Noname manuscript No.
(will be inserted by the editor)

High-dimensional quadratic classifiers in non-sparse settings

Makoto Aoshima · Kazuyoshi Yata

Received: date / Accepted: date

Abstract In this paper, we consider high-dimensional quadratic classifiers in non-sparse settings. The quadratic classifiers proposed in this paper draw information about heterogeneity effectively through both the differences of growing mean vectors and covariance matrices. We show that they hold a consistency property in which misclassification rates tend to zero as the dimension goes to infinity under non-sparse settings. We also propose a quadratic classifier after feature selection by using both the differences of mean vectors and covariance matrices. We discuss the performance of the classifiers in numerical simulations and actual data analyses. Finally, we give concluding remarks about the choice of the classifiers for high-dimensional, non-sparse data.

Keywords Asymptotic normality · Bayes error rate · Feature selection · Heterogeneity · Large p small n

Mathematics Subject Classification (2000) 62H30 · 62H10

1 Introduction

Globally, there is an ever increasing need for fast, accurate and cost effective analysis of high-dimensional data in many fields, including academia, medicine and business. However, existing classifiers for high-dimensional data are often complex, time consuming and have no guarantee of accuracy. In this paper we hope to provide better options. A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. This is the so-called "HDLSS" or "large p, small n" data situation, here p is the data dimension and n is the sample size. In this paper, we mainly focus on the case when "$n/p \to$

Makoto Aoshima
Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan
E-mail: aoshima@math.tsukuba.ac.jp

Kazuyoshi Yata
Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan
E-mail: yata@math.tsukuba.ac.jp

0". Suppose we have independent and p-variate two populations, π_i, $i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{ip})^T$ and unknown positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each i. Let

$$\boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\mu_{121}, ..., \mu_{12p})^T \quad \text{and} \quad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2.$$

We assume $\limsup_{p\to\infty} |\mu_{12j}| < \infty$ for all j. Note that $\limsup_{p\to\infty} ||\boldsymbol{\mu}_{12}||^2/p < \infty$, where $||\cdot||$ denotes the Euclidean norm. Let $\sigma_{i(j)}$ be the j-th diagonal element of $\boldsymbol{\Sigma}_i$ for $j = 1, ..., p$ $(i = 1, 2)$. We assume that $\sigma_{i(j)} \in (0, \infty)$ as $p \to \infty$ for all i, j. For a function, $f(\cdot)$, "$f(p) \in (0, \infty)$ as $p \to \infty$" implies that $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$. Here, "$\liminf_{p\to\infty} f(p)$" and "$\limsup_{p\to\infty} f(p)$" are the limit inferior and the limit superior of $f(p)$, respectively. Then, it holds that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/p \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$. We do not assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, ..., \lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$, and $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, ..., \boldsymbol{h}_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$, from each π_i, where $\boldsymbol{x}_{ik} = (x_{i1k}, ..., x_{ipk})^T$, $k = 1, ..., n_i$. We assume $n_i \geq 2$, $i = 1, 2$. Let $n_{\min} = \min\{n_1, n_2\}$. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_{in_i} = (\overline{x}_{i1n_i}, ..., \overline{x}_{ipn_i})^T = \sum_{k=1}^{n_i} \boldsymbol{x}_{ik}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{k=1}^{n_i} (\boldsymbol{x}_{ik} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ik} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$. Let $s_{in_i(j)}$ be the j-th diagonal element of $\boldsymbol{S}_{in_i}$ for $j = 1, ..., p$ $(i = 1, 2)$.

In this paper, we consider high-dimensional quadratic classifiers in non-sparse settings. Let $\boldsymbol{x}_0 = (x_{01}, ..., x_{0p})^T$ be an observation vector of an individual belonging to one of the two populations. We assume that $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. Let $|\boldsymbol{M}|$ be the determinant of a square matrix $\boldsymbol{M}$. When π_is are Gaussian, the Bayes optimal rule (the minimum-error rate discriminant function) is given as follows: One classifies the individual into π_1 if

$$(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_1) - \log |\boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1}| < (\boldsymbol{x}_0 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_2) \qquad (1)$$

and into π_2 otherwise. Since $\boldsymbol{\mu}_i$s and $\boldsymbol{\Sigma}_i$s are unknown, one usually considers the following typical classifier:

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \boldsymbol{S}_{1n_1}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - \log |\boldsymbol{S}_{2n_2} \boldsymbol{S}_{1n_1}^{-1}| < (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \boldsymbol{S}_{2n_2}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}).$$

The classifier usually converges to the Bayes optimal classifier when $n_{\min} \to \infty$ while p is fixed or $n_{\min}/p \to \infty$. However, in the HDLSS context, the inverse matrix of $\boldsymbol{S}_{in_i}$ does not exist. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Bickel and Levina (2004) considered an inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. Fan and Fan (2008) considered a classification after feature selection. Fan et al. (2012) proposed the regularized optimal affine discriminant (ROAD). When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Dudoit et al. (2002) considered an inverse matrix defined by only diagonal elements of $\boldsymbol{S}_{in_i}$. Aoshima and Yata (2011, 2015b) considered using $\{\mathrm{tr}(\boldsymbol{S}_{in_i})/p\}^{-1} \boldsymbol{I}_p$ instead of $\boldsymbol{S}_{in_i}^{-1}$ from a geometrical background of HDLSS data and proposed geometric classifiers. Here, $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. Hall et al. (2005) and Marron et al. (2007) considered distance weighted classifiers. Chan and Hall (2009) and Aoshima and Yata (2014) considered distance-based classifiers and Aoshima and Yata (2014) gave the misclassification rate adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds.

Recently, Cai and Liu (2011), Shao et al. (2011) and Li and Shao (2015) gave sparse linear or quadratic classification rules for high-dimensional data. They showed that their classification rules have Bayes error rates when π_is are Gaussian. They assumed that λ_{ij}s are bounded under some sparsity conditions that $\boldsymbol{\mu}_{12}$, $\boldsymbol{\Sigma}_i$s and $\boldsymbol{\Sigma}_{12}$ (or $\boldsymbol{\Sigma}_i^{-1}$s and $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$) are sparse. See also Donoho and Jin (2015) for the concepts of sparsity. For example, when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ ($= \boldsymbol{\Sigma}$, say), the error rate of their classification rules is given by $\Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ as $p \to \infty$, where $\Delta_{MD} = \boldsymbol{\mu}_{12}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{12}$ that is the (squared) Mahalanobis distance and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Here, $\Phi(-\Delta_{MD}^{1/2}/2)$ is the Bayes error rate. In this paper, we investigate quadratic classifiers from a perspective that is different from the sparse discriminant analysis. We *do not assume that* $\boldsymbol{\mu}_{12}$, $\boldsymbol{\Sigma}_i$*s and* $\boldsymbol{\Sigma}_{12}$ *are sparse*. In such a context, the target of classification rules is not Bayes error rates as in $\Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ as $p \to \infty$. We consider the following consistency property: misclassification rates tend to 0 as p increases, i.e.,

$$e(i) \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2,$$

where $e(i)$ denotes the error rate of misclassifying an individual from π_i into the other class. For example, if one can assume that π_is are Gaussian and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the Bayes rule (1) has such consistency when $\Delta_{MD} \to \infty$ as $p \to \infty$. It is likely that $\Delta_{MD} \to \infty$ as $p \to \infty$ when $\boldsymbol{\mu}_{12}$ is non-sparse in the sense that $||\boldsymbol{\mu}_{12}|| \to \infty$ as $p \to \infty$. We emphasize that such non-sparse situations often occur in high-dimensional settings. For example, see Hall et al. (2005) or (A.1) to (A.4) and Table 2 in Appendix A of the online supplementary. In this paper, we do not consider the following sparse situations:

$$\limsup_{p\to\infty} ||\boldsymbol{\mu}_{12}|| < \infty \ \text{ and } \ \limsup_{p\to\infty} ||\boldsymbol{\Sigma}_{12}||_F < \infty,$$

where $||\cdot||_F$ is the Frobenius norm. For example, most elements of $\boldsymbol{\mu}_{12}$ are 0 or very small if $\limsup_{p\to\infty} ||\boldsymbol{\mu}_{12}|| < \infty$. See Section 5.2 for sparsity of $\boldsymbol{\Sigma}_i$s. We will show that quadratic classifiers hold the consistency property when $\boldsymbol{\mu}_{12}$ or $\boldsymbol{\Sigma}_{12}$ is non-sparse in the sense that

$$||\boldsymbol{\mu}_{12}|| \to \infty \ \text{ or } \ ||\boldsymbol{\Sigma}_{12}||_F \to \infty \ \text{ as } p \to \infty.$$

Based on (1), we consider the following function of $\boldsymbol{A}_i$ to discriminate π_is in general:

$$W_i(\boldsymbol{A}_i) = (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i})^T \boldsymbol{A}_i (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}) - \mathrm{tr}(\boldsymbol{S}_{in_i} \boldsymbol{A}_i)/n_i - \log |\boldsymbol{A}_i|, \tag{2}$$

where $\boldsymbol{A}_i$ is a positive definite matrix satisfying the equation that $\mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} = \mathrm{tr}(\boldsymbol{A}_i^{-1}\boldsymbol{A}_{i'}) - p$ for $i = 1, 2;\ i' \neq i$. Note that $E\{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i})^T \boldsymbol{A}_i (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i})\} = (\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i)^T \boldsymbol{A}_i (\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i) + \mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{A}_i)/n_i$ when $\boldsymbol{x}_0 \in \pi_{i'}$. Thus, $\mathrm{tr}(\boldsymbol{S}_{in_i} \boldsymbol{A}_i)/n_i$ operates as a bias correction term in (2). See Section 2 for typical $\boldsymbol{A}_i$s. We consider a quadratic classification rule in which one classifies the individual into π_1 if

$$W_1(\boldsymbol{A}_1) - W_2(\boldsymbol{A}_2) < 0 \tag{3}$$

and into π_2 otherwise. Note that (3) becomes a linear classifier when $\boldsymbol{A}_1 = \boldsymbol{A}_2$.

Remark 1 As for l (≥ 3)-class classification, one may consider a classification rule that one classifies the individual into π_i if $\mathrm{argmin}_{i'=1,...,l} W_{i'}(\boldsymbol{A}_{i'}) = i$.

In this paper, we pay special attention to the difference of covariance matrices in classification for high-dimensional data. In Section 2, we give a motivation of quadratic classifiers given by (2) via numerical examples in high-dimensional settings. In Section 3, we discuss asymptotic properties of quadratic classifiers given by (2). We show that the classification rule (3) holds the consistency property under non-sparse settings. We verify that a quadratic classifier given by (2) is asymptotically distributed as the normal distribution under certain conditions. In Section 4, we consider estimation of $\boldsymbol{A}_i$s and give asymptotic properties of estimated classifiers. In Section 5, we propose a quadratic classifier after feature selection by using both the differences of mean vectors and covariance matrices. We discuss the performance of the classifiers in numerical simulations and actual data analyses. Finally, in Section 6, we give several suggestions about how to select $\boldsymbol{A}_i$ as concluding remarks.

2 Motivation of quadratic classifiers

We have that $E\{W_{i'}(\boldsymbol{A}_{i'})\} - E\{W_i(\boldsymbol{A}_i)\} = \Delta_i$ when $\boldsymbol{x}_0 \in \pi_i$, where

$$\Delta_i = \boldsymbol{\mu}_{12}^T \boldsymbol{A}_{i'} \boldsymbol{\mu}_{12} + \mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} + \log |\boldsymbol{A}_{i'}^{-1}\boldsymbol{A}_i| \quad \text{for } i = 1, 2;\ i' \neq i. \tag{4}$$

Proposition 1 *(i)* $\Delta_i \geq 0$. *(ii)* $\Delta_i > 0$ *when* $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ *or* $\boldsymbol{A}_1 \neq \boldsymbol{A}_2$.

We note that $\Delta_i \geq \mathrm{tr}(\boldsymbol{A}_i^{-1}\boldsymbol{A}_{i'}) - p + \log |\boldsymbol{A}_{i'}^{-1}\boldsymbol{A}_i| \geq 0$ when $\mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} = \mathrm{tr}(\boldsymbol{A}_i^{-1}\boldsymbol{A}_{i'}) - p$ for $i = 1, 2;\ i' \neq i$. See (B.7) in Appendix B of the online supplementary. However, it does not always satisfy $\Delta_i \geq 0$ if $\mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} \neq \mathrm{tr}(\boldsymbol{A}_i^{-1}\boldsymbol{A}_{i'}) - p$ for $i = 1, 2;\ i' \neq i$. Also note that if one ignores the bias correction term in (2), $W_i(\boldsymbol{A}_i)$ becomes $W_i(\boldsymbol{A}_i) = (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i})^T \boldsymbol{A}_i (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}) - \log |\boldsymbol{A}_i|$. Then, it does not always satisfy $\Delta_i \geq 0$.

In this paper, we specially consider the following four typical $\boldsymbol{A}_i$s in (2):

$$\text{(I) } \boldsymbol{A}_i = \boldsymbol{I}_p, \quad \text{(II) } \boldsymbol{A}_i = \frac{p}{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\boldsymbol{I}_p, \quad \text{(III) } \boldsymbol{A}_i = \boldsymbol{\Sigma}_{i(d)}^{-1}, \text{ and } \quad \text{(IV) } \boldsymbol{A}_i = \boldsymbol{\Sigma}_i^{-1},$$

where $\boldsymbol{\Sigma}_{i(d)} = \mathrm{diag}(\sigma_{i(1)}, ..., \sigma_{i(p)})$. These four $\boldsymbol{A}_i$s satisfy the condition that $\mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} = \mathrm{tr}(\boldsymbol{A}_i^{-1}\boldsymbol{A}_{i'}) - p$ for $i = 1, 2;\ i' \neq i$, and they provide historical background of discriminant analysis. Note that $||\boldsymbol{\Sigma}_{12}||_F \geq ||\boldsymbol{A}_1^{-1} - \boldsymbol{A}_2^{-1}||_F$ for these four $\boldsymbol{A}_i$s. Also, under (I) to (IV), we note that $\Delta_i \to \infty$ as $p \to \infty$ when $\boldsymbol{\mu}_{12}$ or $\boldsymbol{\Sigma}_{12}$ is non-sparse. Practically, $\boldsymbol{A}_i$s should be estimated except for (I). We will consider quadratic classifiers by estimating $\boldsymbol{A}_i$s in Section 4.

Now, let us see an easy example to check the performance of (I) to (IV) in (3). We plugged the true $\boldsymbol{\Sigma}_i$s in $\boldsymbol{A}_i$s for (II) to (IV). We set $p = 2^s,\ s = 3, ..., 11$. Independent pseudo random observations were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{B}_1(0.3^{|i-j|^{1/3}})\boldsymbol{B}_1$, where $\boldsymbol{B}_1 = \mathrm{diag}[\{0.5 + 1/(p+1)\}^{1/2}, ..., \{0.5 + p/(p+1)\}^{1/2}]$. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_1) = p$ and $\boldsymbol{\Sigma}_{1(d)} = \boldsymbol{B}_1^2$. Let $p_\star = \lceil p^{1/2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $(n_1, n_2) = (\log_2 p, 2\log_2 p)$, we considered three cases:

(a) $\boldsymbol{\mu}_2 = (1, ..., 1, -1, ..., -1, 0, ..., 0)^T$ whose first $2p_\star$ elements are not 0. The first $p_\star$ elements are 1 and the next $p_\star$ elements are -1;
(b) $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1, -1...., -1)^T$ whose last $2p_\star$ elements are not 0. The last $p_\star$ elements are -1 and the previous $p_\star$ elements are 1; and
(c) $\boldsymbol{\mu}_2 = (0, ..., 0, 1, -1, ..., 1, -1)^T$ whose last $2p_\star$ elements are not 0, where the j-element is $(-1)^{j+1}$ for $j = p - 2p_\star + 1, ..., p$.

Note that $||\boldsymbol{\mu}_{12}||^2 = 2p_\star$ for (a) to (c). Next, when $\boldsymbol{\mu}_2 = \mathbf{0}$ (i.e., $\boldsymbol{\mu}_{12} = \mathbf{0}$) and $(n_1, n_2) = (5, 10)$, we considered two cases:

(d) $\boldsymbol{\Sigma}_2 = 1.5\boldsymbol{\Sigma}_1$ and (e) $\boldsymbol{\Sigma}_2 = 1.3(0.3^{|i-j|^{1/3}})$ $(\neq \boldsymbol{\Sigma}_1)$.

Note that $\boldsymbol{\mu}_{12}$ or $\boldsymbol{\Sigma}_{12}$ is non-sparse for (a) to (e) because $||\boldsymbol{\mu}_{12}|| \to \infty$ or $||\boldsymbol{\Sigma}_{12}||_F \to \infty$ as $p \to \infty$. For $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$ we repeated 2000 times to confirm if the classification rule (3) with either of (I) to (IV) does (or does not) classify $\boldsymbol{x}_0$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i. We calculated the error rates, $\overline{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Also, we calculated the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. Their standard deviations are less than 0.0112 from the fact that $\mathrm{Var}\{\overline{e}(i)\} = e(i)\{1 - e(i)\}/2000 \leq 1/8000$. In Fig. 1, we plotted $\overline{e}$ for (a) to (e). Note that (I) is equivalent to (II) for (a) to (c).

We observed that (IV) gives the best performance for (c) in Fig. 1. However, (IV) gave the worst performance for (b) contrary to expectations. In general, one would think that the classifier based on the Mahalanobis distance such as (2) with (IV) is the best when π_is are Gaussian and $n_{\min} \to \infty$. We emphasize that it is not true for high-dimensional data. We will explain its theoretical reason in Section 3.3. We observed that (I) (or (II)) gives a better performance than (III) for (b) and (c) in Fig. 1. We will discuss the reasons in Section 3.4. As for (d) and (e) in Fig. 1, the error rates of (I) were close to 0.5 because of $\boldsymbol{\mu}_{12} = \mathbf{0}$. On the other hand, the quadratic classifiers, with (II), (III) and (IV), gave good performances as $||\boldsymbol{\Sigma}_{12}||_F$ increases. From these observations, we pay special attention to the difference of covariance matrices in classification for high-dimensional data. We will give their theoretical backgrounds in Sections 3.1 and 3.4.

3 Asymptotic properties of quadratic classifiers

In this section, we discuss asymptotic properties of quadratic classifiers given by (2) without estimating $\boldsymbol{A}_i$s. We will consider estimated quadratic classifiers in Section 4 by using the results in this section. Similar to Aoshima and Yata (2015a) and Bai and Saranadasa (1996), we assume the following assumption about population distributions as necessary:

(A-i) Let $\boldsymbol{y}_{ik}$, $k = 1, ..., n_i$, be i.i.d. random q_i-vectors having $E(\boldsymbol{y}_{ik}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{y}_{ik}) = \boldsymbol{I}_{q_i}$ for each i $(= 1, 2)$, where $q_i \geq p$. Let $\boldsymbol{y}_{ik} = (y_{i1k}, ..., y_{iq_ik})^T$ whose components satisfy that $\limsup_{p\to\infty} E(y_{ijk}^4) < \infty$ for all j and

$$E(y_{ijk}^2 y_{irk}^2) = E(y_{ijk}^2)E(y_{irk}^2) = 1 \quad \text{and} \quad E(y_{ijk}y_{irk}y_{isk}y_{itk}) = 0 \qquad (5)$$

for all $j \neq r, s, t$. Then, the observations, $\boldsymbol{x}_{ik}$s, from each π_i $(i = 1, 2)$ are given by $\boldsymbol{x}_{ik} = \boldsymbol{\Gamma}_i\boldsymbol{y}_{ik} + \boldsymbol{\mu}_i$, $k = 1, ..., n_i$, where $\boldsymbol{\Gamma}_i = [\boldsymbol{\gamma}_{i1}, ..., \boldsymbol{\gamma}_{iq_i}]$ is a $p \times q_i$ matrix such that $\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$.

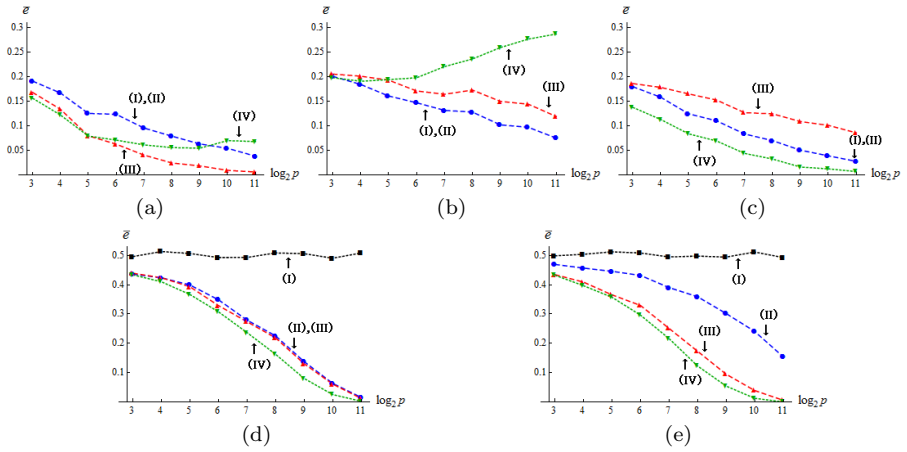


Fig. 1 The average error rates of the classification rule (3) for (I) to (IV). The top panels (a), (b) and (c) display $\overline{e}$ when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ (but $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$). The bottom panels (d) and (e) display $\overline{e}$ when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ (but $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$).

Note that $\boldsymbol{\Gamma}_i$ includes the case that $\boldsymbol{\Gamma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2} = [\lambda_{i1}^{1/2} \boldsymbol{h}_{i1}, ..., \lambda_{ip}^{1/2} \boldsymbol{h}_{ip}]$. We assume the following assumption instead of (A-i) as necessary:

(A-ii) Replace (5) with the independence of y_{ijk}, $j = 1, ..., q_i$ $(i = 1, 2;\ k = 1, ..., n_i)$ in (A-i).

Note that (A-ii) is a special case of (A-i). When π_i has $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, (A-ii) naturally holds.

3.1 Consistency property of quadratic classifiers

In this section, we discuss consistency properties of quadratic classifiers given by (2). We consider the following divergence condition for p and n_is:

$(\star)$ $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$.

Let $\Delta_{iA} = \boldsymbol{\mu}_{12}^T \boldsymbol{A}_{i'} \boldsymbol{\Sigma}_i \boldsymbol{A}_{i'} \boldsymbol{\mu}_{12}$ for $i = 1, 2;\ i' \neq i$. Recall that $\Delta_i = \boldsymbol{\mu}_{12}^T \boldsymbol{A}_{i'} \boldsymbol{\mu}_{12} + \mathrm{tr}\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_{i'} - \boldsymbol{A}_i)\} + \log |\boldsymbol{A}_{i'}^{-1} \boldsymbol{A}_i|$. We consider the following conditions under $(\star)$ for $i = 1, 2$ $(i' \neq i)$:

$$\text{(C-i)}\ \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_i \boldsymbol{A}_i)^2\}}{n_i \Delta_i^2} = o(1)\ \text{and}\ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{A}_{i'} \boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'}) + \mathrm{tr}\{(\boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'})^2\}/n_{i'}}{n_{i'} \Delta_i^2} = o(1),$$

$$\text{(C-ii)}\ \frac{\Delta_{iA}}{\Delta_i^2} = o(1),\ \text{and}\ \text{(C-iii)}\ \frac{\mathrm{tr}[\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_1 - \boldsymbol{A}_2)\}^2]}{\Delta_i^2} = o(1).$$

We note that (C-i) and (C-iii) are regularity conditions. On the other hand, (C-ii) is a non-sparsity condition for high-dimensional data. If p is fixed, Δ_i is bounded typically, so that (C-ii) is not met. (C-i) to (C-iii) are somewhat complicated. See Proposition 2 for simpler conditions than (C-i) to (C-iii).

We claim consistency for (3) with (2) as follows:

Theorem 1 *Assume (A-i). Assume also (C-i) to (C-iii). Then, we have that*

$$\frac{W_{i'}(\boldsymbol{A}_{i'}) - W_i(\boldsymbol{A}_i)}{\Delta_i} = 1 + o_P(1) \quad \textit{under } (\star) \textit{ when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2\textit{; } i' \neq i.$$

Furthermore, for the classification rule (3) with (2), we have that

$$e(i) \to 0, \ \ i = 1, 2, \ \ \textit{under } (\star). \tag{6}$$

Remark 2 When $\boldsymbol{A}_1 = \boldsymbol{A}_2$, we can claim Theorem 1 without (A-i) and (C-iii).

Let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ be the smallest and the largest eigenvalues of any positive definite matrix, $\boldsymbol{M}$. We use the phrase "$\lambda(\boldsymbol{M}) \in (0, \infty)$ as $p \to \infty$" in the sense that $\liminf_{p\to\infty} \lambda_{\min}(\boldsymbol{M}) > 0$ and $\limsup_{p\to\infty} \lambda_{\max}(\boldsymbol{M}) < \infty$. We note that $\boldsymbol{A}_i$s in (I) to (III) satisfy the condition "$\lambda(\boldsymbol{A}_i) \in (0, \infty)$ as $p \to \infty$". Let $\Delta_{\min} = \min\{\Delta_1, \Delta_2\}$, $\lambda_{\max} = \max\{\lambda_{\max}(\boldsymbol{\Sigma}_1), \lambda_{\max}(\boldsymbol{\Sigma}_2)\}$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{\max}^2) = \max\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2), \mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}$. Now, instead of (C-i) and (C-ii), we consider the following simpler conditions under $(\star)$:

(C-i') $\mathrm{tr}(\boldsymbol{\Sigma}_{\max}^2)/(n_{\min}\Delta_{\min}^2) = o(1)$ and (C-ii') $\lambda_{\max}/\Delta_{\min} = o(1)$.

If $\mathrm{tr}(\boldsymbol{\Sigma}_{\max}^2) = O(p)$, (C-i') is equivalent to $p/(n_{\min}\Delta_{\min}^2) = o(1)$. We note that "$\mathrm{tr}(\boldsymbol{\Sigma}_{\max}^2) = O(p)$" is a natural condition when $\lambda_{\max} = O(p^{1/2})$. See Remark 1.1 in Aoshima and Yata (2011).

Proposition 2 *Assume that* $\limsup_{p\to\infty} \lambda_{\max}(\boldsymbol{A}_i) < \infty$ *for* $i = 1, 2$. *Then, (C-i') and (C-ii') imply (C-i) and (C-ii), respectively. Furthermore, if* $\lambda(\boldsymbol{A}_i) \in (0, \infty)$ *as* $p \to \infty$ *for* $i = 1, 2$, *and* $\boldsymbol{A}_i$, $i = 1, 2$, *are diagonal matrices such as in (I) to (III) in Section 2, (C-ii') implies (C-iii).*

From the fact that $\lambda_{\max}(\boldsymbol{\Sigma}_i) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$ for $i = 1, 2$, we note that (C-i') and (C-ii') hold even when $n_{\min}$ is fixed under

$$\mathrm{tr}(\boldsymbol{\Sigma}_{\max}^2)/\Delta_{\min}^2 \to 0 \ \ \text{as } p \to \infty. \tag{7}$$

As mentioned in Section 2, four typical $\boldsymbol{A}_i$s were specifically selected. We first consider (I). By using $\boldsymbol{A}_i = \boldsymbol{I}_p, i = 1, 2$, (2) and (4) are given as

$$\begin{aligned} &W_i(\boldsymbol{I}_p) = ||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}||^2 - \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i \\ &\text{and } \Delta_1 = \Delta_2 = ||\boldsymbol{\mu}_{12}||^2 \ \ \text{(hereafter called } \Delta_{(I)}\text{)}. \end{aligned} \tag{8}$$

We note that

$$\frac{W_1(\boldsymbol{I}_p) - W_2(\boldsymbol{I}_p)}{2} = \Big(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\Big)^T (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2}$$

and $E\{W_1(\boldsymbol{I}_p) - W_2(\boldsymbol{I}_p)\} = (-1)^i||\boldsymbol{\mu}_{12}||^2 = \Delta_{(I)}$ when $\boldsymbol{x}_0 \in \pi_i$. Thus, the linear classifier by (8) is equivalent to the distance-based classifier by Aoshima and Yata (2014). Hereafter, we call the classifier by (8) the "distance-based discriminant analysis (DBDA)". From Theorem 1 and Proposition 2, we have the following result.

Corollary 1 *Assume (C-i') and (C-ii'). Then, for the classification rule (3) with (8), we have consistency (6).*

From Corollary 1, under (7), the classification rule (3) with (8) has consistency (6) even when n_is are fixed. Note that DBDA has the consistency property without (A-i), so that DBDA is quite robust for non-Gaussian data. See Aoshima and Yata (2014) for the details. When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, DBDA does not satisfy (C-i') and (C-ii'), whereas the quadratic classifiers with (II) to (IV) still satisfy them.

Next, we consider (II) to (IV). For (II), by using $\boldsymbol{A}_i = \{p/\mathrm{tr}(\boldsymbol{\Sigma}_i)\}\boldsymbol{I}_p, i = 1, 2$, (4) is given as

$$\Delta_i = \frac{p\Delta_{(I)}}{\mathrm{tr}(\boldsymbol{\Sigma}_{i'})} + \frac{p\mathrm{tr}(\boldsymbol{\Sigma}_i)}{\mathrm{tr}(\boldsymbol{\Sigma}_{i'})} - p + p\log\Big\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i'})}{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\Big\} \quad \text{(hereafter called } \Delta_{i(II)}\text{)}.$$

For (III), by using $\boldsymbol{A}_i = \boldsymbol{\Sigma}_{i(d)}^{-1}, i = 1, 2$, it is given as

$$\Delta_i = \sum_{j=1}^{p}\Big\{\frac{\mu_{12j}^2}{\sigma_{i'(j)}} + \frac{\sigma_{i(j)}}{\sigma_{i'(j)}} - 1 + \log\Big(\frac{\sigma_{i'(j)}}{\sigma_{i(j)}}\Big)\Big\} \quad \text{(hereafter called } \Delta_{i(III)}\text{)}.$$

For (IV), by using $\boldsymbol{A}_i = \boldsymbol{\Sigma}_i^{-1}, i = 1, 2$, it is given as $\Delta_i = \boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\mu}_{12} + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}^{-1}) - p + \sum_{j=1}^{p}\log(\lambda_{i'j}/\lambda_{ij})$ (hereafter called $\Delta_{i(IV)}$).

Now, we consider the following condition for $\boldsymbol{\Sigma}_i, i = 1, 2$:

$$\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2 \to 0 \text{ as } p \to \infty. \tag{9}$$

We note that $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2$ is a measure of sphericity. Under (A-i) and (9), from the fact that $\mathrm{Var}(||\boldsymbol{x}_0 - \boldsymbol{\mu}_i||^2) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ when $\boldsymbol{x}_0 \in \pi_i$, we have that

$$||\boldsymbol{x}_0 - \boldsymbol{\mu}_i|| = \mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}\{1 + o_P(1)\} \text{ when } \boldsymbol{x}_0 \in \pi_i \text{ as } p \to \infty.$$

Thus the centroid data lies near the surface of an expanding sphere. See Hall et al. (2005) for the details of the geometric representation. We emphasize that (II) draws information about heteroscedasticity thorough the geometric representation having different radii, $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$s, of expanding two spheres. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = o(p^2)$ under (9). Hence, for (II), (7) holds under $\liminf_{p\to\infty}\Delta_{\min(II)}/p > 0$ and (9), where $\Delta_{\min(II)} = \min\{\Delta_{1(II)}, \Delta_{2(II)}\}$. We note that $\Delta_{\min(II)} > 0$ when $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ in view of Proposition 1. If one can assume that $\liminf_{p\to\infty}|\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$, it follows $\liminf_{p\to\infty}\Delta_{\min(II)}/p > 0$, so that (7) holds under (9). Hence, for the classification rule (3) with (II), we have consistency (6) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and n_is are fixed. See (II) in (d) and (e) of Fig. 1. The accuracy becomes higher as the difference between $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s grows.

Similarly, for (III), it follows that (7) holds under $\liminf_{p\to\infty}\Delta_{\min(III)}/p > 0$ and (9), where $\Delta_{\min(III)} = \min\{\Delta_{1(III)}, \Delta_{2(III)}\}$. If one can assume that $\liminf_{p\to\infty}\sum_{j=1}^{p}|\sigma_{1(j)}/\sigma_{2(j)} - 1|/p > 0$, it follows $\liminf_{p\to\infty}\Delta_{\min(III)}/p > 0$, so that the classification rule (3) with (III) has consistency (6) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and n_is are fixed. (III) draws information about heteroscedasticity via the difference of diagonal elements between the two covariance matrices. The accuracy becomes higher as the difference of those diagonal elements grows. See (III) in (d) and (e) of Fig. 1.

For (IV), we have the following result.

Proposition 3 *When* $\liminf_{p\to\infty}|tr(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}^{-1})/p - 1| > 0$ *or* $\liminf_{p\to\infty}\sum_{j=1}^{p}|\lambda_{ij}/\lambda_{i'j} - 1|/p > 0$ $(i \neq i')$*, it follows that* $\liminf_{p\to\infty}\Delta_{i(IV)}/p > 0$.

From Theorem 1 and Proposition 2, for the classification rule (3) with (IV), we have consistency (6) under $\liminf_{p\to\infty} \Delta_{\min(IV)}/p > 0$ and some conditions, where $\Delta_{\min(IV)} = \min\{\Delta_{1(IV)}, \Delta_{2(IV)}\}$. Thus from Proposition 3, the accuracy becomes higher as the difference of eigenvalues or eigenvectors between the two covariance matrices grows. See (IV) in (d) and (e) of Fig. 1.

3.2 Asymptotic normality of quadratic classifiers

In this section, we discuss the asymptotic normality of quadratic classifiers given by (2). We further discuss Bayes error rates for high-dimensional data. Let $m = \min\{p, n_{\min}\}$ and

$$\delta_i = 2\Big\{ \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_i \boldsymbol{A}_i)^2\}}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{A}_{i'} \boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'})}{n_{i'}} + \Delta_{iA} \Big\}^{1/2} \quad \text{for } i = 1, 2;\ i' \neq i. \tag{10}$$

Note that $\delta_i^2 = \mathrm{Var}[2(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T \{\boldsymbol{A}_i(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i) - \boldsymbol{A}_{i'}(\overline{\boldsymbol{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12})\}]$ when $\boldsymbol{x}_0 \in \pi_i$. We assume the following conditions when $m \to \infty$ for $i = 1, 2;\ i' \neq i$:

(C-iv) $\dfrac{\boldsymbol{\mu}_{12}^T \boldsymbol{A}_{i'} \boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'} \boldsymbol{\mu}_{12} + \mathrm{tr}\{(\boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'})^2\}/n_{i'}}{n_{i'} \delta_i^2} = o(1), \quad \dfrac{\mathrm{tr}\{(\boldsymbol{\Sigma}_i \boldsymbol{A}_i)^4\}}{n_i^2 \delta_i^4} = o(1)$
and $\dfrac{\mathrm{tr}\{(\boldsymbol{\Sigma}_i \boldsymbol{A}_{i'} \boldsymbol{\Sigma}_{i'} \boldsymbol{A}_{i'})^2\}}{n_{i'}^2 \delta_i^4} = o(1)$;

(C-v) $\dfrac{\mathrm{tr}[\{\boldsymbol{\Sigma}_i(\boldsymbol{A}_1 - \boldsymbol{A}_2)\}^2]}{\delta_i^2} = o(1)$; and (C-vi) $\dfrac{\Delta_{iA}}{\delta_i^2} = o(1)$.

We note that (C-iv) and (C-v) are regularity conditions. From (B.15) in Appendix B, under (A-i), (C-iv) and (C-v), it holds that

$$\begin{aligned} &W_{i'}(\boldsymbol{A}_{i'}) - W_i(\boldsymbol{A}_i) - \Delta_i \\ &= 2(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T \big\{ \boldsymbol{A}_i(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i) - \boldsymbol{A}_{i'}(\overline{\boldsymbol{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12}) \big\} + o_P(\delta_i) \end{aligned}$$

as $m \to \infty$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2;\ i' \neq i$. Under (C-vi), it holds that $(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T \boldsymbol{A}_{i'} \boldsymbol{\mu}_{12} = o_P(\delta_i)$ as $m \to \infty$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Then, we claim the asymptotic normality of (2) under (A-i) as follows:

Theorem 2 *Assume (A-i). Assume also (C-iv) to (C-vi). Then, we have that*

$$\begin{aligned} &\frac{W_{i'}(\boldsymbol{A}_{i'}) - W_i(\boldsymbol{A}_i) - \Delta_i}{\delta_i} \Rightarrow N(0, 1) \quad \textit{as } m \to \infty \\ &\textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2 \ (i' \neq i), \end{aligned} \tag{11}$$

where "$\Rightarrow$" denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution. Furthermore, for the classification rule (3) with (2), it holds that

$$e(i) = \Phi\Big(\frac{-\Delta_i}{\delta_i}\Big) + o(1) \quad \textit{as } m \to \infty \textit{ for } i = 1, 2. \tag{12}$$

Next, we consider the asymptotic normality of (2) under (A-ii). We assume the following condition instead of (C-vi) when $m \to \infty$ for $i = 1, 2;\ i' \neq i$:

(C-vii) $\dfrac{\sum_{j=1}^{q_i}(\boldsymbol{\gamma}_{ij}^T\boldsymbol{A}_{i'}\boldsymbol{\mu}_{12})^4}{\delta_i^4} = o(1)$.

Remark 3 (C-vii) can be written as a condition concerning eigenvalues and eigenvectors. If $\boldsymbol{\Gamma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}$, $\boldsymbol{A}_i = \boldsymbol{\Sigma}_i^{-1}$, $i = 1, 2$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, it holds that $\sum_{j=1}^{q_i}\{\boldsymbol{\gamma}_{ij}^T\boldsymbol{A}_{i'}\boldsymbol{\mu}_{12}\}^4 = \sum_{j=1}^{p}\psi_j^2$ and $\Delta_{iA} = \boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_{12} = \sum_{j=1}^{p}\psi_j$, where $\psi_j = (\boldsymbol{\mu}_{12}^T\boldsymbol{h}_{ij})^2/\lambda_{ij}$. Hence, the condition " $\sum_{j=1}^{p}\psi_j^2/(\sum_{j=1}^{p}\psi_j)^2 \to 0$ as $p \to \infty$" implies (C-vii).

Note that $\sum_{j=1}^{q_i}(\boldsymbol{\gamma}_{ij}^T\boldsymbol{A}_{i'}\boldsymbol{\mu}_{12})^4 \le \sum_{j,j'=1}^{q_i}(\boldsymbol{\gamma}_{ij}^T\boldsymbol{A}_{i'}\boldsymbol{\mu}_{12})^2(\boldsymbol{\gamma}_{ij'}^T\boldsymbol{A}_{i'}\boldsymbol{\mu}_{12})^2 = \Delta_{iA}^2$. Thus, (C-vii) is milder than (C-vi).

Now, we claim the asymptotic normality of (2) under (A-ii) as follows:

Theorem 3 *Assume (A-ii). Assume also (C-iv), (C-v) and (C-vii). Then, we have (11). Furthermore, for the classification rule (3) with (2), we have (12).*

When considering the classifier (8), from Theorems 2 and 3, we have the following result.

Corollary 2 *Assume (C-iv). Assume either (A-i) and (C-vi) or (A-ii) and (C-vii). Then, for the classification rule (3) with (8), we have (12).*

3.3 Bayes error rates

When considering Theorem 3 under the situation that

$$\mathrm{tr}\{(\boldsymbol{\Sigma}_i\boldsymbol{A}_i)^2\}/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{A}_{i'}\boldsymbol{\Sigma}_{i'}\boldsymbol{A}_{i'})/n_{i'} = o(\Delta_{iA}) \ \text{ as } m \to \infty \tag{13}$$

for $i = 1, 2$; $i' \neq i$, one has (12) as

$$e(i) = \Phi\{-\Delta_i/(2\Delta_{iA}^{1/2})\} + o(1) \text{ as } m \to \infty \text{ for } i = 1, 2.$$

Note that $\delta_i/(2\Delta_{iA}^{1/2}) = 1 + o(1)$ under (13). If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 (= \boldsymbol{\Sigma})$, the ratio $\Delta_i/\Delta_{iA}^{1/2}$ has a maximum when $\boldsymbol{A}_1 = \boldsymbol{A}_2 = \boldsymbol{\Sigma}^{-1}$. Then, the ratio becomes the Mahalanobis distance such as $\Delta_i/\Delta_{iA}^{1/2} = \Delta_{MD}^{1/2}$. The classification rule (3) with (2) has an error rate converging to the Bayes error rate in the sense that $e(i) = \Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ for $i = 1, 2$. On the other hand, if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and π_is are Gaussian, under (C-iii) for (IV), the Bayes optimal classifier by (1) becomes as follows:

$$2(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\mu}_{12} + o_P(\Delta_{i(IV)}) > (-1)^i\Delta_{i(IV)}$$

when $\boldsymbol{x}_0 \in \pi_i$; $i' \neq i$. Note that $\mathrm{Var}\{(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\mu}_{12}\} = \boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\mu}_{12}$ (hereafter called $\Delta_{iA(IV)}$) when $\boldsymbol{x}_0 \in \pi_i$ and $\Delta_{iA(IV)}$ is the same as Δ_{iA} for (IV). Hence, $(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_{i'}^{-1}\boldsymbol{\mu}_{12}/\Delta_{iA(IV)}^{1/2}$ is distributed as $N(0, 1)$ when $\boldsymbol{x}_0 \in \pi_i$: $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Then, the Bayes error rate becomes $e(i) = \Phi\{-\Delta_{i(IV)}/(2\Delta_{iA(IV)}^{1/2})\} + o(1)$ for $i = 1, 2$, under some conditions.

When considering Theorem 3 under the situation that

$$p/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}^{-1})/n_{i'} = o(\Delta_{iA(IV)}) \ \text{ as } m \to \infty \text{ for } i = 1, 2;\ i' \neq i, \tag{14}$$

one can claim that the classification rule (3) for (IV) has the Bayes error rate asymptotically even when π_is are non-Gaussian. Note that (14) is equivalent to (13) for (IV) and (14) usually holds when $n_{\min} \to \infty$ while p is fixed or $p \to \infty$ but $n_{\min}/p \to \infty$. If (14) is not met, the classifier for (IV) is not optimal. We emphasize that (14) does not always hold for high-dimensional settings that $n_{\min}/p \to 0$ or $n_{\min}/p \to c\ (> 0)$. For example, let us consider the setup of (a) to (c) in Fig. 1. The condition "$p/n_i = o(\Delta_{iA(IV)})$" is not met from the facts that $\Delta_{iA(IV)} = O(p^{1/2})$ and $n_1 = n_2 = o(p^{1/2})$, so that (14) does not hold. On the other hand, (C-iv) to (C-vi) hold, so that one can claim the asymptotic normality in Theorem 2. Note that (14) does not hold under (C-vi) for (IV). Thus, the error rate of the classifier based on the Mahalanobis distance does not converge to the Bayes error rate when Theorem 2 is claimed. Such situations frequently occur in HDLSS settings such as $n_{\min}/p \to 0$. This is the reason why the classifier based on the Mahalanobis distance does not always give a preferable performance for high-dimensional data even when $n_{\min} \to \infty$ and π_is are Gaussian.

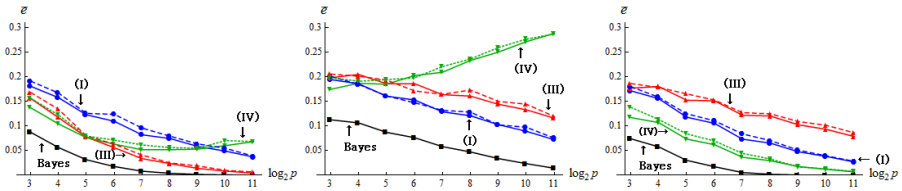
3.4 Comparisons of the classifiers

In this section, we investigate the performance of the classifiers given by (2) for (I) to (IV) in terms of consistency and asymptotic normality. For (I), by using $\boldsymbol{A}_i = \boldsymbol{I}_p, i = 1, 2$, (10) is given as

$$\delta_i = 2\big\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'})/n_{i'} + \boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_{12}\big\}^{1/2} \text{ (hereafter called } \delta_{i(I)}).$$

Similarly, for (II), (III) and (IV), we write δ_i as $\delta_{i(II)}$, $\delta_{i(III)}$ and $\delta_{i(IV)}$. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, we consider (I), (III) and (IV) in the setup of (a) to (c) in Fig. 1. Note that (I), (III) and (IV) satisfy (C-iv) to (C-vi) from the facts that $n_{\min} = o(p^{1/2})$, $\Delta_{iA} = O(||\boldsymbol{\mu}_{12}||^2) = O(p^{1/2})$, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/p \in (0, \infty)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i^4) = o(p^2)$ as $p \to \infty$ for $i = 1, 2$. Thus, Theorem 2 holds for (I), (III) and (IV). We plotted the asymptotic error rates, $\Phi(-\Delta_{(I)}/\delta_{1(I)})$, $\Phi(-\Delta_{1(III)}/\delta_{1(III)})$ and $\Phi(-\Delta_{1(IV)}/\delta_{1(IV)})$ in Fig. 2. From (12) we note that $e(1) - e(2) = o(1)$ when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Thus, the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$, is regarded as an estimate of $e(1)$. We laid $\overline{e}$ for (I), (III) and (IV) by borrowing from Fig. 1. We observed that $\overline{e}$ behaves very close to the asymptotic error rate as expected theoretically. We also plotted the Bayes error rate, $\Phi(-\Delta_{MD}^{1/2}/2)$. We observed that (IV) does not converge to the Bayes error rate when Theorem 2 is claimed. See Section 3.3 for the details. When p is sufficiently large, we note that $\Delta_{(IV)} = \boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_{12}$ $(= \Delta_{MD})$ is small for (b) compared to (c) because $|\boldsymbol{\mu}_{12}^T\boldsymbol{h}_{i1}|$ becomes large for (b) compared to (c), where $\boldsymbol{h}_{i1}$ is the first eigenvector of $\boldsymbol{\Sigma}_i$. Thus (IV) gave a worse performance for (b) than (c). As for (I) and (III), the difference of the performances depends on the configuration of μ_{ij}s and $\sigma_{i(j)}$s. When p is sufficiently large, we note that $\Delta_{(I)} = \sum_{j=1}^p \mu_{12j}^2 < \Delta_{1(III)} = \sum_{j=1}^p \mu_{12j}^2/\sigma_{2(j)}$ for (a) and $\Delta_{(I)} > \Delta_{1(III)}$ for (b) and (c) because $\sigma_{2(j)} = 0.5 + j/(p+1)$, $j = 1, ..., p$ for (a) to (c). It follows that $\Delta_{(I)}/\delta_{i(I)} < \Delta_{i(III)}/\delta_{i(III)}$ for (a) and $\Delta_{(I)}/\delta_{i(I)} > \Delta_{i(III)}/\delta_{i(III)}$ for (b) and (c). Thus (III) is better than (I) for (a), on the other hand, they trade places for (b) and (c).

When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, (II), (III) and (IV) draw information about heteroscedasticity through the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s, $\boldsymbol{\Sigma}_{i(d)}$s or $\boldsymbol{\Sigma}_i$s, respectively. See Section 3.1 for



(a) (b) (c)

Fig. 2 The asymptotic error rates (solid lines) by $\Phi(-\Delta_{(I)}/\delta_{1(I)})$, $\Phi(-\Delta_{1(III)}/\delta_{1(III)})$ and $\Phi(-\Delta_{1(IV)}/\delta_{1(IV)})$, together with the corresponding $\overline{e}$ (dashed lines) by (I), (III) and (IV) in the setup of (a) to (c) in Fig. 1. Also, Bayes optimal error rate was given by $\Phi(-\Delta_{MD}^{1/2}/2)$.

the details. We consider them in the setup of (d) and (e) in Fig. 1. For (d), note that $\Delta_{(I)} = 0$ but $\Delta_{i(II)} = \Delta_{i(III)} = \Delta_{i(IV)} > cp$ for some constant $c > 0$. (II), (III) and (IV) hold the consistency property even when n_is are fixed because (C-i) to (C-iii) are satisfied. Actually, for (d) and (e) in Fig. 1, we observed that the three classifiers gave preferable performances by using the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s, $\boldsymbol{\Sigma}_{i(d)}$s or $\boldsymbol{\Sigma}_i$s as p increases. For (e), note that the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s is smaller than that for (d). Actually, in Fig. 1, we observed that (II) gives a worse performance for (e) than (d). On the other hand, (III) gave a better performance than (II) because $\Delta_{i(III)}$ is sufficiently larger than $\Delta_{i(II)}$ for (e) when p is large. (IV) draws information about heteroscedasticity from the difference of the covariance matrices themselves, so that it gave the best performance in this case. However, we note that it is difficult to estimate $\boldsymbol{\Sigma}_i^{-1}$s feasibly for high-dimensional data. See Section 5.2 for the details.

4 Estimated quadratic classifiers

We denote an estimator of $\boldsymbol{A}_i$ by $\hat{\boldsymbol{A}}_i$. We consider estimating the quadratic classifier by $W_i(\hat{\boldsymbol{A}}_i)$.

4.1 Preliminary

Let $||\boldsymbol{M}|| = \lambda_{\max}^{1/2}(\boldsymbol{M}^T\boldsymbol{M})$ for any square matrix $\boldsymbol{M}$. Let κ be a constant such as $\kappa = \Delta_{\min}$ or $\kappa = \delta_{\min}$, where $\delta_{\min} = \min\{\delta_1, \delta_2\}$. We consider the following condition for $\hat{\boldsymbol{A}}_i$s under $(\star)$:

(C-viii) $p||\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i|| = o_P(\kappa)$ for $i = 1, 2$.

Proposition 4 *Assume (C-viii). Assume also that* $\lambda(\boldsymbol{A}_i) \in (0, \infty)$ *as* $p \to \infty$ *for* $i = 1, 2$. *Then, we have that*

$$W_1(\hat{\boldsymbol{A}}_1) - W_2(\hat{\boldsymbol{A}}_2) = W_1(\boldsymbol{A}_1) - W_2(\boldsymbol{A}_2) + o_P(\kappa) \quad (15)$$

under $(\star)$ *when* $\boldsymbol{x}_0 \in \pi_i$ *for* $i = 1, 2$.

When one chooses $\boldsymbol{A}_i$s as $\boldsymbol{A}_1 = \boldsymbol{A}_2$ $(= \boldsymbol{A})$, $W(\hat{\boldsymbol{A}})$ gives a linear classifier. We consider the following condition for $\hat{\boldsymbol{A}}$ under $(\star)$:

(C-ix) $(p/n_{\min}^{1/2} + p^{1/2}||\boldsymbol{\mu}_{12}||)||\hat{\boldsymbol{A}} - \boldsymbol{A}|| = o_P(\kappa)$.

We have the following result.

Proposition 5 *Assume (C-ix). Then, we have (15).*

We note that (C-ix) is milder than (C-viii) from the fact that $||\boldsymbol{\mu}_{12}|| = O(p^{1/2})$. Hence, we recommend to use a linear classifier such as (8) or (19). The quadratic classifiers should be used when the difference of covariance matrices is considerably large. See Section 4.3 for the details.

4.2 Quadratic classifier by $\hat{\boldsymbol{A}}_i = \{p/\mathrm{tr}(\boldsymbol{S}_i)\}\boldsymbol{I}_p$

We consider the classifier by

$$W_i(\{p/\mathrm{tr}(\boldsymbol{S}_{in_i})\}\boldsymbol{I}_p) = \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}||^2}{\mathrm{tr}(\boldsymbol{S}_{in_i})} - \frac{p}{n_i} + p\log\{\mathrm{tr}(\boldsymbol{S}_{in_i})/p\}. \tag{16}$$

Note that $\Delta_i = \Delta_{i(II)}$ and $\boldsymbol{A}_i = \{p/\mathrm{tr}(\boldsymbol{\Sigma}_i)\}\boldsymbol{I}_p$. From Theorem 1, Propositions 2 and 4, we have the following result.

Corollary 3 *Assume (A-i). Assume also (C-i') and (C-ii'). Then, for the classification rule (3) with (16), we have consistency (6).*

The classifier (16) is equivalent to the geometric classifier by Aoshima and Yata (2011, 2015b). Aoshima and Yata (2011, 2015b) discussed sample size determination to ensure prespecified high accuracy for the classifier. Hereafter, we call the classifier (16) the "geometrical quadratic discriminant analysis (GQDA)". Similar to Section 3.1, we have consistency (6) for GQDA under (A-i) and (7) even when $n_{\min}$ is fixed. If one can assume that $\liminf_{p\to\infty} |\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$, we have consistency (6) for GQDA under (A-i) and (9) even when $n_{\min}$ is fixed and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

As for asymptotic normality, from Corollary B.1 in Appendix B, for the classification rule (3) with (16), we have (12) under the regularity conditions given in Corollary B.1.

Now, we compare DBDA with GQDA. We have that

$$\begin{aligned}\hat{\Delta}_{(I)} =& ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2 \quad \text{and} \\ \hat{\Delta}_{i(II)} =& \frac{p}{\mathrm{tr}(\boldsymbol{S}_{i'n_{i'}})}\Big[\hat{\Delta}_{(I)} + \mathrm{tr}(\boldsymbol{S}_{in_i}) - \mathrm{tr}(\boldsymbol{S}_{i'n_{i'}}) + \mathrm{tr}(\boldsymbol{S}_{i'n_{i'}})\log\Big\{\frac{\mathrm{tr}(\boldsymbol{S}_{i'n_{i'}})}{\mathrm{tr}(\boldsymbol{S}_{in_i})}\Big\}\Big]\end{aligned}$$

for $i = 1, 2$; $i' \neq i$. We note that $E(\hat{\Delta}_{(I)}) = \Delta_{(I)}$. When $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) \to 1$ as $p \to \infty$, it holds $\{\delta_{i(I)}p/\mathrm{tr}(\boldsymbol{\Sigma}_{i'})\}/\delta_{i(II)} = 1 + o(1)$. Then, it follows that $\Delta_{(I)}/\delta_{i(I)} \leq \{1 + o(1)\}\Delta_{i(II)}/\delta_{i(II)}$ in (12) from the fact that $\Delta_{i(II)}\mathrm{tr}(\boldsymbol{\Sigma}_{i'})/p \geq \Delta_{(I)}$. Thus from Corollaries 2 and B.1, if $\hat{\Delta}_{i(II)}\mathrm{tr}(\boldsymbol{S}_{i'n_{i'}})/p$ is sufficiently larger than $\hat{\Delta}_{(I)}$ for some i, we recommend to use GQDA. Otherwise one may use DBDA free from (A-i). See Corollary 1 for the details.

4.3 Quadratic classifier by $\hat{\boldsymbol{A}}_i = \boldsymbol{S}_{in_i(d)}^{-1}$

Let $\boldsymbol{S}_{in_i(d)} = \mathrm{diag}(s_{in_i(1)}, ..., s_{in_i(p)})$ for $i = 1, 2$. We consider the classifier by

$$W_i(\boldsymbol{S}_{in_i(d)}^{-1}) = \sum_{j=1}^{p} \Big(\frac{(x_{0j} - \overline{x}_{ijn_i})^2}{s_{in_i(j)}} - \frac{1}{n_i} + \log s_{in_i(j)} \Big). \tag{17}$$

Note that $\Delta_i = \Delta_{i(III)}$ and $\boldsymbol{A}_i = \boldsymbol{\Sigma}_{i(d)}^{-1}$. Dudoit et al. (2002) considered the quadratic classifier without the bias correction term. That was called the diagonal quadratic discriminant analysis (DQDA). Hereafter, we call the classifier (17) "DQDA-bc". Let $\eta_{i(j)} = \mathrm{Var}\{(x_{ijk} - \mu_{ij})^2\}$ for $i = 1, 2$, and $j = 1, ..., p$ ($k = 1, ..., n_i$). Since $\hat{\boldsymbol{A}}_i = \boldsymbol{S}_{in_i(d)}^{-1}$ does not satisfy (C-viii) in that shape, we consider the following assumption:

(A-iii) $\eta_{i(j)} \in (0, \infty)$ as $p \to \infty$ and $\limsup_{p\to\infty} E\{\exp(t_{ij}|x_{ijk} - \mu_{ij}|^2/\eta_{i(j)}^{1/2})\} < \infty$ for some $t_{ij} > 0$, $i = 1, 2$, and $j = 1, ..., p$ ($k = 1, ..., n_i$).

Note that (A-iii) holds when π_i has $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$. From Theorem 1, Propositions 2 and 4, we have the following result.

Corollary 4 *Assume (A-i) and (A-iii). Assume also (C-ii'). Then, for the classification rule (3) with (17), we have consistency (6) under the condition that*

$$(n_{\min}\Delta_{\min(III)}^2)^{-1} p^2 \log p = o(1). \tag{18}$$

Note that (C-i') holds under (18). From the fact that $\Delta_{i(III)} = O(p)$, it follows that $n_{\min}^{-1} \log p = o(1)$ under (18). Similar to Section 3.1, if one can assume that $\liminf_{p\to\infty} ||\boldsymbol{\mu}_{12}||^2/p > 0$ or $\liminf_{p\to\infty} \sum_{j=1}^{p} |\sigma_{1(j)}/\sigma_{2(j)} - 1|/p > 0$, DQDA-bc holds consistency (6) under (A-i), (A-iii), (9) and $n_{\min}^{-1} \log p = o(1)$. When $\Delta_{\min(III)}$ is not sufficiently large, say $\Delta_{\min(III)} = O(p^{1/2})$, we can claim Corollary 4 in high-dimension, large-sample-size settings such as $n_{\min}/p \to \infty$. In Section 5, we shall provide a DQDA type classifier by feature selection and show that it has the consistency property even when $n_{\min}/p \to 0$ and $\Delta_{\min(III)}$ is not sufficiently large.

Next, we consider the pooled sample diagonal matrix,

$$\boldsymbol{S}_{n(d)} = \frac{\sum_{i=1}^{2}(n_i - 1)\boldsymbol{S}_{in_i(d)}}{\sum_{i=1}^{2} n_i - 2}.$$

Note that $E(\boldsymbol{S}_{n(d)}) = \sum_{i=1}^{2}(n_i - 1)\boldsymbol{\Sigma}_{i(d)}/(\sum_{i=1}^{2} n_i - 2)$ (hereafter called $\boldsymbol{\Sigma}_{(d)}$). When $\boldsymbol{\Sigma}_{1(d)} = \boldsymbol{\Sigma}_{2(d)}$, it follows that $\boldsymbol{\Sigma}_{(d)} = \boldsymbol{\Sigma}_{i(d)}$, $i = 1, 2$. Let us write $\boldsymbol{S}_{n(d)} = \mathrm{diag}(s_{n(1)}, ..., s_{n(p)})$ and $\boldsymbol{\Sigma}_{(d)} = \mathrm{diag}(\sigma_{(1)}, ..., \sigma_{(p)})$. We consider the classifier by

$$W_i(\boldsymbol{S}_{n(d)}^{-1}) = \sum_{j=1}^{p} \Big(\frac{(x_{0j} - \overline{x}_{ijn_i})^2}{s_{n(j)}} - \frac{s_{in_i(j)}}{n_i s_{n(j)}} \Big). \tag{19}$$

We note that the classification rule (3) with (19) becomes a linear classifier. Bickel and Levina (2004) and Dudoit et al. (2002) considered the linear classifier without the bias correction term. That was called the diagonal linear discriminant analysis (DLDA). Hereafter, we call the classifier (19) "DLDA-bc". Although Huang et al. (2010) gave bias corrected versions of DLDA and DQDA,

they considered a bias correction only when π_is are Gaussian. We note that $\Delta_1 = \Delta_2 = \sum_{j=1}^{p} \mu_{12j}^2/\sigma_{(j)}$ (hereafter called $\Delta_{(III')}$) and $\boldsymbol{A}_1 = \boldsymbol{A}_2 = \boldsymbol{\Sigma}_{(d)}^{-1}$. Then, by combining Theorem 1 with Propositions 2 and 5, we have the following result.

Corollary 5 *Assume (A-iii). Assume also (C-i') and (C-ii'). Then, for the classification rule (3) with (19), we have consistency (6) under the condition that*

$$(n_{\min}\Delta_{(III')})^{-1} p \log p = o(1). \tag{20}$$

Under $n_{\min}^{-1} \log p = o(1)$, one may claim that (20) is milder than (18) if $\Delta_{\min(III)}$ and $\Delta_{(III')}$ are of the same order. Hence, DQDA-bc might be considered when $\Delta_{\min(III)}$ is considerably larger than $\Delta_{(III')}$, otherwise DLDA-bc even when $\boldsymbol{\Sigma}_{i(d)}$s are not common. See Appendix A. However, we do not recommend to use both DQDA-bc and DLDA-bc in practice. We shall improve DQDA-bc (or DLDA-bc) by feature selection in Section 5.

4.4 Quadratic classifier by $\hat{\boldsymbol{A}}_i = \boldsymbol{S}_{in_i}^{-1}$

We consider the classifier by

$$W_i(\boldsymbol{S}_{in_i}^{-1}) = (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i})^T \boldsymbol{S}_{in_i}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{in_i}) - p/n_i + \log |\boldsymbol{S}_{in_i}|. \tag{21}$$

In the HDLSS situation where $n_i/p \to 0$, $\boldsymbol{S}_{in_i}^{-1}$ (or $\log |\boldsymbol{S}_{in_i}|$) does not exist. If we suppose high-dimension, large-sample-size situations such as $n_{\min}/p \to \infty$, one can claim from Corollary B.2 in Appendix B that the classification rule (3) with (21) has consistency (6) under the conditions that $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$, and

$$(n_{\min}\Delta_{\min(IV)}^2)^{-1} p^4 \log p = o(1) \tag{22}$$

and the regularity conditions given in Corollary B.2. From the fact that $\Delta_{i(IV)} = O(p)$ when $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$, it follows that $n_{\min}^{-1} p^2 \log p = o(1)$ under (22). Note that the condition "$n_{\min}^{-1} p^2 \log p = o(1)$" is quite strict for high-dimensional data. In Section 5, we shall discuss a classifier by thresholding covariance matrix estimation when $n_{\min}/p \to 0$.

5 Quadratic classifiers by feature selection and thresholding covariance matrix estimation

In this section, we propose a new quadratic classifier by feature selection for (17) and discuss a quadratic classifier by thresholding covariance matrix estimation for (21).

5.1 Quadratic classifier after feature selection

From Corollary 4, DQDA-bc does not always hold consistency (6) even in non-sparse situations (i.e., $\Delta_{\min(III)} \to \infty$). Hence, we consider applying a variable selection procedure to classification. Fan and Fan (2008) proposed the feature annealed independent rule based on the difference of mean vectors. However, we give a different type of feature selection by using both the differences of mean vectors and covariance matrices. We have that

$$\Delta_{1(III)} + \Delta_{2(III)} = \sum_{j=1}^{p} \Big(\frac{\mu_{12j}^2 + \sigma_{1(j)}}{\sigma_{2(j)}} + \frac{\mu_{12j}^2 + \sigma_{2(j)}}{\sigma_{1(j)}} - 2 \Big).$$

Let $\theta_j = (\mu_{12j}^2 + \sigma_{1(j)})/(2\sigma_{2(j)}) + (\mu_{12j}^2 + \sigma_{2(j)})/(2\sigma_{1(j)}) - 1$ for $j = 1, ..., p$. Note that $\Delta_{1(III)} + \Delta_{2(III)} = 2\sum_{j=1}^{p} \theta_j$. Also, note that $\theta_j > 0$ when $\mu_{1j} \neq \mu_{2j}$ or $\sigma_{1(j)} \neq \sigma_{2(j)}$. Now, we give an estimator of θ_j $(j = 1, ..., p)$ by

$$\hat{\theta}_j = \frac{(\overline{x}_{1jn_1} - \overline{x}_{2jn_2})^2 + s_{1n_1(j)}}{2s_{2n_2(j)}} + \frac{(\overline{x}_{1jn_1} - \overline{x}_{2jn_2})^2 + s_{2n_2(j)}}{2s_{1n_1(j)}} - 1.$$

Then, we have the following result.

Theorem 4 *Assume (A-iii). Assume also* $n_{\min}^{-1} \log p = o(1)$. *Then, we have that as* $p \to \infty$

$$\max_{j=1,...,p} |\hat{\theta}_j - \theta_j| = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}.$$

Let $\boldsymbol{D} = \{j \mid \theta_j > 0 \text{ for } j = 1, ..., p\}$ and $p_* = \#\boldsymbol{D}$, where $\#\boldsymbol{S}$ denotes the number of elements in a set $\boldsymbol{S}$. Let $\xi = (n_{\min}^{-1} \log p)^{1/2}$. We select a set of significant variables by

$$\widehat{\boldsymbol{D}} = \{j \mid \hat{\theta}_j > \xi^{\gamma} \text{ for } j = 1, ..., p\}, \tag{23}$$

where $\gamma \in (0, 1)$ is a chosen constant. Then, from Theorem 4, we have the following result.

Corollary 6 *Assume (A-iii),* $n_{\min}^{-1} \log p = o(1)$ *and* $\liminf_{p\to\infty} \theta_j > 0$ *for all* $j \in \boldsymbol{D}$. *Then, we have that* $P(\boldsymbol{D} = \widehat{\boldsymbol{D}}) \to 1$ *as* $p \to \infty$.

Remark 4 As for l (≥ 3)-class classification, one may consider $\hat{\theta}_j$ given by $\hat{\theta}_j = \sum_{i\neq i'}^{k} \{(\overline{x}_{ijn_i} - \overline{x}_{i'jn_{i'}})^2 + s_{in_i(j)}\}/\{k(k-1)s_{i'n_{i'}(j)}\} - 1$ for $j = 1, ..., p$.

Now, we consider a classifier using only the variables in $\widehat{\boldsymbol{D}}$. We define the classifier by

$$W_i(\boldsymbol{S}_{in_i(d)}^{-1})_{FS} = \sum_{j\in\widehat{\boldsymbol{D}}} \Big(\frac{(x_{0j} - \overline{x}_{ijn_i})^2}{s_{in_i(j)}} - \frac{1}{n_i} + \log s_{in_i(j)} \Big) \tag{24}$$

for $i = 1, 2$. We consider the classification rule (3) with (24). We call this feature selected DQDA "FS-DQDA". Let us write that $\boldsymbol{x}_{i*k} = (x_{ij_1k}, ..., x_{ij_{p_*}k})^T$ for all i, k, where $\boldsymbol{D} = \{j_1, ..., j_{p_*}\}$. Let $\boldsymbol{\Sigma}_{i*} = \mathrm{Var}(\boldsymbol{x}_{i*k})$ for $i = 1, 2$ $(k = 1, ..., n_i)$. Then, from Theorem 1 and Corollary 6, we have the following result in non-sparse situations (i.e., $\Delta_{\min(III)} \to \infty$ or $p_* \to \infty$).

Corollary 7 *Assume (A-i) and (A-iii). Assume also* $\lambda_{\max}(\boldsymbol{\Sigma}_{i*})/p_* = o(1)$ *for* $i = 1, 2$, *and* $\liminf_{p\to\infty} \theta_j > 0$ *for all* $j \in \boldsymbol{D}$. *Then, for the classification rule (3) with (24), we have consistency (6) under* $n_{\min}^{-1} \log p = o(1)$.

By comparing Corollaries 7 with 4, note that the condition "$n_{\min}^{-1} \log p = o(1)$" is much milder than (18). Thus we recommend FS-DQDA more than DQDA-bc (or the original DQDA). For the choice of $\gamma \in (0, 1)$ in (23), we recommend applying cross-validation procedures or simply choosing $\gamma = 0.5$ because Corollary 7 is claimed for any $\gamma \in (0, 1)$. See Section 5.3 and Appendix A for the performance of FS-DQDA with $\gamma = 0.5$. In addition, we emphasize that the computational cost of FS-DQDA is quite low even when $p \geq 10,000$.

5.2 Quadratic classifier by thresholding covariance matrix estimation

We consider applying a thresholding estimation of covariance matrices to classification. Bickel and Levina (2008) gave an estimator of $\boldsymbol{\Sigma}_i^{-1}$ for high-dimensional data. Let $\sigma_{i(st)}$ be the (s, t) element of $\boldsymbol{\Sigma}_i$ for $s, t = 1, ..., p$ $(i = 1, 2)$. A sparsity measure of $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ is given by $c_{p,h_i} = \max_{1\leq t\leq p} \sum_{s=1}^{p} |\sigma_{i(st)}|^{h_i}$, where $h_i \in [0, 1)$ is a constant not depending on p and 0^0 is defined as 0. Note that $\lambda_{\max}(\boldsymbol{\Sigma}_i) \leq M c_{p,h_i}$ for some constant $M > 0$. If c_{p,h_i} is much smaller than p for a constant $h_i \in [0, 1)$, $\boldsymbol{\Sigma}_i$ is considered as sparse in the sense that many elements of $\boldsymbol{\Sigma}_i$ are very small. See Section 3 in Shao et al. (2011) for the details. Let $I(\cdot)$ be the indicator function. A thresholding operator is defined by $T_\tau(\boldsymbol{M}) = [m_{st} I(|m_{st}| \geq \tau)]$ for any $\tau > 0$ and any symmetric matrix $\boldsymbol{M} = [m_{st}]$. Let $\tau_{n_i} = M'(n_i^{-1} \log p)^{1/2}$ for some constant $M' > 0$. Then, Bickel and Levina (2008) gave the following result.

Theorem 5 *Assume (A-iii),* $n_i^{-1} \log p = o(1)$ *and* $\liminf_{p\to\infty} \lambda_{\min}(\boldsymbol{\Sigma}_i) > 0$. *For a sufficiently large* $M'(> 0)$, *it holds that as* $p \to \infty$

$$||\{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1} - \boldsymbol{\Sigma}_i^{-1}|| = O_P\big(c_{p,h_i}(n_i^{-1} \log p)^{(1-h_i)/2}\big).$$

Remark 5 Theorem 5 is obtained by Theorem 1 and Section 2.3 in Bickel and Levina (2008).

We use $\hat{\boldsymbol{A}}_i = \{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1}$ as an estimator of $\boldsymbol{\Sigma}_i^{-1}$ and consider the classifier by $W_i(\{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1})$. By combining Theorem 5 and Proposition 4, if it holds that $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \to \infty$ and

$$p c_{p,h_i}(n_i^{-1} \log p)^{(1-h_i)/2} / \Delta_{\min(IV)} = o(1), \tag{25}$$

the classification rule (3) with $W_i(\{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1})$ has consistency (6) under some regularity conditions. When $\boldsymbol{\Sigma}_i$s are sparse as $c_{p,h_i} = O(1)$ for some $h_i (i = 1, 2)$ and $\liminf_{p\to\infty} \Delta_{\min(IV)}/p > 0$, (25) holds in HDLSS situations such as $n_{\min}^{-1} \log p = o(1)$. Shao et al. (2011) and Li and Shao (2015) considered a linear and a quadratic classifier by the thresholding estimation of $\boldsymbol{\Sigma}_i^{-1}$s under some sparsity conditions. On the other hand, Cai et al. (2011) gave the constrained ℓ_1-minimization for inverse matrix estimation (CLIME). One may apply the CLIME to the classification rule (3). However, one should note that the computational cost for the thresholding (or sparse) estimation of $\boldsymbol{\Sigma}_i^{-1}$s is very high even when

$p \approx 1,000$. It is quite unrealistic to apply the estimation to classification when p is very high, say $p \geq 10,000$. Also, the sparsity condition "$\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \to \infty$" is quite severe for high-dimensional data. In actual data analyses, we often encounter the situation that $\lambda_{ij} \to \infty$ as $p \to \infty$ for the first several js. See Yata and Aoshima (2013) for the details.

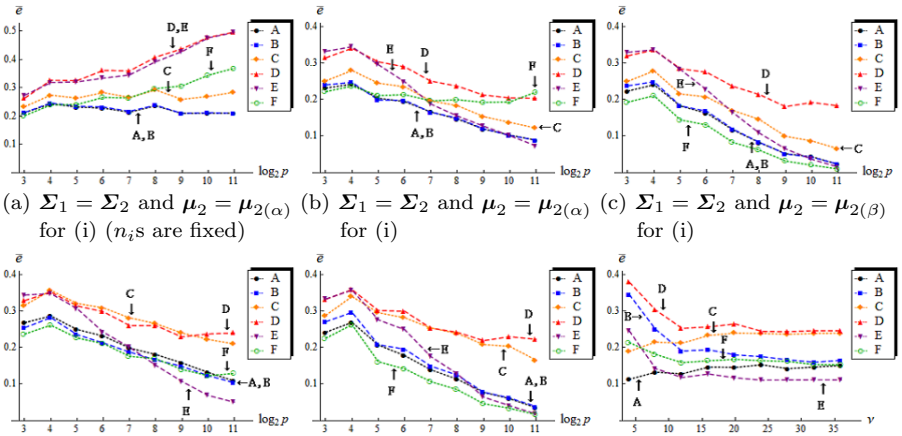
5.3 Simulation

We used computer simulations to compare the performance of the classifiers: DBDA by (8), GQDA by (16), DLDA-bc by (19), DQDA-bc by (17) and FS-DQDA by (24). We did not compare those classifiers with $W_i(\{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1})$ which is given by thresholding estimation of $\boldsymbol{\Sigma}_i^{-1}$s in Section 5.2 because the computational cost of the thresholding estimation is very high when p is large. Thus we considered the classifier (2) with (IV) instead of using the thresholding estimation, provided that $\boldsymbol{\Sigma}_i$s were known. We set $\gamma = 0.5$ in (23). We considered the case of $p_* = 2\lceil p^{1/2}/2 \rceil$. We generated $\boldsymbol{x}_{ik} - \boldsymbol{\mu}_i$, $k = 1, 2, ...,$ $(i = 1, 2)$ independently from (i) $N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, (ii) $\boldsymbol{\Gamma}_i \boldsymbol{y}_{ik}$ with $\boldsymbol{\Gamma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2}$ and $y_{ijk} = (v_{ijk} - 2)/2$ $(j = 1, ..., p)$ in which v_{ijk}s are i.i.d. as the chi-squared distribution with 2 degrees of freedom, or (iii) 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 (A-i) is met in (i) and (ii). We set $p = 2^s$, $s = 3, ..., 11$ for (i) and (ii), and $p = 500$ and $\nu = 4s$, $s = 1, ..., 9$ for (iii). We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{B}_1(0.3^{|i-j|^{1/3}})\boldsymbol{B}_1$, where $\boldsymbol{B}_1$ is defined in Section 2. We considered two cases: $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1, -1...., -1)^T$ $(= \boldsymbol{\mu}_{2(\alpha)}$, say) whose last p_* elements are not 0, where the last $p_*/2$ elements are -1 and the previous $p_*/2$ elements are 1; and $\boldsymbol{\mu}_2 = (0, ..., 0, 1, -1, ..., 1, -1)^T$ $(= \boldsymbol{\mu}_{2(\beta)}$, say) whose last p_* elements are not 0, where the j-element is $(-1)^{j+1}$ for $j = p - p_* + 1, ..., p$. Let $\boldsymbol{B}_2 = \mathrm{diag}(1, ..., 1, 2^{1/2}, ..., 2^{1/2})$ whose last p_* diagonal elements are $2^{1/2}$. We considered six cases:

(a) $n_1 = 10$, $n_2 = 20$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i);
(b) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i);
(c) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\beta)}$ for (i);
(d) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{B}_2 \boldsymbol{\Sigma}_1 \boldsymbol{B}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i);
(e) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{B}_2 \boldsymbol{\Sigma}_1 \boldsymbol{B}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\beta)}$ for (ii); and
(f) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{B}_2 \boldsymbol{\Sigma}_1 \boldsymbol{B}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (iii).

It holds that $n_{\min}^{-1} \log p = o(1)$ for (b) to (f), $\liminf_{p \to \infty} \Delta_{\min}/p_* > 0$ for (a) to (f), and $\liminf_{p \to \infty} |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|/p_* > 0$ for (d), (e) and (f). Similar to Section 2, we calculated the average error rate, $\overline{e}$, by 2000 replications and plotted the results in Fig. 3 (a) to (f).

We observed from (a) in Fig. 3 that DBDA and GQDA give preferable performances when n_is are fixed. DLDA-bc gave a moderate performance because $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. However, the other classifiers did not give preferable performances when p is large. This is probably due to the consistency property of those classifiers (except (IV)) which is claimed under at least $n_{\min}^{-1} \log p = o(1)$. Actually, as for (b) and (c), they gave moderate performances because $n_{\min}^{-1} \log p = o(1)$. Thus, we do not recommend to use quadratic classifiers such as DQDA-bc and FS-DQDA, that consider all the elements (or the diagonal elements) of sample



(a) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i) (n_is are fixed) (b) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i) (c) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\beta)}$ for (i)

(d) $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (i) (e) $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\beta)}$ for (ii) (f) $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2(\alpha)}$ for (iii)

Fig. 3 The average error rates of the classifiers: A: DBDA, B: GQDA, C: DLDA-bc, D: DQDA-bc, E: FS-DQDA, and F: the classifier by (2) with (IV). Their standard deviations are less than 0.0112.

covariance matrices, unless $n_{\min}^{-1} \log p = o(1)$. When $n_{\min}^{-1} \log p \neq o(1)$ or n_is are fixed, we recommend to use DBDA and GQDA. On the other hand, FS-DQDA gave a good performance for (d) and (e) as p increases because the difference of the covariance matrices becomes large as p increases. We note that from Corollary 7 FS-DQDA holds the consistency property for (d) and (e). However, DQDA-bc did not give a preferable performance because $\Delta_{\min(III)} = O(p^{1/2})$, so that DQDA-bc does not hold the consistency property from Corollary 4. We note that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ but $\Delta_{(I)}/\delta_{i(I)} \approx \Delta_{i(II)}/\delta_{i(II)}$ for (d) and (e). Thus GQDA gave a similar performance to DBDA for (d) and (e). As for (f), DBDA gave a preferable performance even when ν is small because DBDA holds the consistency property without (A-i). The other classifiers did not give preferable performances when ν is small. However, they gave moderate performances when ν becomes large because $t_p(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu) \Rightarrow N_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$ as $\nu \to \infty$. Especially, FS-DQDA gave a good performance when ν is not small. This is probably because FS-DQDA has smaller variance by feature selection as $p_*/p \to 0$ than the other classifiers. On the other hand, for several cases, the classifier by (2) with (IV) did not give preferable performances in spite of known $\boldsymbol{\Sigma}_i$s. See Section 3.3 for the theoretical reasons. It is likely that the classifier by $W_i(\{T_{\tau_{n_i}}(\boldsymbol{S}_{in_i})\}^{-1})$ gives poor performances for the high-dimensional settings.

5.4 Example: gene expression data sets

By using gene expression data sets given by Golub et al. (1999) and Armstrong et al. (2002), we compared the performance of the classifiers: DBDA, GQDA, DLDA-bc, DQDA-bc, FS-DQDA and a support vector machine. We summarized the results in Appendix A of the online supplementary.

6 Concluding remarks

In this paper, we considered high-dimensional quadratic classifiers in non-sparse settings. The classifier based on the Mahalanobis distance does not always give a preferable performance even when $n_{\min} \to \infty$ and π_is are assumed Gaussian. See Sections 2 and 3. We emphasize that the quadratic classifiers proposed in this paper draw information about heterogeneity effectively through both the differences of mean vectors and covariance matrices. See Section 3.4 for the details. If the difference of covariance matrices is not sufficiently large, one may use the linear classifier, DBDA. It is quite flexible about the conditions to claim the consistency property. See Sections 4.2 for the details. We emphasize that DLDA-bc, DQDA-bc and FS-DQDA can hold the consistency property under at least "$n_{\min}^{-1} \log p = o(1)$". Thus we do not recommend to use those classifiers when "$n_{\min}^{-1} \log p \neq o(1)$". See Section 5.3 and Appendix A for the details. In such cases, one should use DBDA and GQDA because they hold the consistency property even when n_is are fixed. See Section 4.2 about the choice between DBDA and GQDA. When "$n_{\min}^{-1} \log p = o(1)$", we recommend DQDA-bc and FS-DQDA. Especially, FS-DQDA can claim the consistency property even when $n_{\min}/p \to 0$ and $\Delta_{\min}$ is not sufficiently large. See Section 5.1 for the details. For the choice of $\gamma \in (0, 1)$ in (23), one may apply cross-validation procedures or simply choose as $\gamma = 0.5$. Actually, FS-DQDA with $\gamma = 0.5$ gave preferable performances throughout our simulations and real data analyses. On the other hand, even when $n_{\min}^{-1} \log p = o(1)$, we do not recommend to use the classifier by the thresholding (or sparse) estimation of $\boldsymbol{\Sigma}_i^{-1}$ unless (i) the eigenvalues are bounded in the sense that $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \to \infty$, and (ii) $\boldsymbol{\Sigma}_i$s are sparse in the sense that many elements of $\boldsymbol{\Sigma}_i$s are very small. We emphasize that "$\lambda_{\max}(\boldsymbol{\Sigma}_i)$s are bounded" is a strict condition since the eigenvalues should depend on p and it is probable that $\lambda_{ij} \to \infty$ as $p \to \infty$ for the first several js. See Yata and Aoshima (2013) for the details. Also, the computational cost for the thresholding (or sparse) estimation is very high.

In conclusion, we hope we have given simpler classifiers which will be more effective in the real world analysis of high-dimensional data.

7 Proofs

We give all proofs of the theoretical results together with additional corollaries in Appendix B of the online supplementary.

Acknowledgements We would like to thank the reviewers for their constructive comments. The research of the first author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Exploratory Research, Japan Society for the Promotion of Science (JSPS), under Contract Numbers 15H01678 and 26540010. The research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078.

References

1. Aoshima M, Yata K (2011) Two-stage procedures for high-dimensional data. Seq Anal (Editor's special invited paper) 30:356–399

2. Aoshima M, Yata K (2014) A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. Ann I Stat Math 66:983–1010
3. Aoshima M, Yata K (2015a) Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. Methodol Comput Appl 17:419-439
4. Aoshima M, Yata K (2015b) Geometric classifier for multiclass, high-Dimensional data. Seq Anal 34:279-294
5. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30:41–47
6. Bai Z, Saranadasa H (1996) Effect of high dimension: by an example of a two sample problem. Stat Sinica 6:311–329
7. Bickel PJ, 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
8. Bickel PJ, Levina E (2008) Covariance regularization by thresholding. Ann Stat 36:2577–2604
9. Cai TT, Liu W (2011) A direct estimation approach to sparse linear discriminant analysis. J Am Stat Assoc 106:1566–1577
10. Cai TT, Liu W, Luo X (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J Am Stat Assoc 106:594–607
11. Chan YB, Hall P (2009) Scale adjustments for classifiers in high-dimensional, low sample size settings. Biometrika 96:469–478
12. Donoho D, Jin J (2015) Higher criticism for large-scale inference, especially for rare and weak effects. Stat Sci 30:1–25.
13. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97:77–87
14. Fan J, Fan Y (2008) High-dimensional classification using features annealed independence rules. Ann Stat 36:2605–2637
15. Fan J, Feng Y, Tong X (2012) A road to classification in high dimensional space: the regularized optimal affine discriminant. J Roy Stat Soc B 74:745–771
16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
17. Hall P, Marron JS, Neeman A (2005) Geometric representation of high dimension, low sample size data. J Roy Stat Soc B 67:427–444
18. Huang S, Tong T, Zhao H (2010) Bias-corrected diagonal discriminant rules for high-dimensional classification. Biometrics 66:1096–1106
19. Li Q, Shao J (2015) Sparse quadratic dicriminant analysis for high dimensional data. Stat Sinica 25:457–473
20. Marron JS, Todd MJ, Ahn J (2007) Distance-weighted discrimination. J Am Stat Assoc 102:1267–1271
21. Shao J, Wang Y, Deng X, Wang S (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. Ann Stat 39:1241–1265
22. Yata K, Aoshima M (2013) Correlation tests for high-dimensional data using extended cross-data-matrix methodology. J Multivariate Anal 117:313-331