,2}\mathrm{tr}(\boldsymbol{S}_{in_i})\}$. Note that $E_{\boldsymbol{\theta}}(\widehat{\Delta}) = \Delta$. Under (A-iv), it holds that $\widehat{\Delta}/\Delta = 1 + o_p(1)$. Under the assumptions of Theorem 2, it holds that $\widehat{\Delta}_\star/\Delta_\star = 1 + o_p(1)$. Thus by using $\widehat{\Delta}$ and $\widehat{\Delta}_\star$, one can check whether $\Delta_\star$ is sufficiently larger than Δ or not.

2.2 Simulation

We used computer simulations to compare the performance of the classifiers. We generated $\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i,\ j = 1, 2, \ldots,\ (i = 1, 2)$ independently from a pseudorandom p-variate t-distribution, $t_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i, \nu)$ with mean zero, covariance matrix $\boldsymbol{\Sigma}_i$ and degrees of freedom ν. Note that $t_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i, \nu)$ converges to $N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ as $\nu \to \infty$. We set $\boldsymbol{\mu}_2 = \boldsymbol{0},\ \boldsymbol{\Sigma}_1 = c_1\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = c_2\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, where

$$\boldsymbol{B} = \mathrm{diag}[\{0.5 + 1/(p+1)\}^{1/2}, \ldots, \{0.5 + p/(p+1)\}^{1/2}]. \qquad (7)$$

We considered two cases for $\boldsymbol{\mu}_1$: (a) $\boldsymbol{\mu}_1 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ whose first $\lceil p^{2/3} \rceil$ elements are 1, and (b) $\boldsymbol{\mu}_1 = (0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}}$ whose last $\lceil p^{2/3} \rceil$ elements are 1. Here, $\lceil x \rceil$ denotes the smallest integer $\geq x$. Note that $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 = \lceil p^{2/3} \rceil$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = O(p),\ i = 1, 2$. One can check that (A-iii) and (A-iv) are met even for fixed n_is from the fact that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \Delta\lambda_{i1} \leq \Delta\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. We considered three cases: (I) $p = 2^s,\ s = 5, \ldots, 10,\ (n_1, n_2) = (10, 10),\ (c_1, c_2) = (1, 1)$ and $\nu = 25$ for (a) and (b); (II) $p = 2^s,\ s = 5, \ldots, 10,\ (n_1, n_2) = (10, 20),\ (c_1, c_2) = (0.8, 1.2)$ and $\nu = 25$ for (b); and (III) $p = 500,\ (n_1, n_2) = (10, 20),\ (c_1, c_2) = (0.8, 1.2)$ and $\nu = 10(10)60$ for (b). Note that $\Delta_\star = \Delta$ in case (I) and $\Delta_\star = \Delta + p/15$ in cases (II) and (III). We compared the classifiers, (3), (6) and the following three classifiers: (i) DLDA given by Dudoit et al. (2002) and Bickel and Levina (2004); (ii) DQDA given by Dudoit et al. (2002); and (iii) the hard-margin linear support vector machine (HM-LSVM) given by Vapnic (1999). The rule of diagonal linear discriminant analysis (DLDA) is given for $\boldsymbol{x}_0 \in \pi_1$ (or π_2) by

$$\{\boldsymbol{x}_0 - (\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2})/2\}^{\mathrm{T}}\boldsymbol{S}_d^{-1}(\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) < 0 \quad (\text{or} \geq 0),$$

where $\boldsymbol{S}_d = \mathrm{diag}(s_{1n}, \ldots, s_{pn}),\ s_{jn} = \sum_{i=1}^{2}\sum_{l=1}^{n_i}(x_{ijl} - \overline{x}_{ijn_i})^2/(n_1 + n_2 - 2)$ and $\overline{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl}/n_i$. The rule of diagonal quadratic discriminant analysis (DQDA) is given for $\boldsymbol{x}_0 \in \pi_1$ (or π_2) by

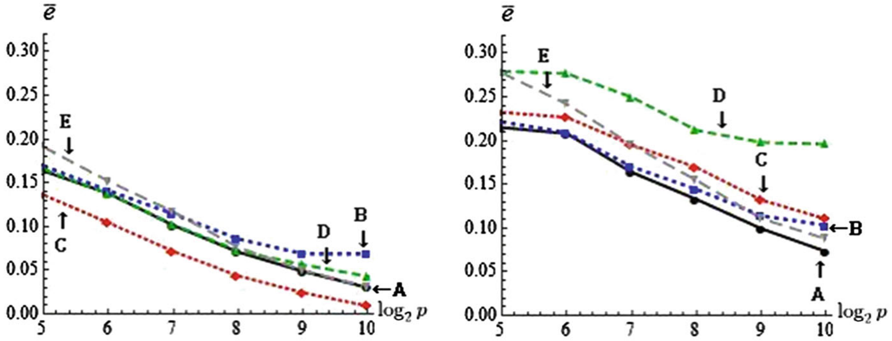

Fig. 1 The *left panel* displays $\overline{e}$ in case of (**a**) $\boldsymbol{\mu}_1 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ and the *right panel* displays $\overline{e}$ in case of (**b**) $\boldsymbol{\mu}_1 = (0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}}$. In each diagram, *A*, *B*, *C*, *D* and *E* denote (3), (6), DLDA, DQDA and HM-LSVM, respectively, for $\nu = 25$, $(c_1, c_2) = (1, 1)$, $(n_1, n_2) = (10, 10)$ and $p = 2^s$ $(s = 5, \ldots, 10)$

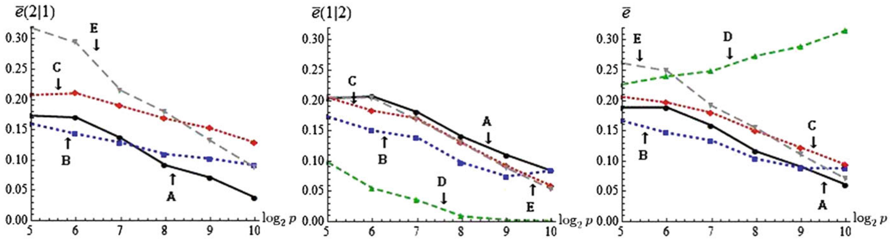

Fig. 2 In case of (**b**) $\boldsymbol{\mu}_1 = (0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}}$, the *left panel* displays $\overline{e}(2|1)$, the *middle panel* displays $\overline{e}(1|2)$ and the *right panel* displays $\overline{e}$. In each diagram, *A*, *B*, *C*, *D* and *E* denote (3), (6), DLDA, DQDA and HM-LSVM, respectively, for $\nu = 25$, $(c_1, c_2) = (0.8, 1.2)$, $(n_1, n_2) = (10, 20)$ and $p = 2^s$ $(s = 5, \ldots, 10)$. In the *left panel*, $\overline{e}(2|1)$ is not described for *D* because the rate was too high

$$
\begin{aligned}
&(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^{\mathrm{T}} \boldsymbol{S}_{d(1)}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^{\mathrm{T}} \boldsymbol{S}_{d(2)}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}) \\
&- \log \left\{ \frac{\det(\boldsymbol{S}_{d(2)})}{\det(\boldsymbol{S}_{d(1)})} \right\} < 0 \quad (\text{or} \geq 0),
\end{aligned}
$$

where $\boldsymbol{S}_{d(i)} = \operatorname{diag}(s_{(i)1n_i}, \ldots, s_{(i)pn_i})$ and $s_{(i)jn_i} = \sum_{l=1}^{n_i} (x_{ijl} - \overline{x}_{ijn_i})^2/(n_i - 1)$. Note that the HDLSS data ($p > n_1 + n_2$) are linearly separable by a hyperplane. Thus we used the hard-margin support vector machine. We checked 2000 times for $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$ about whether each rule does (or does not) classify $\boldsymbol{x}_0$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i. We calculated $\overline{e}(2|1) = \sum_{r=1}^{2000} P_{1r}/2000$ and $\overline{e}(1|2) = \sum_{r=1}^{2000} P_{2r}/2000$ as estimates of $e(2|1)$ and $e(1|2)$. Note that the standard deviation of the estimates are less than 0.011. Also, we calculated an error rate, $\overline{e} = \{\overline{e}(2|1) + \overline{e}(1|2)\}/2$.

In Fig. 1, we plotted only $\overline{e}$ for case (I) since $\overline{e}(1|2)$ and $\overline{e}(2|1)$ were of almost the same rate under the settings as $n_1 = n_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. In Figs. 2 and 3, we plotted $\overline{e}(2|1)$, $\overline{e}(1|2)$ and $\overline{e}$ for cases (II) and (III), respectively. In each diagram, A, B, C, D and E denote (3), (6), DLDA, DQDA, and HM-LSVM, respectively. We observed that (3) gives a preferable performance for both $e(2|1)$ and $e(1|2)$ in those cases.

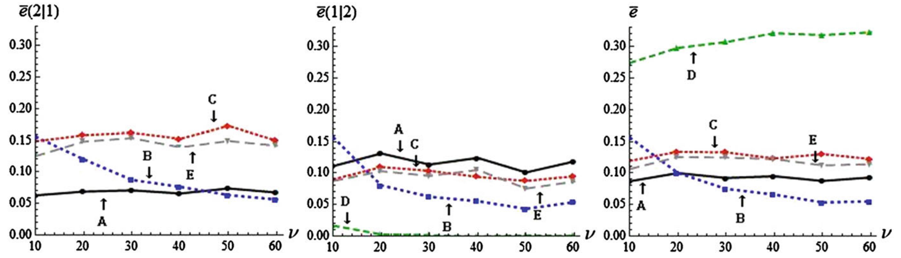

Fig. 3 In case of (**b**) $\boldsymbol{\mu}_1 = (0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}}$, the *left panel* displays $\bar{e}(2|1)$, the *middle panel* displays $\bar{e}(1|2)$ and the *right panel* displays $\bar{e}$. In each diagram, *A*, *B*, *C*, *D* and *E* denote (3), (6), DLDA, DQDA and HM-LSVM, respectively, for $p = 500$, $(c_1, c_2) = (0.8, 1.2)$, $(n_1, n_2) = (10, 20)$ and $\nu = 10(10)60$. In the *left panel*, $\bar{e}(2|1)$ is not described for D because the rate was too high

In Fig. 1, (6) gave a little worse performance compared to (3). On the other hand, (6) gave a preferable performance for the unbalanced cases as observed in Figs. 2 and 3. In Fig. 3, (6) gave a better performance compared to (3) when ν is large and it gave a bad performance when ν is not large enough for π_is to satisfy (A-i). However, (3) seems to perform well even when $\nu = 10$.

We observed that DLDA gives a better performance compared to (3) for case (a) in Fig. 1. However, DLDA gave a worse performance for case (b) even in Fig. 1. This is probably due to the distance between the two populations. As for (3), $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 = \lceil p^{2/3} \rceil$ is considered as the distance for both the cases, (a) and (b). Let $\boldsymbol{\Sigma}_d = \mathrm{diag}(\sigma_{1(d)}, \ldots, \sigma_{p(d)})$ having $\sigma_{j(d)} = E_{\boldsymbol{\theta}}(s_{jn})$. As for DLDA, $\Delta_d = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}_d^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is considered as the distance. Note that $E_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0)_{DL} | \boldsymbol{x}_0 \in \pi_2\} - E_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0)_{DL} | \boldsymbol{x}_0 \in \pi_1\} = \Delta_d$, where $w(\boldsymbol{x}_0)_{DL} = \{\boldsymbol{x}_0 - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\}^{\mathrm{T}} \boldsymbol{\Sigma}_d^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Then, $\Delta_d = \sum_{j=1}^{\lceil p^{2/3} \rceil} \sigma_{j(d)}^{-1} > \Delta$ for (a) and $\Delta_d = \sum_{j=p+1-\lceil p^{2/3} \rceil}^{p} \sigma_{j(d)}^{-1} < \Delta$ for (b). This is probably the reason about (3) vs. DLDA.

We observed from Figs. 2 and 3 that DLDA and HM-LSVM give an unbalanced performance between $e(2|1)$ and $e(1|2)$ and DQDA leads an undesirable performance for $e(2|1)$ as p increases. This is probably due to the bias of those classifiers when p is large. For example, let us denote $\boldsymbol{\Sigma}_{d(i)} = \mathrm{diag}(\sigma_{(i)1}, \ldots, \sigma_{(i)p})$, $i = 1, 2$, having $\sigma_{(i)j} = E_{\boldsymbol{\theta}}(s_{(i)jn_i})$, and assume $\sigma_{(i)j}$s are known. Then, the classifier for DQDA is given by $w(\boldsymbol{x}_0)_{DQ} = (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^{\mathrm{T}} \boldsymbol{\Sigma}_{d(1)}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^{\mathrm{T}} \boldsymbol{\Sigma}_{d(2)}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}) - \log\{\det(\boldsymbol{\Sigma}_{d(2)})\} + \log\{\det(\boldsymbol{\Sigma}_{d(1)})\}$. It holds that $E_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0)_{DQ}\} = p + p/n_1 - p/n_2 - \mathrm{tr}(\boldsymbol{\Sigma}_{d(1)} \boldsymbol{\Sigma}_{d(2)}^{-1}) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}_{d(2)}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \log\{\det(\boldsymbol{\Sigma}_{d(2)})\} + \log\{\det(\boldsymbol{\Sigma}_{d(1)})\}$ when $\boldsymbol{x}_0 \in \pi_1$. Hence, the bias of $w(\boldsymbol{x}_0)_{DQ}$ is $p/n_1 - p/n_2$ which becomes formidably large as p increases. Furthermore, if $\sigma_{(i)j}$s are unknown, DQDA would cause extra bias for the estimation. Recently, Huang et al. (2010) gave bias corrected DLDA and DQDA. However, they gave a bias correction only when the population is Gaussian. As for HM-LSVM, Chan and Hall (2009) gave a scale-adjusted SVM. On the other hand, the classifier by (3) is unbiased even when the population is non-Gaussian and it always holds that $E_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0 | n_1, n_2)\} = (-1)^i \Delta/2$ for $\boldsymbol{x}_0 \in \pi_i$. This is probably the main reason why the classifier by (3) gives a preferable performance for both $e(2|1)$ and $e(1|2)$ as p increases. See Chan and Hall (2009) for numerical comparisons among the classifiers by (3) (or (4)) and other distance-based classifiers including the scale-adjusted SVM.

2.3 Asymptotic normality of the classifier

Let

$$\delta_i = \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_{i'}} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}{2n_1(n_1-1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}{2n_2(n_2-1)} \right\}^{1/2} \tag{8}$$

for $i(\neq i') = 1, 2$. We assume an extra assumption:

(A-v) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\delta_i^2} \to 0$ as $p \to \infty$ and $n_j \to \infty,\ j = 1, 2$, for $i = 1, 2$.

Under (A-v), note that $\mathrm{Var}_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0|n_1, n_2)\}/\delta_i^2 = 1 + o(1)$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. We have the following theorem.

Theorem 3 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and (*A-v*). Then, we have as* $p \to \infty$ *and* $n_i \to \infty,\ i = 1, 2$, *that*

$$\frac{w(\boldsymbol{x}_0|n_1, n_2) - (-1)^i\Delta/2}{\delta_i} \Rightarrow N(0, 1) \quad \textit{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \textit{for} \quad i = 1, 2, \tag{9}$$

where "⇒" denotes the convergence in distribution and $N(0, 1)$ *denotes a random variable distributed as the standard normal distribution.*

Remark 6 From Theorem 3, for the classifier given by (3), we have as $p \to \infty$ and $n_i \to \infty,\ i = 1, 2$, that

$$e(2|1) = \Phi\left(\frac{-\Delta}{2\delta_1}\right) + o(1) \quad \text{and} \quad e(1|2) = \Phi\left(\frac{-\Delta}{2\delta_2}\right) + o(1) \tag{10}$$

under (A-ii), (A-v) and (1) with (A-i), where $\Phi(\cdot)$ denotes the cumulative distribution function of a $N(0, 1)$ random variable.

Remark 7 Yata and Aoshima (2012b) showed asymptotic normality for (3) under different conditions such as n_is are fixed. Chan and Hall (2009) showed asymptotic normality for the distance-based classifier given by (4) under different assumptions.

Let us consider an example such as $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\ i = 1, 2$, having $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ whose first $\lceil \mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2} \rceil$ elements are 1. Here, $\boldsymbol{\Sigma}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B}$ is defined by (7). We set $n_1 = \log_2 p$ and $n_2 = 2\log_2 p$. Note that $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 = \lceil \mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2} \rceil = \mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2}\{1+o(1)\}$, $\Delta/\delta_1 = (1/n_1 + 1.2/n_2)^{-1/2}\{1+o(1)\}$, $\Delta/\delta_2 = (1.2/n_1 + 1.44/n_2)^{-1/2}\{1 + o(1)\}$ and $\delta_2^2/\delta_1^2 = 1.2\{1 + o(1)\}$ as $p \to \infty$. We considered the cases of $p = 2^s,\ s = 3, \ldots, 14$. Independent pseudorandom 2000 observations of $w(\boldsymbol{x}_0|n_1, n_2)$ were generated when $\boldsymbol{x}_0 \in \pi_1$ or π_2. In the end of the rth replication, we checked whether the rule (3) does (or does not) classify

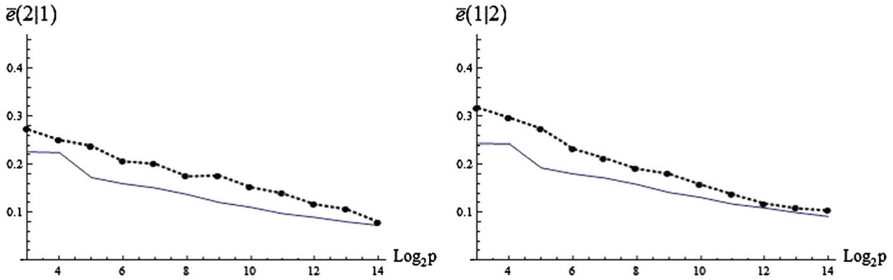

Fig. 4 When $(n_1, n_2) = (\log_2 p, 2\log_2 p)$ and $p = 2^s$ $(s = 3, \ldots, 14)$, the *left panel* displays $\overline{e}(2|1)$ (*dashed line*) and $\Phi\{-\Delta/(2\delta_1)\}$ (*solid line*) and the *right panel* displays $\overline{e}(1|2)$ (*dashed line*) and $\Phi\{-\Delta/(2\delta_2)\}$ (*solid line*)

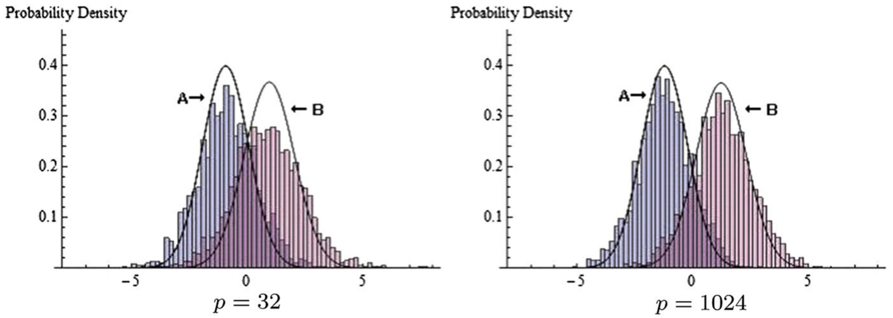

Fig. 5 The histograms of $w(\boldsymbol{x}_0|n_1, n_2)/\delta_1$ for $\boldsymbol{x}_0 \in \pi_1$ or π_2 together with the probability densities of A: $N\{-\Delta/(2\delta_1), 1\}$ and B: $N\{\Delta/(2\delta_1), \delta_2^2/\delta_1^2\}$ when $p = 32$ and $p = 1024$

$\boldsymbol{x}_0$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i. We calculated $\overline{e}(2|1) = \sum_{r=1}^{2000} P_{1r}/2000$ and $\overline{e}(1|2) = \sum_{r=1}^{2000} P_{2r}/2000$ as estimates of $e(2|1)$ and $e(1|2)$. Note that the standard deviation of the estimates is less than 0.011. From Remark 6, we also calculated $\Phi\{-\Delta/(2\delta_1)\}$ and $\Phi\{-\Delta/(2\delta_2)\}$. In Fig. 4, we plotted $\overline{e}(2|1)$ and $\overline{e}(1|2)$ together with $\Phi\{-\Delta/(2\delta_i)\}$, $i = 1, 2$. As expected theoretically, we observed that the plots become close to $\Phi\{-\Delta/(2\delta_i)\}$ as p increases. In Fig. 5, we gave two histograms of $w(\boldsymbol{x}_0|n_1, n_2)/\delta_1$ for $\boldsymbol{x}_0 \in \pi_1$ or π_2 when $p = 32$ $(= 2^5)$ and $p = 1024$ $(= 2^{10})$. From Theorem 3, we also displayed the asymptotic probability densities of $w(\boldsymbol{x}_0|n_1, n_2)/\delta_1$, $N\{-\Delta/(2\delta_1), 1\}$ when $\boldsymbol{x}_0 \in \pi_1$ and $N\{\Delta/(2\delta_1), \delta_2^2/\delta_1^2\}$ when $\boldsymbol{x}_0 \in \pi_2$. When $p = 32$, the histogram appears different from the probability density. However, when $p = 1024$, it becomes close to the probability density.

3 Misclassification rate adjusted classifier for two-class classification

In this section, we develop a scale-adjusted distance-based classifier that ensures misclassification rates are no more than a prespecified value. The advantage of the classifier is quite robust and applicable to multiclass classification.

3.1 Sample size determination to control misclassification rates

We are interested in designing a classifier that ensures both $e(2|1) \le \alpha$ and $e(1|2) \le \beta$ when $\Delta \ge \Delta_L$ for prespecified constants, $\alpha,\ \beta \in (0, 1/2)$ and $\Delta_L\ (>0)$. We assume $\Delta_L = o\{\min_{i=1,2} \operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}\}$. We adjust the classification rule in (3) by using some tuning parameter, γ, as follows: one classifies an individual into π_1 if

$$w(\boldsymbol{x}_0|n_1, n_2) < \gamma \tag{11}$$

and into π_2 otherwise. Let z_α be a constant such that $P\{N(0,1) \ge z_\alpha\} = \alpha$. We consider n_is satisfying $\delta_i \le \Delta_L/(z_\alpha + z_\beta)$ for $i = 1, 2$, where δ_i is defined by (8). From the fact that $\operatorname{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \le \{\operatorname{tr}(\boldsymbol{\Sigma}_1^2)\operatorname{tr}(\boldsymbol{\Sigma}_2^2)\}^{1/2}$, it holds for $i' \ne i$ that

$$\delta_i^2 \le \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)}{n_i - 1} + \frac{\operatorname{tr}(\boldsymbol{\Sigma}_{i'}^2)^{1/2} \max_{j=1,2} \operatorname{tr}(\boldsymbol{\Sigma}_j^2)^{1/2}}{n_{i'} - 1} \le \max_{j=1,2} \operatorname{tr}(\boldsymbol{\Sigma}_j^2)^{1/2} \sum_{i=1}^{2} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}}{n_i - 1}.$$

Let $\sigma = \max_{i=1,2} \operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. Then, we find the sample size for each π_i as

$$n_i \ge \frac{(z_\alpha + z_\beta)^2 \sigma}{\Delta_L^2} \operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \sum_{j=1}^{2} \operatorname{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} + 1 \quad (= C_i,\ \text{say}). \tag{12}$$

Note that $C_i/p \to 0,\ i = 1, 2$, as $p \to \infty$ under the condition that $\max_{i=1,2} \{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)\}/\Delta_L^2 = o(p)$. We also note that $n_i \to \infty,\ i = 1, 2$, as $p \to \infty$ from the fact that $\Delta_L = o\{\min_{i=1,2} \operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}\}$. Then, we have the following theorem.

Theorem 4 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-iii). *Let* $\gamma = \Delta_L(z_\alpha - z_\beta)/\{2(z_\alpha + z_\beta)\}$ *in* (11). *Then, for the classification rule given by* (11) *with* (12), *it holds as* $p \to \infty$ *that*

$$\limsup e(2|1) \le \alpha \quad \textit{and} \quad \limsup e(1|2) \le \beta \quad \textit{when } \Delta \ge \Delta_L.$$

Remark 8 One can design Δ_L by using the two sample test given by Aoshima and Yata (2011a,b) or Chen and Qin (2010). Under the regularity conditions, it holds as $p \to \infty$ and $n_i \to \infty,\ i = 1, 2$, that

$$\frac{\widehat{\Delta} - \Delta}{\kappa} \Rightarrow N(0, 1),$$

where $\widehat{\Delta}$ is defined in Remark 5 and

$$\kappa = \left\{ 2\sum_{i=1}^{2} \frac{W_{in_i}}{n_i(n_i - 1)} + 4\frac{\operatorname{tr}(\boldsymbol{S}_{1n_1}\boldsymbol{S}_{2n_2})}{n_1 n_2} \right\}^{1/2}$$

having W_{in_i} defined by (16). It follows that $P_{\boldsymbol{\theta}}(\widehat{\Delta}/\kappa - z_{\alpha'} \le \Delta/\kappa) \to 1-\alpha'$ for given $\alpha' \in (0, 1/2)$. Thus, one may design a lower bound of Δ by

$$\Delta_L = \widehat{\Delta} - \kappa z_{\alpha'}$$

for sufficiently small α'. Then, it holds that $\Delta_L/\Delta = 1 + o_p(1)$ under (A-iii) and (1) with (A-i) when $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/(n_i^2\Delta^2) \to 0,\ i = 1, 2$.

3.2 Two-stage procedure

Since $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unknown, it is necessary to estimate C_i in (12) with some pilot samples. We proceed with the following two steps:

1. Choose $m_i (\ge 4)$ satisfying

$$\frac{m_i}{C_i} \le 1, \quad \frac{C_i}{m_i^2} \to 0 \ \text{ and } \ \frac{C_i}{m_i}\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^4)}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^2} \to 0 \ \text{ as } \ p \to \infty \ \text{ under (A-ii)} \tag{13}$$

for $i = 1, 2$. Note that m_i holds (13) when it follows that $m_i/C_i \in (0, 1)$ as $p \to \infty$. Take pilot samples, $\boldsymbol{x}_{ij},\ j = 1, \ldots, m_i$, of size m_i from each π_i. Then, calculate W_{im_i} for each π_i according to (16). Let $\widehat{\sigma} = \max_{i=1,2} W_{im_i}^{1/2}$. We recall that $\lceil x \rceil$ denotes the smallest integer $\ge x$. Define the total sample size for each π_i by

$$N_i = \max\left\{ m_i,\ \left\lceil \frac{(z_\alpha + z_\beta)^2 \widehat{\sigma}}{\Delta_L^2} W_{im_i}^{1/4} \sum_{j=1}^{2} W_{jm_j}^{1/4} \right\rceil + 1 \right\}. \tag{14}$$

2. For each i, if $N_i = m_i$, do not take any additional samples from π_i and otherwise, that is if $N_i > m_i$, take additional samples, $\boldsymbol{x}_{ij},\ j = m_i + 1, \ldots, N_i$, of size $N_i - m_i$ from π_i. By combining the initial samples and the additional samples, calculate $\overline{\boldsymbol{x}}_{iN_i}$ and $\boldsymbol{S}_{iN_i},\ i = 1, 2$. Then, we classify an individual into π_1 if

$$w(\boldsymbol{x}_0 | N_1, N_2) < \gamma, \tag{15}$$

and into π_2 otherwise, where $\gamma = \Delta_L(z_\alpha - z_\beta)/\{2(z_\alpha + z_\beta)\}$.

We have the following theorem.

Theorem 5 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-iii). *Then, for the classification rule given by* (15) *with* (13) *and* (14), *it holds as* $p \to \infty$ *that*

$$\limsup e(2|1) \le \alpha \quad \textit{and} \quad \limsup e(1|2) \le \beta \quad \textit{when } \Delta \ge \Delta_L.$$

Remark 9 If $\Delta_\star$ is sufficiently larger than Δ, we recommend to use the two-stage classification procedure based on (6) that was developed by Aoshima and Yata (2011a).

Remark 10 Under (A-ii), (13) and (1) with (A-i), it holds as $p \to \infty$ that $\mathrm{Var}_{\boldsymbol{\theta}}\{W_{im_i}/\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\} = o(C_i^{-1})$. Then, we claim as $p \to \infty$ that $N_i/C_i = 1 + o_p(1)$, which is in the HDLSS situation in the sense that $N_i/p = o_p(1)$ under the condition that $\max_{i=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}/\Delta_L^2 = o(p)$.

Remark 11 Under (1) with (A-i), it holds as $m_i \to \infty$, $i = 1, 2$, that $N_i / \max\{C_i, m_i\} = 1 + o_p(1)$. Hence, by comparing m_i with N_i, one may check whether (13) holds or not from the fact that m_i holds (13) when $m_i/C_i \in (0, 1)$ as $p \to \infty$.

Remark 12 One may choose $m_i(\geq 4)$ such as satisfies $m_i/C_i > 1$ for some i. Then, the assertion in Theorem 5 is still claimed. However, it may cause over-sampling in the sense that $N_i/C_i > 1$ w.p.1.

3.3 Simulation

In order to examine the performance of the classifier given by (15) with (13) and (14), we used computer simulations. Independent pseudo random observations were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We considered $\boldsymbol{\Sigma}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B}$ is defined by (7). We set $\boldsymbol{\mu}_1 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ whose first 30 elements are 1 and $\boldsymbol{\mu}_2 = (0, \ldots, 0)^{\mathrm{T}}$, so that $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 = 30$. We prespecified $\Delta_L = 30$. We set $m_i = \lceil 0.5 \times (C_i - 1)\rceil + 1$, $i = 1, 2$, where C_i is defined by (12). We considered four cases: (a) $p = 400$ when $(\alpha, \beta) = (0.1, 0.1)$, (b) $p = 1200$ when $(\alpha, \beta) = (0.1, 0.1)$, (c) $p = 400$ when $(\alpha, \beta) = (0.05, 0.15)$, and (d) $p = 1200$ when $(\alpha, \beta) = (0.05, 0.15)$.

By averaging the outcomes from 2000 ($= R$, say) replications, the findings were summarized in Table 1. Under a fixed scenario, suppose that the rth replication ends with $N_i = n_{ir}$ $(i = 1, 2)$ observations for $r = 1, \ldots, R$. Let $\overline{n}_i = R^{-1}\sum_{r=1}^{R} n_{ir}$ and

Table 1 Accuracy of the classifier given by (15) with (13) and (14)

	C_i	$\overline{n}_i$	$\overline{n}_i - C_i$	$V(n_i)$	$\overline{e}(j\|i)$	$s\{\overline{e}(j\|i)\}$
When $(\alpha, \beta) = (0.1, 0.1)$						
$p = 400$: $(m_1, m_2) = (8, 9)$						
π_1	14.43	14.85	0.42	16.12	0.085	0.00624
π_2	16.1	16.74	0.64	29.85	0.102	0.00675
$p = 1200$: $(m_1, m_2) = (22, 24)$						
π_1	41.72	41.97	0.26	14.89	0.084	0.00619
π_2	46.92	47.16	0.24	30.31	0.099	0.00668
When $(\alpha, \beta) = (0.05, 0.15)$						
$p = 400$: $(m_1, m_2) = (9, 10)$						
π_1	15.69	15.95	0.25	14.28	0.044	0.00456
π_2	17.53	17.99	0.46	26.92	0.152	0.00802
$p = 1200$: $(m_1, m_2) = (24, 27)$						
π_1	45.56	46.0	0.45	16.11	0.033	0.00397
π_2	51.25	51.77	0.52	33.23	0.133	0.00758

$V(n_i) = (R-1)^{-1}\sum_{r=1}^{R}(n_{ir} - \overline{n}_i)^2$. In the end of the rth replication, we checked whether the classifier does (or does not) classify $\boldsymbol{x}_0$ from π_i correctly and defined $P_{ir} = 0$ (or 1) accordingly for each i. We calculated $\overline{e}(2|1) = R^{-1}\sum_{r=1}^{R} P_{1r}$ and $\overline{e}(1|2) = R^{-1}\sum_{r=1}^{R} P_{2r}$ as estimates of $e(2|1)$ and $e(1|2)$. Their estimated standard errors were given by $s\{\overline{e}(j|i)\}$ for $i \neq j$, where $s^2\{\overline{e}(j|i)\} = R^{-1}\overline{e}(j|i)\{1 - \overline{e}(j|i)\}$. Throughout, the classifier given by (15) with (13) and (14) gave adequate performances for all the cases when considered those standard errors.

3.4 Estimation of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$

Throughout this section, we omit the subscript with regard to the class. Yata and Aoshima (2013) gave a method called the extended cross-data-matrix (ECDM) methodology that is an extension of the CDM methodology developed by Yata and Aoshima (2010). The ECDM methodology can be applied to obtain an unbiased estimator of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ as follows: we assume $n \geq 4$. Let $n_{(1)} = \lceil n/2 \rceil$ and $n_{(2)} = n - n_{(1)}$. Let

$$\boldsymbol{V}_{n(1)(k)} = \begin{cases} \{\lfloor k/2 \rfloor - n_{(1)} + 1, \ldots, \lfloor k/2 \rfloor\} & \text{if } \lfloor k/2 \rfloor \geq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor\} \cup \{\lfloor k/2 \rfloor + n_{(2)} + 1, \ldots, n\} & \text{otherwise;} \end{cases}$$

$$\boldsymbol{V}_{n(2)(k)} = \begin{cases} \{\lfloor k/2 \rfloor + 1, \ldots, \lfloor k/2 \rfloor + n_{(2)}\} & \text{if } \lfloor k/2 \rfloor \leq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor - n_{(1)}\} \cup \{\lfloor k/2 \rfloor + 1, \ldots, n\} & \text{otherwise} \end{cases}$$

for $k = 3, \ldots, 2n-1$, where $\lfloor x \rfloor$ denotes the largest integer $\leq x$. Let $\#(\boldsymbol{S})$ denote the number of elements in a set $\boldsymbol{S}$. Note that $\#(\boldsymbol{V}_{n(l)(k)}) = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{n(1)(k)} \cap \boldsymbol{V}_{n(2)(k)} = \boldsymbol{\emptyset}$ and $\boldsymbol{V}_{n(1)(k)} \cup \boldsymbol{V}_{n(2)(k)} = \{1, \ldots, n\}$ for $k = 3, \ldots, 2n-1$. Also, note that $i \in \boldsymbol{V}_{n(1)(i+j)}$ and $j \in \boldsymbol{V}_{n(2)(i+j)}$ for $i < j$ $(\leq n)$. Let

$$\overline{\boldsymbol{x}}_{n(1)(k)} = n_{(1)}^{-1} \sum_{j \in \boldsymbol{V}_{n(1)(k)}} \boldsymbol{x}_j \quad \text{and} \quad \overline{\boldsymbol{x}}_{n(2)(k)} = n_{(2)}^{-1} \sum_{j \in \boldsymbol{V}_{n(2)(k)}} \boldsymbol{x}_j$$

for $k = 3, \ldots, 2n-1$. Then, Yata and Aoshima (2013) gave an estimator of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ by

$$W_n = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} \left\{(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{n(1)(i+j)})^{\mathrm{T}}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n(2)(i+j)})\right\}^2, \tag{16}$$

where $u_n = n_{(1)}n_{(2)}/\{(n_{(1)} - 1)(n_{(2)} - 1)\}$. Note that $E_{\boldsymbol{\theta}}(W_n) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$. We have the following theorem.

Theorem 6 *Assume* (1) *with* (A-i). *Then, it holds that*

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(\frac{W_n}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right) = \frac{4}{n^2}\{1 + o(1)\} + O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n}\right\}$$

as $n \to \infty$ either when $p \to \infty$ or p is fixed.

Further, if $\boldsymbol{x}_j$ is Gaussian, it holds that $\mathrm{Var}_{\boldsymbol{\theta}}\{W_n/\mathrm{tr}(\boldsymbol{\Sigma}^2)\} = 4\{1+o(1)\}/n^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}^4)\{1+o(1)\}/\{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n\}$ as $n \to \infty$ either when $p \to \infty$ or p is fixed. Bai and Saranadasa (1996) and Srivastava (2005) considered an estimator of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ by $V_n = c_n^{-1}\{\mathrm{tr}(\boldsymbol{S}_n^2) - \mathrm{tr}(\boldsymbol{S}_n)^2/(n-1)\}$ with $c_n = (n-2)(n+1)/(n-1)^2$. They showed that, when $\boldsymbol{x}_j$ is Gaussian, it holds that $E_{\boldsymbol{\theta}}(V_n) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$ and $\mathrm{Var}_{\boldsymbol{\theta}}\{V_n/\mathrm{tr}(\boldsymbol{\Sigma}^2)\} = 4\{1+o(1)\}/n^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}^4)\{1+o(1)\}/\{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n\}$. It should be noted that V_n is not an unbiased estimator unless $\boldsymbol{x}_j$s are Gaussian. In addition, one cannot claim $\mathrm{Var}_{\boldsymbol{\theta}}\{V_n/\mathrm{tr}(\boldsymbol{\Sigma}^2)\} < \infty$ unless the eighth moments of each variable in $\boldsymbol{y}_j$ are uniformly bounded.

4 Multiclass classification

In this section, we consider k ($\geq$ 3)-class classification for high-dimensional data. Let

$$Y_i(\boldsymbol{x}_0|n_i) = ||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}||^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i}$$

for $i = 1, \ldots, k$. We consider a classification rule in which one classifies an individual into π_i if

$$\max\left\{\underset{j=1,\ldots,k}{\mathrm{argmin}}\, Y_j(\boldsymbol{x}_0|n_j)\right\} = i. \tag{17}$$

For the case that $\mathrm{argmin}_{j=1,\ldots,k} Y_j(\boldsymbol{x}_0|n_j) = \{i_1, \ldots, i_l\}$ with integers $l \in [2,k]$ and $i_1 < \cdots < i_l$, we have that $\max\{\mathrm{argmin}_{j=1,\ldots,k} Y_j(\boldsymbol{x}_0|n_j)\} = i_l$. Note that the difference, $Y_1(\boldsymbol{x}_0|n_1)/2 - Y_2(\boldsymbol{x}_0|n_2)/2$, coincides with the classifier, $w(\boldsymbol{x}_0|n_1,n_2)$, discussed in Sect. 2.1.

4.1 Asymptotic properties of the classifier

Let $\Delta_{ij} = ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2$ for $i, j = 1, \ldots, k;\ i \neq j$. We assume the followings:

(A-vi) $\dfrac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\Delta_{ij}^2} \to 0$ as $p \to \infty$ for $i, j = 1, \ldots, k;\ i \neq j$;

(A-vii) $\dfrac{\max_{i'=1,\ldots,k}\mathrm{tr}(\boldsymbol{\Sigma}_{i'}^2)}{n_i\Delta_{ij}^2} \to 0$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i, j = 1, \ldots, k;\ i \neq j$.

We denote the error of misclassifying an individual from π_i (into another class) by $e(i)$.

Theorem 7 *Assume* (A-vi) *and* (A-vii). *Then, for the classification rule given by* (17), *we have as* $p \to \infty$ *that*

$$e(i) \to 0 \quad \textit{for} \quad i = 1, \ldots, k.$$

Remark 13 If one can assume that $\max_{i'=1,\ldots,k}\{\mathrm{tr}(\boldsymbol{\Sigma}_{i'}^2)\}/\Delta_{ij}^2 \to 0$ as $p \to \infty$ for $i, j = 1, \ldots, k;\ i \neq j$, it follows naturally that (A-vi) and (A-vii) hold. Then, one can claim Theorem 7 even when n_i is fixed for $i = 1, \ldots, k$.

Let

$$\delta_{ij} = \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}{n_j} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{2n_i(n_i-1)} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{2n_j(n_j-1)} \right\}^{1/2}$$

for $i, j = 1, \ldots, k;\ i \neq j$. We assume an extra assumption:

(A-viii) $\dfrac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\delta_{ij}^2} \to 0$ as $p \to \infty$ and $n_{i'} \to \infty,\ i' = 1, \ldots, k,$ for $i, j = 1, \ldots, k;\ i \neq j$.

Theorem 8 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-viii). *We have as* $p \to \infty$ *and* $n_i \to \infty,\ i = 1, \ldots, k$, *that*

$$\frac{Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j) + \Delta_{ij}}{2\delta_{ij}} \Rightarrow N(0,1) \quad \textit{when} \quad \boldsymbol{x}_0 \in \pi_i$$

for $i, j = 1, \ldots, k;\ i \neq j$.

Corollary 1 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-viii). *For the classification rule given by* (17)*, we have as* $p \to \infty$ *and* $n_i \to \infty,\ i = 1, \ldots, k$, *that*

$$e(i) \le \sum_{j(\neq i)=1}^{k} \Phi\{-\Delta_{ij}/(2\delta_{ij})\} + o(1) \quad \textit{for} \quad i = 1, \ldots, k.$$

4.2 Sample size determination to control misclassification rates

Let $\Delta_i = \min_{j(\neq i)=1,\ldots,k} \Delta_{ij}$ for $i = 1, \ldots, k$. We are interested in designing a classifier having $e(i) \le \alpha_i$ when $\Delta_i \ge \Delta_{iL}$ for all $i = 1, \ldots, k$, where $\alpha_i \in (0, 1/2)$ and $\Delta_{iL}\ (> 0),\ i = 1, \ldots, k$, are prespecified constants. We assume $\Delta_{iL} = o\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}\},\ i = 1, \ldots, k$. Let $\sigma = \max_{i=1,\ldots,k} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$ and $\sigma_i = \max_{j(\neq i)=1,\ldots,k} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/2}$. Let $\alpha_{(i)} = \min_{j(\neq i)=1,\ldots,k} \alpha_j$. Then, we find the sample size for each π_i as

$$n_i \ge \frac{(z_{\alpha_i/(k-1)} + z_{\alpha_{(i)}/(k-1)})^2 \sigma}{\Delta_{iL}^2} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} + \sigma_i^{1/2}\} + 1 \quad (= C_i,\ \text{say}). \tag{18}$$

According to (18), we take samples from each π_i in order to calculate $Y_i(\boldsymbol{x}_0|n_i)$.

We consider the following classification procedure:

[Misclassification rate adjusted classifier (MRAC)]

(Step 1) Set $i = 0$.

(Step 2) Put $i = i + 1$. If $i = k$, go to Step 4; otherwise go to Step 3.

(Step 3) If it holds that

$$Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j) < \max(\Delta_{iL}, \Delta_{jL}) \frac{z_{\alpha_i/(k-1)} - z_{\alpha_j/(k-1)}}{z_{\alpha_i/(k-1)} + z_{\alpha_j/(k-1)}}$$

for all $j = i+1, \ldots, k$, go to Step 4; otherwise go to Step 2.

(Step 4) Classify $\boldsymbol{x}_0$ into π_i.

Note that $P\{N(0, 1) \ge z_{\alpha_i/(k-1)}\} = \alpha_i/(k-1)$. We have the following theorem.

Theorem 9 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-vi). *Then, for the MRAC with* (18)*, it holds as $p \to \infty$ that*

$$\limsup e(i) \le \alpha_i \tag{19}$$

when $\Delta_i \ge \Delta_{iL}$ for $i = 1, \ldots, k$.

Remark 14 If we consider the classification satisfying $e(i) \le \alpha \in (0, 1/2)$ when $\Delta_i \ge \Delta_{iL}$ for $i = 1, \ldots, k$, one can find the sample size for each π_i as $n_i \ge \Delta_{iL}^{-2} 4 z_{\alpha/(k-1)}^2 \sigma \operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/4}\{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} + \sigma_i^{1/2}\} + 1$. Then, under (A-ii), (A-vi) and (1) with (A-i), for the classifier given by (17), it holds as $p \to \infty$ that $\limsup e(i) \le \alpha$ when $\Delta_i \ge \Delta_{iL}$ for $i = 1, \ldots, k$.

Remark 15 From Remark 8, we have a lower bound of Δ_{ij} for $i, j = 1, \ldots, k;\ i \ne j$, by

$$\Delta_{ijL} = ||\overline{\boldsymbol{x}}_{in_i} - \overline{\boldsymbol{x}}_{jn_j}||^2 - \operatorname{tr}(\boldsymbol{S}_{in_i})/n_i - \operatorname{tr}(\boldsymbol{S}_{jn_j})/n_j - \kappa_{ij} z_{\alpha'},$$

where $\alpha' \in (0, 1/2)$ and $\kappa_{ij} = [2W_{in_i}/\{n_i(n_i-1)\} + 2W_{jn_j}/\{n_j(n_j-1)\} + 4\operatorname{tr}(\boldsymbol{S}_{in_i}\boldsymbol{S}_{jn_j})/(n_i n_j)]^{1/2}$ having W_{in_i} according to (16). Thus, one may design a lower bound of Δ_i by $\Delta_{iL} = \min_{j(\ne i)=1,\ldots,k} \Delta_{ijL}$ for sufficiently small α'. Then, it holds that $\Delta_{iL}/\Delta_i = 1 + o_p(1)$ under (A-vi) and (1) with (A-i) when $\operatorname{tr}(\boldsymbol{\Sigma}_i^2)/(n_i^2\Delta_i^2) \to 0$ for $i = 1, \ldots, k$.

4.3 Two-stage procedure

In order to estimate C_is in (18), we proceed with the following two steps:

1. Choose $m_i (\ge 4)$ satisfying (13) for each π_i. Take pilot samples, $\boldsymbol{x}_{ij},\ j = 1, \ldots, m_i$, of size m_i from each π_i. Then, calculate W_{im_i} for each π_i according to (16). Let $\widehat{\sigma} = \max_{i=1,\ldots,k} W_{im_i}^{1/2}$ and $\widehat{\sigma}_i = \max_{j(\ne i)=1,\ldots,k} W_{jm_j}^{1/2}$. Define the total sample size for each π_i by

$$N_i = \max\left\{m_i, \left\lceil \frac{(z_{\alpha_i/(k-1)} + z_{\alpha_{(i)}/(k-1)})^2 \widehat{\sigma}}{\Delta_{iL}^2} W_{im_i}^{1/4}\left(W_{im_i}^{1/4} + \widehat{\sigma}_i^{1/2}\right)\right\rceil + 1\right\}. \tag{20}$$

2. For each i, if $N_i = m_i$, do not take any additional samples from π_i and otherwise, that is if $N_i > m_i$, take additional samples, $\boldsymbol{x}_{ij},\ j = m_i+1, \ldots, N_i$, of size $N_i - m_i$ from π_i. By combining the initial samples and the additional samples, calculate $\overline{\boldsymbol{x}}_{iN_i}$ and $\boldsymbol{S}_{iN_i},\ i = 1, \ldots, k$. Then, follow MRAC by using $Y_i(\boldsymbol{x}_0|N_i)$ instead of $Y_i(\boldsymbol{x}_0|n_i)$.

Theorem 10 *Assume* (1) *with* (A-i). *Assume also* (A-ii) *and* (A-vi). *Then, for the MRAC with* (13) *and* (20), (19) *holds as* $p \to \infty$ *when* $\Delta_i \ge \Delta_{iL}$ *for* $i = 1, \ldots, k$.

Remark 16 Under (A-ii), (13) and (1) with (A-i), it holds as $p \to \infty$ that $N_i/C_i = 1 + o_p(1)$, which is in the HDLSS situation in the sense that $N_i/p = o_p(1)$ under the condition that $\max_{j=1,\ldots,k}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)\}/\Delta_{iL}^2 = o(p)$.

Remark 17 One may take $m_i(\ge 4)$ such as satisfies $m_i/C_i > 1$ for some i. Then, the assertion in Theorem 10 is still claimed. However, it may cause over-sampling in the sense that $N_i/C_i > 1$ w.p.1.

4.4 Simulation

In order to examine the performance of the MRAC with (13) and (20), we used computer simulations. We considered three classes having a non-Gaussian distribution generated by $y_{ijl} = (8/10)^{1/2} w_{ijl}$, where $w_{ijl},\ j = 1, \ldots, p\ (l = 1, 2, \ldots)$ are independently distributed as t-distribution with 10 degrees of freedom for each $\pi_i\ (i = 1, 2, 3)$. Note that $E(y_{ijl}) = 0,\ E(y_{ijl}^2) = 1$, and $y_{ijl},\ j = 1, \ldots, p\ (i = 1, 2, 3;\ l = 1, 2, \ldots)$ are independent. Let $\boldsymbol{x}_{il} = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2}(y_{i1l}, \ldots, y_{ipl})^{\mathrm{T}} + \boldsymbol{\mu}_i\ (i = 1, 2, 3;\ l = 1, 2, \ldots)$, where $\boldsymbol{\Lambda}_i = \boldsymbol{H}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i \boldsymbol{H}_i$. Then, the population distribution of $\boldsymbol{x}_{il}$ satisfies (A-i) for each π_i. We considered $\boldsymbol{\Sigma}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B},\ \boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_3 = 1.2\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B}$ is defined by (7). We set $\boldsymbol{\mu}_1 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ whose first 40 elements are 1, $\boldsymbol{\mu}_2 = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ whose 40 elements (21st to 60th) are 1, and $\boldsymbol{\mu}_3 = (0, \ldots, 0)^{\mathrm{T}}$. Then, we had $\Delta_i = 40$ for $i = 1, 2, 3$. We prespecified $\Delta_{iL} = 40,\ i = 1, 2, 3$. We set $m_i = \lceil 0.5 \times (C_i - 1)\rceil + 1$ for each π_i, where C_i is defined by (18). We considered four cases: (a) $p = 400$ when $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.1, 0.1)$, (b) $p = 1200$ when $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.1, 0.1)$, (c) $p = 400$ when $(\alpha_1, \alpha_2, \alpha_3) = (0.05, 0.1, 0.15)$, and (d) $p = 1200$ when $(\alpha_1, \alpha_2, \alpha_3) = (0.05, 0.1, 0.15)$.

In Table 2, we summarized the findings obtained by averaging the outcomes from 2000 ($= R$, say) replications in each case. Under a fixed scenario, suppose that the rth replication ends with $N_i = n_{ir}\ (i = 1, 2, 3)$ observations for $r = 1, \ldots, R$. Let $\overline{n}_i = R^{-1}\sum_{r=1}^{R} n_{ir}$ and $V(n_i) = (R-1)^{-1}\sum_{r=1}^{R}(n_{ir} - \overline{n}_i)^2$. In the end of the rth replication, we checked whether the MRAC does (or does not) classify an individual from π_i correctly and defined $P_{ir} = 0$ (or 1) accordingly for each i. We calculated $\overline{e}(i) = R^{-1}\sum_{r=1}^{R} P_{ir},\ i = 1, 2, 3$, as estimates of $e(i)$s. Their estimated standard errors were given by $s\{\overline{e}(i)\},\ i = 1, 2, 3$, where $s^2\{\overline{e}(i)\} = R^{-1}\overline{e}(i)\{1 - \overline{e}(i)\}$. Throughout, the MRAC with (13) and (20) gave adequate performances for all the cases when considered those standard errors.

Table 2 Accuracy of the MRAC with (13) and (20)

	C_i	$\overline{n}_i$	$\overline{n}_i - C_i$	$V(n_i)$	$\overline{e}(i)$	$s\{\overline{e}(i)\}$
When $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.1, 0.1)$						
$p = 400$: $(m_1, m_2, m_3) = (8, 9, 9)$						
π_1	13.44	14.57	1.13	14.06	0.044	0.00456
π_2	15.62	16.77	1.14	27.78	0.06	0.00531
π_3	15.24	16.4	1.16	18.78	0.114	0.00711
$p = 1200$: $(m_1, m_2, m_3) = (20, 24, 23)$						
π_1	38.73	39.34	0.61	14.96	0.029	0.00375
π_2	45.46	46.32	0.87	30.8	0.082	0.00612
π_3	44.18	45.01	0.83	18.7	0.083	0.00615
When $(\alpha_1, \alpha_2, \alpha_3) = (0.05, 0.1, 0.15)$						
$p = 400$: $(m_1, m_2, m_3) = (9, 10, 9)$						
π_1	15.94	17.03	1.1	12.73	0.034	0.00402
π_2	18.56	19.58	1.02	26.81	0.061	0.00535
π_3	16.21	17.02	0.81	15.7	0.143	0.00782
$p = 1200$: $(m_1, m_2, m_3) = (24, 28, 25)$						
π_1	46.3	46.96	0.65	16.95	0.032	0.00391
π_2	54.38	55.0	0.62	39.22	0.051	0.0049
π_3	47.11	47.65	0.54	19.87	0.129	0.00748

5 Example

We analyzed gene expression data given by Armstrong et al. (2002) in which the data set consists of 12582 ($= p$) genes. We had three classes of leukemia subtypes, that is, π_1: acute lymphoblastic leukemia (24 samples), π_2: mixed-lineage leukemia (20 samples), and π_3: acute myeloid leukemia (28 samples). We prespecified $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.2, 0.05)$, so that $\alpha_{(1)} = 0.05$, $\alpha_{(2)} = 0.05$, and $\alpha_{(3)} = 0.1$. We set $m_1 = m_2 = m_3 = 10$. According to Remark 15, by setting $\alpha' = 0.05$ and $n_i = m_i (= 10)$, $i = 1, 2, 3$, we had $\Delta_{12L} = 5.96 \times 10^9$, $\Delta_{13L} = 2.37 \times 10^{10}$, and $\Delta_{23L} = 7.81 \times 10^9$. Thus, we prespecified $\Delta_{1L} = \min(\Delta_{12L}, \ \Delta_{13L}) = 5.96 \times 10^9$, $\Delta_{2L} = \min(\Delta_{12L}, \ \Delta_{23L}) = 5.96 \times 10^9$, and $\Delta_{3L} = \min(\Delta_{13L}, \ \Delta_{23L}) = 7.81 \times 10^9$.

By using pilot samples of size $m_1 = m_2 = m_3 = 10$, we calculated $W_{1m_1} = 2.59 \times 10^{19}$, $W_{2m_2} = 2.16 \times 10^{19}$, $W_{3m_3} = 2.51 \times 10^{19}$, $\widehat{\sigma}^2 = 2.59 \times 10^{19}$, $\widehat{\sigma}_1^2 = 2.51 \times 10^{19}$, and $\widehat{\sigma}_2^2 = \widehat{\sigma}_3^2 = 2.59 \times 10^{19}$ according to (16). From (20), the total sample size for π_1 was given by

$$N_1 = \max\left\{10, \left\lceil \frac{(z_{\alpha_1/(k-1)} + z_{\alpha_{(1)}/(k-1)})^2 \widehat{\sigma}}{\Delta_{1L}^2} W_{1m_1}^{1/4} \left(W_{1m_1}^{1/4} + \widehat{\sigma}_1^{1/2}\right) \right\rceil + 1 \right\} = 20.$$

Similarly, we had $N_2 = 16$ and $N_3 = 12$. We investigated the accuracy of the MRAC with $(N_1, N_2, N_3) = (20, 16, 12)$ by using remaining samples of sizes

Table 3 Average misclassification rates of the MRAC for $m_1 = m_2 = m_3 = 10$ and for $(\varDelta_{1L}, \varDelta_{2L}, \varDelta_{3L}) = (5.96 \times 10^9, 5.96 \times 10^9, 7.81 \times 10^9)$

$(\alpha_1, \alpha_2, \alpha_3)$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}(3)$	(N_1, N_2, N_3)
(0.2, 0.2, 0.2)	0.074	0.127	0.075	(11, 10, 10)
(0.2, 0.2, 0.1)	0.058	0.093	0.069	(14, 13, 10)
(0.2, 0.1, 0.1)	0.072	0.093	0.072	(14, 16, 11)
(0.1, 0.2, 0.05)	0.053	0.133	0.056	(20, 16, 12)
(0.1, 0.1, 0.1)	0.054	0.108	0.072	(17, 16, 11)
(0.1, 0.1, 0.05)	0.045	0.11	0.059	(20, 19, 12)
(0.1, 0.05, 0.1)	0.06	0.06	0.073	(20, 19, 12)
(0.05, 0.1, 0.1)	0.045	0.08	0.064	(20, 19, 12)

When $\alpha_i \le 0.05$ for at least two π_is, the result was not available within the data sets

$24 - N_1 = 4,\ 20 - N_2 = 4$ and $28 - N_3 = 16$ for $\pi_i,\ i = 1, 2, 3$. We randomly split the data sets for $\pi_i,\ i = 1, 2, 3$, into training sets of sizes $(N_1, N_2, N_3) = (20, 16, 12)$ and test sets of sizes $(4, 4, 16)$. We proceeded with the MRAC by calculating $Y_i(\boldsymbol{x}_0|N_i),\ i = 1, 2, 3$, based on a training data set and checked the accuracy by using a test data set for each π_i. We repeated this procedure 100 times. Then, we had the average of the misclassification rates as $\overline{e}(1) = 0.053$, $\overline{e}(2) = 0.133$, and $\overline{e}(3) = 0.056$ for $\pi_i,\ i = 1, 2, 3$. Similarly, for various settings of α_is, we investigated the performance of the MRAC for $m_1 = m_2 = m_3 = 10$ and for $(\varDelta_{1L}, \varDelta_{2L}, \varDelta_{3L}) = (5.96 \times 10^9, 5.96 \times 10^9, 7.81 \times 10^9)$. We summarized the results in Table 3. The MRAC seems to give good performances in such a HDLSS situation.

6 Proofs

6.1 Proof of Theorem 1

We have for $\boldsymbol{x}_0 \in \pi_i$ that

$$E_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0|n_1, n_2)\} = (-1)^i \frac{\varDelta}{2} \quad \text{and}$$

$$\mathrm{Var}_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0|n_1, n_2)\} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_j} + \sum_{i=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{2n_i(n_i - 1)} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\big(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j/n_j\big)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

where $j \neq i$. From the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \le \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}^{1/2}$, it holds that $\mathrm{Var}_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0|n_1, n_2)/\varDelta\} \to 0$ under (A-iii) and (A-iv). Thus by using Chebyshev's inequality, we obtain as $p \to \infty$ that $w(\boldsymbol{x}_0|n_1, n_2)/\varDelta = (-1)^i/2 + o_p(1)$ for $\boldsymbol{x}_0 \in \pi_i,\ i = 1, 2$. This concludes the proof. □

6.2 Proof of Theorem 2

We first consider the case when $\boldsymbol{x}_0 \in \pi_1$. Under (AY-ii) and (AY-iii), it holds that $\mathrm{Var}_{\boldsymbol{\theta}}\{||\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i^2\} = o(\Delta_\star^2)$, $\mathrm{Var}_{\boldsymbol{\theta}}\{(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_1)/n_i\} = O\{\max_{j=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_j^2)/n_i\} = o(\Delta_\star^2)$, $i = 1, 2$, and $\mathrm{Var}_{\boldsymbol{\theta}}\{(\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - \overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2/n_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(\Delta_\star^2)$. Thus by using Chebyshev's inequality, we obtain that

$$\begin{aligned}
&||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)||^2 - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 = ||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + o_p(\Delta_\star);\\
&||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2\\
&\quad = ||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + \Delta + o_p(\Delta_\star).
\end{aligned}$$

We have under (AY-iii) and (1) with (A-i) that $\mathrm{Var}_{\boldsymbol{\theta}}\{\mathrm{tr}(\boldsymbol{S}_{in_i})\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i\} = o(\Delta_\star^2)$, $i = 1, 2$, and $\mathrm{Var}_{\boldsymbol{\theta}}(||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\}$, so that $\mathrm{tr}(\boldsymbol{S}_{in_i}) = \mathrm{tr}(\boldsymbol{\Sigma}_i) + o_p(\Delta_\star)$, $i = 1, 2$, and $||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 = tr(\boldsymbol{\Sigma}_1) + O_p\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2}\}$. Note that $\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_i) = O(1)$, $i = 1, 2$, under $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \in (0, \infty)$ as $p \to \infty$ and $\limsup_{p\to\infty}\{\Delta/\mathrm{tr}(\boldsymbol{\Sigma}_i)\} < \infty$ for $i = 1, 2$. Let $w(\boldsymbol{x}_0)_{AY} = p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}||^2/\mathrm{tr}(\boldsymbol{S}_{1n_1}) - p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}||^2/\mathrm{tr}(\boldsymbol{S}_{2n_2}) - p\log\{\mathrm{tr}(\boldsymbol{S}_{2n_2})/\mathrm{tr}(\boldsymbol{S}_{1n_1})\} - p/n_1 + p/n_2$. Then, under (AY-i) to (AY-iii) and (1) with (A-i), we have that

$$\begin{aligned}
\frac{w(\boldsymbol{x}_0)_{AY}}{p} &= \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)||^2 - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\}\\
&\quad - \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})}\\
&= \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_1) + o_p(\Delta_\star)} - \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + \Delta + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_2) + o_p(\Delta_\star)}\\
&\quad - \log\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2) + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_1) + o_p(\Delta_\star)}\right\}\\
&= \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2\{\mathrm{tr}(\boldsymbol{\Sigma}_2) - \mathrm{tr}(\boldsymbol{\Sigma}_1)\}}{\mathrm{tr}(\boldsymbol{\Sigma}_1)\mathrm{tr}(\boldsymbol{\Sigma}_2)} - \frac{\Delta}{\mathrm{tr}(\boldsymbol{\Sigma}_2)} - \log\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2)}{\mathrm{tr}(\boldsymbol{\Sigma}_1)}\right\}\\
&\quad + o_p\{\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_1)\} + o_p\{\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_2)\}\\
&= \frac{\mathrm{tr}(\boldsymbol{\Sigma}_2) - \mathrm{tr}(\boldsymbol{\Sigma}_1)}{\mathrm{tr}(\boldsymbol{\Sigma}_2)} - \frac{\Delta}{\mathrm{tr}(\boldsymbol{\Sigma}_2)} + \log\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{\mathrm{tr}(\boldsymbol{\Sigma}_2)}\right\} + o_p\{\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_1)\}.
\end{aligned} \tag{21}$$

Note that $\log\{\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} = \{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma}_2) - \{1 + o(1)\}\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}^2/\{2\mathrm{tr}(\boldsymbol{\Sigma}_2)^2\}$ under $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \to 1$ as $p \to \infty$. Then, for the case when $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \to 1$ as $p \to \infty$, we have from (21) that

$$\frac{w(\boldsymbol{x}_0)_{AY}\mathrm{tr}(\boldsymbol{\Sigma}_2)}{p\Delta_\star} = -\frac{\Delta}{\Delta_\star} - \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\}^2}{2\Delta_\star \mathrm{tr}(\boldsymbol{\Sigma}_2)} + o_p(1) = -1 + o_p(1).$$

Next, we consider the case when $\liminf_{p\to\infty} |\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$. Note that $1 - \mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) + \log\{\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} < 0$ under $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \neq 1$. Thus we have from (21) that $w(\boldsymbol{x}_0)_{AY}/p \le 1 - \mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) + \log\{\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} + o_p(1) < 0$ w.p.1. Hence, we conclude the result when $\boldsymbol{x}_0 \in \pi_1$. When $\boldsymbol{x}_0 \in \pi_2$, we have the result similarly. Thus the proof is completed. □

6.3 Proof of Theorem 3

We first consider the case when $\boldsymbol{x}_0 \in \pi_1$. We have from (A-ii) and (A-v) that

$$w(\boldsymbol{x}_0|n_1, n_2) + \Delta/2 = (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\{(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) - (\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)\} + o_p(\delta_1). \quad (22)$$

Let

$$v_{1j} = -\frac{(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)/n_1}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^{1/2}}, \quad j = 1, \ldots, n_1;$$

$$v_{1n_1+j} = \frac{(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2)/n_2}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^{1/2}}, \quad j = 1, \ldots, n_2.$$

Note that $\sum_{j=1}^{n_1+n_2} E_{\boldsymbol{\theta}}(v_{1j}^2) = 1$ and

$$\sum_{j=1}^{n_1+n_2} v_{1j} = \frac{(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\{(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) - (\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)\}}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^{1/2}}.$$

Then, it holds for $j = 2, \ldots, n_1 + n_2$, that $E_{\boldsymbol{\theta}}(v_{1j}|v_{1j-1}, \ldots, v_{11}) = 0$. We consider applying the martingale central limit theorem given by McLeish (1974). Refer to Section 2.6 in Ghosh et al. (1997) for the details of the martingale central limit theorem. Let us write that $\boldsymbol{x}_0 - \boldsymbol{\mu}_1 = \boldsymbol{\Gamma}_1\boldsymbol{y}_0 = \boldsymbol{\Gamma}_1(y_{01}, \ldots, y_{0r_1})^{\mathrm{T}}$ and $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{i1}, \ldots, \boldsymbol{\gamma}_{ir_i})$, $i = 1, 2$. Note that

$$v_{1j}^2 = \frac{\sum_{i,i',l,l'}^{r_1} \boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{1i'}y_{0i}y_{1i'j}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{1l'}y_{0l}y_{1l'j}/n_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2}, \quad j = 1, \ldots, n_1;$$

$$v_{1n_1+j}^2 = \frac{\sum_{i,l}^{r_1}\sum_{i',l'}^{r_2} \boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{2i'}y_{0i}y_{2i'j}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{2l'}y_{0l}y_{2l'j}/n_2^2}{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2}, \quad j = 1, \ldots, n_2.$$

Under (1) with (A-i), we can evaluate for $s = 1, 2$, that

$$
\begin{aligned}
&E_{\boldsymbol{\theta}}\left\{\left(\sum_{i,l}^{r_1}\sum_{i',l'}^{r_s}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{si'}y_{0i}y_{si'j}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{sl'}y_{0l}y_{sl'j}\right)^2\right\}\\
&= \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)^2 + E_{\boldsymbol{\theta}}\left\{\left(\sum_{i\neq l}^{r_1}\sum_{i'\neq l'}^{r_s}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{si'}y_{0i}y_{si'j}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{sl'}y_{0l}y_{sl'j}\right)^2\right\}\\
&\quad + O\left\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s) + \sum_{i=1}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i} + \sum_{i=1}^{r_s}\boldsymbol{\gamma}_{si}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\gamma}_{si}\boldsymbol{\gamma}_{si}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\gamma}_{si}\right\}\\
&= 3\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)^2 + O\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)\} \qquad (23)
\end{aligned}
$$

from the facts that

$$
\begin{aligned}
&\sum_{i=1}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i} \le \sum_{i,l}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1l}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i} = \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s);\\
&\sum_{i=1}^{r_s}\boldsymbol{\gamma}_{si}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\gamma}_{si}\boldsymbol{\gamma}_{si}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\gamma}_{si} \le \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s).
\end{aligned}
$$

Let $I(\cdot)$ be the indicator function. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \le \mathrm{tr}(\boldsymbol{\Sigma}_1^2\boldsymbol{\Sigma}_2^2)$. Then, by using Chebyshev's inequality and Schwarz's inequality, from (23) and (A-ii), we have for Lindeberg's condition that

$$
\begin{aligned}
&\sum_{j=1}^{n_1+n_2} E_{\boldsymbol{\theta}}\{v_{1j}^2 I(v_{1j}^2 \ge \tau)\}\\
&\le \sum_{j=1}^{n_1+n_2}\frac{E_{\boldsymbol{\theta}}(v_{1j}^4)}{\tau} = \sum_{j=1}^{n_1} O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1^4)\}/n_1^4}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^2}\right]\\
&\quad + \sum_{j=n_1+1}^{n_1+n_2} O\left[\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}/n_2^4}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^2}\right]\\
&= O\left[\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^2/n_1^3 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^2/n_2^3}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^2}\right] \to 0
\end{aligned}
$$

for any $\tau > 0$. Now, under (1) with (A-i), we can evaluate for $s, s' = 1, 2$, and $j \neq j'$ that

$$
\begin{aligned}
&E_{\boldsymbol{\theta}}\left\{\left(\sum_{i,l}^{r_1}\sum_{i',l'}^{r_s}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{si'}y_{0i}y_{si'j}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{sl'}y_{0l}y_{sl'j}\right)\right.\\
&\quad\times\left.\left(\sum_{i,l}^{r_1}\sum_{i',l'}^{r_{s'}}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\gamma}_{s'i'}y_{0i}y_{s'i'j'}\boldsymbol{\gamma}_{1l}^{\mathrm{T}}\boldsymbol{\gamma}_{s'l'}y_{0l}y_{s'l'j'}\right)\right\}\\
&=E_{\boldsymbol{\theta}}\left\{\left(\sum_{i,l}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1l}y_{0i}y_{0l}\right)\left(\sum_{i,l}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_{s'}\boldsymbol{\gamma}_{1l}y_{0i}y_{0l}\right)\right\}\\
&=\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})+2\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})+O\big(\sum_{i=1}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_{s'}\boldsymbol{\gamma}_{1i}\big)\\
&=\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})+2\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})\\
&\quad+O[\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})\}^{1/2}] \qquad (24)
\end{aligned}
$$

from the fact that

$$
\begin{aligned}
\sum_{i=1}^{r_1}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_{s'}\boldsymbol{\gamma}_{1i}&\le\big\{\sum_{i=1}^{r_1}(\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_s\boldsymbol{\gamma}_{1i})^2\big\}^{1/2}\big\{\sum_{i=1}^{r_1}(\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_{s'}\boldsymbol{\gamma}_{1i})^2\big\}^{1/2}\\
&\le\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_s)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{s'})\}^{1/2}.
\end{aligned}
$$

Note that $\mathrm{tr}(\boldsymbol{\Sigma}_1^3\boldsymbol{\Sigma}_2)\le\{\mathrm{tr}(\boldsymbol{\Sigma}_1^4)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^{1/2}=o\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1^2\boldsymbol{\Sigma}_2^2)\le\mathrm{tr}(\boldsymbol{\Sigma}_1^4)^{1/2}\mathrm{tr}(\boldsymbol{\Sigma}_2^4)^{1/2}=o\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}=o\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^2\}$ under (A-ii). Then, by using Chebyshev's inequality, we have from (23)–(24) and (A-ii) that

$$
\begin{aligned}
&P_{\boldsymbol{\theta}}\left(\Big|\sum_{j=1}^{n_1+n_2}v_{1j}^2-1\Big|\ge\tau\right)\\
&\quad=O\left[\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^4)/n_1^2+\{\mathrm{tr}(\boldsymbol{\Sigma}_1^4)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^{1/2}/(n_1n_2)+\mathrm{tr}(\boldsymbol{\Sigma}_1^2\boldsymbol{\Sigma}_2^2)/n_2^2}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1+\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^2}\right]\\
&\qquad+O\left[\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^2/n_1^3+\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^2/n_2^3}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1+\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^2}\right]\to 0
\end{aligned}
$$

for any $\tau>0$. Thus it holds that $\sum_{j=1}^{n_1+n_2}v_{1j}^2=1+o_p(1)$. Hence, by using the martingale central limit theorem, we obtain that

$$
\sum_{j=1}^{n_1+n_2}v_{1j}\Rightarrow N(0,1). \qquad (25)
$$

Note that $\delta_1/\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1+\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2\}^{1/2}\to 1$ under (A-ii). Then, by combining (22) with (25), we conclude the result when $\boldsymbol{x}_0\in\pi_1$. When $\boldsymbol{x}_0\in\pi_2$, we have the same arguments. The proof is completed. □

6.4 Proof of Theorem 4

We first consider the case when $\boldsymbol{x}_0\in\pi_1$. We have from (12) that $\delta_1\le\Delta_L/(z_\alpha+z_\beta)$. Here, (A-iii) implies (A-v) for $i=1$ when $\liminf_{p\to\infty}\delta_1/\Delta>0$. Then, from Theorem 3, we claim for (11) that

$$\begin{aligned}1-e(2|1)&=P_{\boldsymbol{\theta}}\left\{w(\boldsymbol{x}_0|n_1,n_2)+\frac{\Delta_L(z_\beta-z_\alpha)}{2(z_\alpha+z_\beta)}<0\right\}\\&=P_{\boldsymbol{\theta}}\left\{\frac{w(\boldsymbol{x}_0|n_1,n_2)+\Delta/2}{\delta_1}<\frac{\Delta}{2\delta_1}+\frac{\Delta_L(z_\alpha-z_\beta)}{2\delta_1(z_\alpha+z_\beta)}\right\}\\&\ge P_{\boldsymbol{\theta}}\left\{\frac{w(\boldsymbol{x}_0|n_1,n_2)+\Delta/2}{\delta_1}<\frac{\Delta_L z_\alpha}{\delta_1(z_\alpha+z_\beta)}\right\}\\&\ge P_{\boldsymbol{\theta}}\{N(0,1)<z_\alpha\}+o(1)\to 1-\alpha\end{aligned}$$

when $\Delta\ge\Delta_L$ and $\liminf_{p\to\infty}\delta_1/\Delta>0$. When $\delta_1/\Delta\to 0$ as $p\to\infty$, it holds that $\mathrm{Var}_{\boldsymbol{\theta}}\{w(\boldsymbol{x}_0|n_1,n_2)\}/\Delta^2\to 0$ under (A-iii). Thus we can claim that

$$\begin{aligned}1-e(2|1)&=P_{\boldsymbol{\theta}}\left\{\frac{w(\boldsymbol{x}_0|n_1,n_2)}{\Delta}<\frac{\Delta_L(z_\alpha-z_\beta)}{2\Delta(z_\alpha+z_\beta)}\right\}\\&=P_{\boldsymbol{\theta}}\left\{o_p(1)<\frac{1}{2}+\frac{\Delta_L(z_\alpha-z_\beta)}{2\Delta(z_\alpha+z_\beta)}\right\}\to 1\end{aligned}$$

when $\Delta\ge\Delta_L$ and $\delta_1/\Delta\to 0$ as $p\to\infty$. Next, when $\liminf_{p\to\infty}\delta_1/\Delta=0$ and $\limsup_{p\to\infty}\delta_1/\Delta\ne 0$, one can claim $\limsup e(2|1)\le\alpha$ by considering the convergent subsequence of δ_1/Δ. Hence, we conclude the result when $\boldsymbol{x}_0\in\pi_1$. When $\boldsymbol{x}_0\in\pi_2$, we have the result similarly. Thus the proof is completed. □

6.5 Proof of Theorem 5

From (13), it holds as $p\to\infty$ that $|N_i-C_i|=o_p(C_i^{1/2})$ under (A-ii) and (1) with (A-i). Then, we write that $|N_i-C_i|=O_p(\omega C_i^{1/2})$, where ω (>0) is a variable such that $\omega\to 0$ as $p\to\infty$. Let $C_{iL}=\lfloor C_i-(\omega C_i)^{1/2}\rfloor,\ i=1,2$. From the fact that $|N_i-C_i|=o_p\{(\omega C_i)^{1/2}\}$, we can claim as $p\to\infty$ that $\max\{m_i,C_{iL}\}\le N_i<C_i+(\omega C_i)^{1/2}$ w.p.1. Then, in a way similar to the proofs of Theorems 2.4 and 2.5 in Aoshima and Yata (2011a), we have that

$$w(\boldsymbol{x}_0|N_1,N_2)=w(\boldsymbol{x}_0|C_{1L},C_{2L})+o_p(\Delta_L)+o_p(\Delta)$$

when $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$. From the fact that $C_{iL}/C_i \to 1$ as $p \to \infty$, similarly to the proof of Theorem 4, we can conclude the results. □

6.6 Proof of Theorem 6

From (23) and (24), it holds as $n \to \infty$ that

$$\mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{2\sum_{i<j}^{n}\{(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{\mu})\}^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)n(n-1)}\right] = \frac{4}{n^2}\{1+o(1)\} + O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n}\right\}.$$

Thus we can evaluate as $n \to \infty$ that

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}}\left(\frac{W_n}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right) &= \mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{2\sum_{i<j}^{n}\{(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{\mu})\}^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)n(n-1)}\right]\{1+o(1)\} \\ &= \frac{4}{n^2}\{1+o(1)\} + O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n}\right\}.\end{aligned}$$

This concludes the proof. □

6.7 Proof of Theorem 7

Under (A-vi) and (A-vii), we have as $p \to \infty$ that $\{Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j)\}/\Delta_{ij} = -1 + o_p(1)$ when $\boldsymbol{x}_0 \in \pi_i$ for $j = 1, \ldots, k;\ j \neq i$. Thus it holds as $p \to \infty$ that $Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j) < 0$ w.p.1 when $\boldsymbol{x}_0 \in \pi_i$ for $j = 1, \ldots, k;\ j \neq i$. Thus we have that $P_{\boldsymbol{\theta}}[\max\{\mathrm{argmin}_{j=1,\ldots,k} Y_j(\boldsymbol{x}_0|n_j)\} = i] \to 1$ when $\boldsymbol{x}_0 \in \pi_i$. This concludes the proof. □

6.8 Proofs of Theorem 8 and Corollary 1

From Theorem 3, we have under (A-i), (A-ii) and (A-viii) that $\{Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j) + \Delta_{ij}\}/(2\delta_{ij}) \Rightarrow N(0,1)$ when $\boldsymbol{x}_0 \in \pi_i$ for $j = 1, \ldots, k;\ j \neq i$. Then, from Bonferroni's inequality, we have that $1 - e(i) \geq 1 - \sum_{j(\neq i)=1}^{k} \Phi\{-\Delta_{ij}/(2\delta_{ij})\} + o(1)$ when $\boldsymbol{x}_0 \in \pi_i$. This concludes the proofs. □

6.9 Proof of Theorem 9

From (18), we have that

$$\delta_{ij}^2 \leq \max_{l=1,\ldots,k}\{\mathrm{tr}(\boldsymbol{\Sigma}_l^2)^{1/2}\}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}}{n_i-1} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/2}}{n_j-1}\right\} \leq \frac{\max(\Delta_{iL}^2,\ \Delta_{jL}^2)}{(z_{\alpha_i/(k-1)} + z_{\alpha_j/(k-1)})^2}.$$

Note that $\Delta_{ij} \geq \max(\Delta_{iL},\ \Delta_{jL})$ when $\Delta_i \geq \Delta_{iL}$, $i = 1, \ldots, k$. Then, in a way similar to the proof of Theorem 4, we have for $j \neq i$ that

$$P_{\boldsymbol{\theta}} \left\{ \frac{Y_i(\boldsymbol{x}_0|n_i) - Y_j(\boldsymbol{x}_0|n_j)}{2\delta_{ij}} \geq \frac{\max(\Delta_{iL},\ \Delta_{jL})}{2\delta_{ij}} \frac{z_{\alpha_i/(k-1)} - z_{\alpha_j/(k-1)}}{z_{\alpha_i/(k-1)} + z_{\alpha_j/(k-1)}} \right\}$$
$$(= e(j|i),\ \text{say})$$
$$\leq \frac{\alpha_i}{k-1} + o(1)$$

when $\boldsymbol{x}_0 \in \pi_i$. Then, from Bonferroni's inequality, we have that $1 - e(i) \geq 1 - \sum_{j(\neq i)=1}^{k} e(j|i) \geq 1 - \alpha_i + o(1)$ when $\boldsymbol{x}_0 \in \pi_i$. This concludes the proof. □

6.10 Proof of Theorem 10

In a way similar to the proofs of Theorems 5 and 9, we can conclude the results. □

Acknowledgments We would like to thank an associate editor and two anonymous referees for their valuable comments.

References

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

Aoshima, M., Yata, K. (2011a). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper)*, *30*, 356–399.

Aoshima, M., Yata, K. (2011b). Authors' response. *Sequential Analysis*, *30*, 432–440.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S.E., Lander, E. S., Golub, T. R., Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*, 41–47.

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

Baik, J., Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, *97*, 1382–1408.

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

Chan, Y.-B., Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, *96*, 469–478.

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

Dudoit, S., Fridlyand, J., 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.

Ghosh, M., Mukhopadhyay, N., Sen, P. K. (1997). *Sequential estimation*. New York: Wiley.

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

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

Huang, S., Tong, T., Zhao, H. (2010). Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics*, *66*, 1096–1106.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, *29*, 295–327.

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

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

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability*, *2*, 620–628.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, *17*, 1617–1642.

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

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, *35*, 251–272.

Vapnic, V. N. (1999). *The nature of statistical learning theory* (second ed.). New York: Springer-Verlag.

Yata, K., Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. Communications in Statistics. *Theory and Methods, Special Issue Honoring Zacks, S. (ed. Mukhopadhyay, N.)*, *38*, 2634–2652.

Yata, K., Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, *101*, 2060–2077.

Yata, K., Aoshima, M. (2012a). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, *105*, 193–215.

Yata, K., Aoshima, M. (2012b). Asymptotic properties of a distance-based classifier for high-dimensional data. *RIMS Koukyuroku*, *1804*, 53–64.

Yata, K., Aoshima, M. (2013). Correlation tests for high-dimensional data using extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, *117*, 313–331.