Ann Inst Stat Math (2019) 71:473–503
https://doi.org/10.1007/s10463-018-0655-z

Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models

Makoto Aoshima[1] · **Kazuyoshi Yata**[1]

Received: 26 December 2016 / Revised: 12 January 2018 / Published online: 10 March 2018

Abstract We consider classifiers for high-dimensional data under the strongly spiked eigenvalue (SSE) model. We first show that high-dimensional data often have the SSE model. We consider a distance-based classifier using eigenstructures for the SSE model. We apply the noise-reduction methodology to estimation of the eigenvalues and eigenvectors in the SSE model. We create a new distance-based classifier by transforming data from the SSE model to the non-SSE model. We give simulation studies and discuss the performance of the new classifier. Finally, we demonstrate the new classifier by using microarray data sets.

Keywords Asymptotic normality · Data transformation · Discriminant analysis · Large p small n · Noise-reduction methodology · Spiked model

1 Introduction

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"

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. Research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078.

✉ Makoto Aoshima
aoshima@math.tsukuba.ac.jp
http://www.math.tsukuba.ac.jp/ aoshima-lab/

Kazuyoshi Yata
yata@math.tsukuba.ac.jp

[1] Institute of Mathematics, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan

data situation where $p/n \to \infty$; here p is the data dimension and n is the sample size. Suppose we have independent and p-variate two populations, π_i, $i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ for each i. 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 $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i(1)}, ..., \lambda_{i(p)})$ is a diagonal matrix of eigenvalues, $\lambda_{i(1)} \geq \cdots \geq \lambda_{i(p)} \geq 0$, and $\boldsymbol{H}_i = [\boldsymbol{h}_{i(1)}, ..., \boldsymbol{h}_{i(p)}]$ 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. We assume $n_i \geq 4$, $i = 1, 2$. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)^T/(n_i - 1)$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. When the π_is are Gaussian, a typical classification rule is that one classifies an individual into π_1 if

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}_1^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_1) - \log \left\{ \det(\boldsymbol{S}_2 \boldsymbol{S}_1^{-1}) \right\} < (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_2^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_2),$$

and into π_2 otherwise. However, the inverse matrix of $\boldsymbol{S}_i$ does not exist in the HDLSS context ($p > n_i$). Also, we emphasize that the Gaussian assumption is strict in real high-dimensional data analyses. Bickel and Levina (2004) considered a naive Bayes classifier for high-dimensional data. Fan and Fan (2008) considered classification after feature selection. Cai and Liu (2011), Shao et al. (2011) and Li and Shao (2015) gave sparse linear or quadratic classification rules for high-dimensional data. The above references all assumed the following eigenvalues condition: There is a constant $c_0 > 0$ (not depending on p) such that

$$c_0^{-1} < \lambda_{i(p)} \ \text{ and } \ \lambda_{i(1)} < c_0 \ \text{ for } i = 1, 2. \tag{1}$$

Dudoit et al. (2002) considered using the inverse matrix defined by only diagonal elements of $\boldsymbol{S}_i$. Aoshima and Yata (2011, 2015a) considered substituting $\{\mathrm{tr}(\boldsymbol{S}_i)/p\}\boldsymbol{I}_p$ for $\boldsymbol{S}_i$ by using the difference of a geometric representation of HDLSS data from each π_i. Here, $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. On the other hand, Hall et al. (2005, 2008) and Marron et al. (2007) considered distance weighted classifiers. Ahn and Marron (2010) considered a HDLSS classifier based on the maximal data piling. Hall et al. (2005), Chan and Hall (2009), Aoshima and Yata (2014) and Watanabe et al. (2015) considered distance-based classifiers. Aoshima and Yata (2014) gave the misclassification rate-adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds under the following eigenvalues condition:

$$\frac{\lambda_{i(1)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2. \tag{2}$$

We emphasize that (2) is much milder than (1) because (2) includes the case that $\lambda_{i(1)} \to \infty$ as $p \to \infty$. See Remark 1 for the details. Aoshima and Yata (2014) considered the distance-based classifier as follows: Let

$$W(\boldsymbol{x}_0) = \left(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2}{2} \right)^T (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1) - \frac{\mathrm{tr}(\boldsymbol{S}_1)}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_2)}{2n_2}. \tag{3}$$

Then, one classifies $\boldsymbol{x}_0$ into π_1 if $W(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. Here, $-\mathrm{tr}(\boldsymbol{S}_1)/(2n_1) + \mathrm{tr}(\boldsymbol{S}_2)/(2n_2)$ is a bias-correction term. Note that the classifier (3) is equivalent to the scale-adjusted distance-based classifier given by Chan and Hall (2009). Aoshima and Yata (2015b) called the classification rule (3) the "distance-based discriminant analysis (DBDA)".

Recently, Aoshima and Yata (2018) considered the "strongly spiked eigenvalue (SSE) model" as follows:

$$\liminf_{p\to\infty}\left\{\frac{\lambda_{i(1)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}\right\} > 0 \ \text{ for } i = 1 \text{ or } 2. \tag{4}$$

On the other hand, Aoshima and Yata (2018) called (2) the "non-strongly spiked eigenvalue (NSSE) model". Note that (4) holds under the condition:

$$\liminf_{p\to\infty}\left\{\frac{\lambda_{i(1)}}{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\right\} > 0 \ \text{ for } i = 1 \text{ or } 2, \tag{5}$$

from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \le \mathrm{tr}(\boldsymbol{\Sigma}_i)^2$. Here, $\lambda_{i(1)}/\mathrm{tr}(\boldsymbol{\Sigma}_i)$ is the first contribution ratio. We call (5) the "super strongly spiked eigenvalue (SSSE) model".

Remark 1 Let us consider a spiked model such as

$$\lambda_{i(r)} = a_{i(r)}\, p^{\alpha_{i(r)}} \ (r = 1, ..., t_i) \quad \text{and} \quad \lambda_{i(r)} = c_{i(r)} \ (r = t_i + 1, ..., p) \tag{6}$$

with positive and fixed constants, $a_{i(r)}$s, $c_{i(r)}$s and $\alpha_{i(r)}$s, and a positive and fixed integer t_i. Note that the NSSE condition (2) holds when $\alpha_{i(1)} < 1/2$ for $i = 1, 2$. On the other hand, the SSE condition (4) holds when $\alpha_{i(1)} \ge 1/2$, and further the SSSE condition (5) holds when $\alpha_{i(1)} \ge 1$. See Yata and Aoshima (2012) for the details of the spiked model.

We observed

$$\frac{\lambda_{i(r)}}{\mathrm{tr}(\boldsymbol{\Sigma}_i)} \ (= \varepsilon_{i(r)}, \ \text{say}) \ \text{ and } \ \frac{\lambda_{i(r)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \ (= \eta_{i(r)}, \ \text{say}), \quad i = 1, 2; \ r = 1, 2, ...,$$

for six well-known microarray data sets by using the noise-reduction methodology and the cross-data-matrix methodology. For those methods, see Yata and Aoshima (2010, 2012). Note that $\varepsilon_{i(r)}$ is the contribution ratio and $\eta_{i(r)}$ is a quadratic contribution ratio of the r-th eigenvalue. We estimated $\varepsilon_{i(r)}$ by $\hat{\varepsilon}_{i(r)} = \tilde{\lambda}_{i(r)}/\mathrm{tr}(\boldsymbol{S}_i)$ and $\eta_{i(r)}$ by $\hat{\eta}_{i(r)} = \acute{\lambda}_{i(r)}^2/\widehat{\Psi}_{i(1)}$, where $\tilde{\lambda}_{i(r)}$ is defined by (16), and $\acute{\lambda}_{i(r)}$ and $\widehat{\Psi}_{i(1)}$ are defined in Sect. 4.3. We note that $\hat{\varepsilon}_{i(r)}$ and $\hat{\eta}_{i(r)}$ are consistent estimators of $\varepsilon_{i(r)}$ and $\eta_{i(r)}$ when $p \to \infty$. See (18) and (23) for the details. The six microarray data sets are as follows:

(D-i) Non-pathologic tissues data with 1413 genes, consisting of π_1: placenta or blood (104 samples) and π_2 : other solid tissue (113 samples) given by Christensen et al. (2009);

Table 1 Estimates of $(\varepsilon_{i(1)}, \eta_{i(1)})$ by $(\hat{\varepsilon}_{i(1)}, \hat{\eta}_{i(1)})$ for the six well-known microarray data sets

	(D-i)	(D-ii)	(D-iii)	(D-iv)	(D-v)	(D-vi)
p	1413	2000	2905	7129	12625	47293
(n_1, n_2)	(104, 113)	(40, 22)	(111, 57)	(58, 19)	(36, 137)	(84, 44)
$\hat{\varepsilon}_{1(1)}$	0.636	0.153	0.108	0.22	0.038	0.091
$\hat{\varepsilon}_{2(1)}$	0.233	0.157	0.083	0.386	0.035	0.085
$\hat{\eta}_{1(1)}$	0.995	0.569	0.304	0.71	0.283	0.502
$\hat{\eta}_{2(1)}$	0.582	0.523	0.363	0.963	0.269	0.403

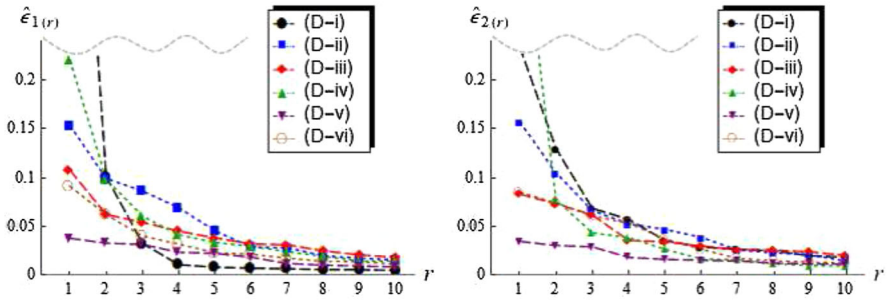


Fig. 1 Estimates of the first ten contribution ratios by $\hat{\varepsilon}_{i(r)}$s for the six well-known microarray data sets

(D-ii) Colon cancer data with 2000 genes, consisting of π_1: colon tumor (40 samples) and π_2 : normal colon (22 samples) given by Alon et al. (1999);

(D-iii) Breast cancer data with 2905 genes, consisting of π_1 : good (111 samples) and π_2 : poor (57 samples) given by Gravier et al. (2010);

(D-iv) Lymphoma data with 7129 genes, consisting of π_1 : DLBCL (58 samples) and π_2 : follicular lymphoma (19 samples) given by Shipp et al. (2002);

(D-v) Myeloma data with 12625 genes, consisting of π_1 : patients without bone lesions (36 samples) and π_2 : patients with bone lesions (137 samples) given by Tian et al. (2003);

(D-vi) Breast cancer data with 47293 genes, consisting of π_1 : luminal group (84 samples) and π_2 : non-luminal group (44 samples) given by Naderi et al. (2007).

The data sets (D-ii), (D-iv) and (D-v) are given in Jeffery et al. (2006), (D-i) and (D-iii) are given in Ramey (2016), and (D-vi) is given in Glaab et al. (2012). We summarized the results for $\hat{\varepsilon}_{i(1)}$ and $\hat{\eta}_{i(1)}$ in Table 1. We also visualized the first ten contribution ratios given by $\hat{\varepsilon}_{i(r)}$ $(r = 1, ..., 10;\ i = 1, 2)$ in Fig. 1 and the first ten quadratic contribution ratios given by $\hat{\eta}_{i(r)}$ $(r = 1, ..., 10;\ i = 1, 2)$ in Fig. 2. See (18) and (23) for the details.

We observed from Fig. 1 that the first several eigenvalues are much larger than the rest for the microarray data sets (except (D-v)). In particular, the first eigenvalues for (D-i) and (D-iv) are extremely large. These data appear to be consistent with the SSSE asymptotic domain given in (5). On the other hand, the first several eigenvalues

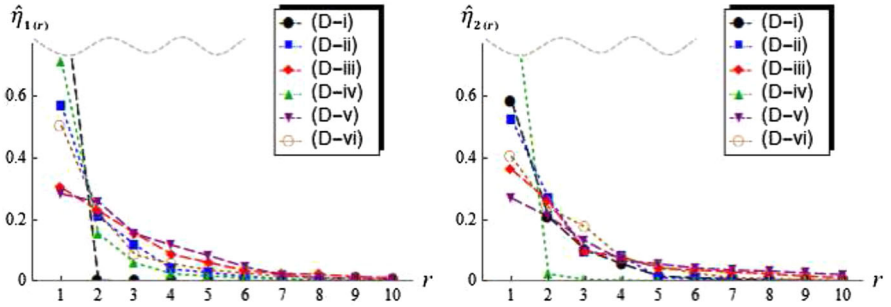


Fig. 2 Estimates of the first ten quadratic contribution ratios by $\hat{\eta}_{i(r)}$s for the six well-known microarray data sets

for (D-v) are relatively small. However, from Table 1 and Fig. 2, $\eta_{i(1)}$s for (D-v) are not sufficiently small. Also, $\eta_{i(1)}$s for (D-ii), (D-iii) and (D-vi) are relatively large in Table 1 and Fig. 2. Hence, the six microarray data appear to be consistent with the SSE asymptotic domain given in (4). See Sect. 4.3. In this paper, we consider classifiers under the SSE model. We do not assume the normality of the population distributions. We propose an effective distance-based classifier for such high-dimensional data sets.

The organization of this paper is as follows. In Sect. 2, we introduce asymptotic properties of the distance-based classifier for high-dimensional data. We discuss the distance-based classifier in the SSE model. In Sect. 3, we consider a distance-based classifier using eigenstructures for the SSE model. In Sect. 4, we discuss estimation of the eigenvalues and eigenvectors for the SSE model. We create a new distance-based classifier by estimating the eigenstructures. In Sect. 5, we give simulation studies and discuss the performance of the new classifier. Finally, we demonstrate the new classifier by using microarray data sets.

2 Distance-based classifier for high-dimensional data

In this section, we introduce asymptotic properties of the distance-based classifier for high-dimensional data. As for any positive-semidefinite matrix $\boldsymbol{M}$, we write the square root of $\boldsymbol{M}$ as $\boldsymbol{M}^{1/2}$. Let

$$\boldsymbol{x}_{ij} = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2} \boldsymbol{z}_{ij} + \boldsymbol{\mu}_i,$$

where $\boldsymbol{z}_{ij} = (z_{ij(1)}, ..., z_{ij(p)})^T$ is considered as a sphered data vector having the zero mean vector and identity covariance matrix. Similar to Bai and Saranadasa (1996) and Chen and Qin (2010), we assume the following assumption for π_i, $i = 1, 2$, as necessary:

$$\text{(A-i)} \quad \limsup_{p\to\infty} E(z_{ij(r)}^4) < \infty \text{ for all } r, \quad E(z_{ij(r)}^2 z_{ij(s)}^2) = E(z_{ij(r)}^2) E(z_{ij(s)}^2) = 1,$$

$$E(z_{ij(r)} z_{ij(s)} z_{ij(t)}) = 0 \text{ and } E(z_{ij(r)} z_{ij(s)} z_{ij(t)} z_{ij(u)}) = 0 \text{ for all } r \neq s, t, u.$$

When the π_is are Gaussian, (A-i) naturally holds. Let

$$\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \quad \Delta = \|\boldsymbol{\mu}\|^2, \quad n_{\min} = \min\{n_1, n_2\} \text{ and } m = \min\{p, n_{\min}\},$$

where $\|\cdot\|$ denotes the Euclidean norm. Note that $E\{W(\boldsymbol{x}_0)\} = (-1)^i \Delta/2$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Also, note that the divergence condition "$p \to \infty$, $n_1 \to \infty$ and $n_2 \to \infty$" is equivalent to "$m \to \infty$". Let

$$\delta_{oi} = \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)}{n_{i'}} + \sum_{l=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_l^2)}{2n_l(n_l - 1)} \right\}^{1/2}$$

and $\delta_i = \{\delta_{oi}^2 + \boldsymbol{\mu}^T(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_{i'}/n_{i'})\boldsymbol{\mu}\}^{1/2}$ for $i = 1, 2$; $i' \neq i$. Note that $\delta_i^2 = \mathrm{Var}\{W(\boldsymbol{x}_0)\}$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$.

Let $e(i)$ denote the error rate of misclassifying an individual from π_i into the other class for $i = 1, 2$. Then, for the classification rule (3) DBDA, Aoshima and Yata (2014) gave the following result.

Theorem 1 (Aoshima and Yata 2014) *Assume the following conditions:*

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

(AY-ii) $\dfrac{\max_{i=1,2} tr(\boldsymbol{\Sigma}_i^2)}{n_{\min}\Delta^2} \to 0$ *as* $m \to \infty$.

Then, for DBDA, we have that as $m \to \infty$

$$e(i) \to 0 \text{ for } i = 1, 2. \tag{7}$$

Remark 2 For DBDA, under (AY-i) and (AY-ii), one may write (7) as

$$e(i) = O(\delta_i^2/\Delta^2) \text{ for } i = 1, 2.$$

Next, we consider the asymptotic normality of the classifier. Hereafter, for a function, $f(\cdot)$, "$f(p) \in (0, \infty)$ as $p \to \infty$" implies $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$. Let "$\Rightarrow$" denote the convergence in distribution, let $N(0, 1)$ denote a random variable distributed as the standard normal distribution and let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution. Aoshima and Yata (2014) gave the following result.

Theorem 2 (Aoshima and Yata 2014) *Assume the following conditions:*

(AY-iii) $\dfrac{\boldsymbol{\mu}^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}}{\delta_{oi}^2} \to 0$ *as* $m \to \infty$, $\liminf_{p\to\infty} \dfrac{tr(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)}{tr(\boldsymbol{\Sigma}_i^2)} > 0$ *for* $i = 1, 2$, *and* $\dfrac{tr(\boldsymbol{\Sigma}_1^2)}{tr(\boldsymbol{\Sigma}_2^2)} \in (0, \infty)$ *as* $p \to \infty$.

Assume also the NSSE condition (2)*. Under a certain assumption milder than (A-i), it holds that as* $m \to \infty$

$$\frac{W(\boldsymbol{x}_0) - (-1)^i \Delta/2}{\delta_{oi}} \Rightarrow N(0,1) \ \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Furthermore, for DBDA, it holds that as $m \to \infty$

$$e(i) - \Phi\left(\frac{-\Delta}{2\delta_{oi}}\right) = o(1) \ \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2. \tag{8}$$

Remark 3 Aoshima and Yata (2015b) gave a different asymptotic normality from Theorem 2 under different conditions. From the facts that $\delta_{oi}/\delta_i \to 1$ as $m \to \infty$ under (AY-iii) and $\mathrm{Var}\{W(\boldsymbol{x}_0)\} = \delta_i^2$ when $\boldsymbol{x}_0 \in \pi_i$, one may write (8) as

$$e(i) - \Phi\{-\Delta/(2\delta_i)\} = o(1) \ \text{ when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2.$$

By using the asymptotic normality, Aoshima and Yata (2014) proposed *the misclassification rate-adjusted classifier (MRAC)* in high-dimensional settings.

In this paper, we consider the distance-based classifier from a different point of view. We consider the classifier under the SSE model. We emphasize that high-dimensional data often have the SSE model. See Table 1, Figs. 1 and 2. If the SSE condition (4) is met, one cannot claim the asymptotic normality in Theorem 2. In addition, if the SSE condition (4) is met, (AY-ii) in Theorem 1 is equivalent to

$$\lambda_{i(1)}^2/(n_{\min}\Delta^2) = o(1) \quad \text{for } i = 1, 2. \tag{9}$$

Thus (AY-ii) in the SSE model is stricter than that in the NSSE model. For example, for the NSSE model as the spiked model in (6) with $\alpha_{i(1)} < 1/2$, $i = 1, 2$, (AY-ii) is equivalent to $p/(n_{\min}\Delta^2) = o(1)$. On the other hand, for the SSE model as (6) with $\alpha_{i(1)} > 1/2$ (and $\alpha_{i(1)} \geq \alpha_{i'(1)}$ for $i' \neq i$), (AY-ii) is equivalent to $p^{2\alpha_{i(1)}}/(n_{\min}\Delta^2) = o(1)$. That means $n_{\min}$ or Δ should be quite large for the SSE model compared to the NSSE model. Thus if the SSE condition (4) is met, DBDA has the classification consistency (7) under (AY-i) in Theorem 1 and the strict condition (9). In order to overcome the difficulties, we propose a new distance-based classifier by estimating eigenstructures for the SSE model.

3 Distance-based classifier using eigenstructures

Let

$$\Psi_{i(r)} = \mathrm{tr}(\boldsymbol{\Sigma}_i^2) - \sum_{s=1}^{r-1} \lambda_{i(s)}^2 = \sum_{s=r}^{p} \lambda_{i(s)}^2 \quad \text{for } i = 1, 2;\ r = 1, \ldots, p.$$

In this section, similar to Aoshima and Yata (2018), we assume the following model for $i = 1, 2$:

(M-i) There exists a fixed integer k_i (≥ 1) such that $\lambda_{i(1)}, ..., \lambda_{i(k_i)}$ are distinct in the sense that $\liminf_{p\to\infty}(\lambda_{i(r)}/\lambda_{i(s)} - 1) > 0$ when $1 \leq r < s \leq k_i$, and $\lambda_{i(k_i)}$ and $\lambda_{i(k_i+1)}$ satisfy

$$\liminf_{p\to\infty} \frac{\lambda^2_{i(k_i)}}{\Psi_{i(k_i)}} > 0 \ \text{ and } \ \frac{\lambda^2_{i(k_i+1)}}{\Psi_{i(k_i+1)}} \to 0 \ \text{ as } p \to \infty.$$

Note that (M-i) implies the SSE condition (4), that is (M-i) is one of the SSE models. For example, (M-i) holds in the spiked model in (6) with

$$\alpha_{i(1)} \geq \cdots \geq \alpha_{i(k_i)} \geq 1/2 > \alpha_{i(k_i+1)} \geq \cdots \geq \alpha_{i(t_i)} \ \text{ and } \ a_{i(r)} \neq a_{i(s)}$$

for $1 \leq r < s \leq k_i$; $i = 1, 2$. We emphasize that (M-i) is a natural model under the SSE condition (4). See Fig. 2. The six microarray data appear to be consistent with (M-i). Similar to (9), we note that the sufficient condition (AY-ii) in Theorem 1 is equivalent to

$$\sum_{r=1}^{k_i} \lambda^2_{i(r)}/(n_{\min}\Delta^2) = o(1) \quad \text{for } i = 1, 2 \tag{10}$$

under (M-i). According to the arguments in the last paragraph of Sect. 2, if (M-i) is met, DBDA has the classification consistency (7) under (AY-i) in Theorem 1 and the strict condition (10). Also, one cannot claim the asymptotic normality in Theorem 2 under (M-i). In order to overcome the difficulties, along the lines of Aoshima and Yata (2018), we consider a data transformation from the SSE model to the NSSE model in this section.

Let us see a toy example of the model (M-i) such as

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2, \ \ \lambda_{i(1)} = p \ \text{ and } \ \lambda_{i(2)} = \cdots = \lambda_{i(p)} = 1.$$

Then, the sufficient condition (10) is equivalent to "$p^2/(n_{\min}\Delta^2) = o(1)$". Now, we consider the following data transformation, based on the first eigenvector of $\boldsymbol{\Sigma}_i$, so as to avoid the strongly spiked eigenspace. We transform $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s into $(\boldsymbol{I}_p - \boldsymbol{h}_{i(1)}\boldsymbol{h}^T_{i(1)})\boldsymbol{x}_0$ ($= \boldsymbol{x}_{0,h}$, say) and $(\boldsymbol{I}_p - \boldsymbol{h}_{i(1)}\boldsymbol{h}^T_{i(1)})\boldsymbol{x}_{ij}$ ($= \boldsymbol{x}_{ij,h}$, say), respectively. Note that $\mathrm{Var}(\boldsymbol{x}_{0,h}) = \mathrm{Var}(\boldsymbol{x}_{ij,h}) = \sum_{r=2}^{p} \lambda_{i(r)}\boldsymbol{h}_{i(r)}\boldsymbol{h}^T_{i(r)}$ ($= \boldsymbol{\Sigma}_{i,h}$, say) when $\boldsymbol{x}_0 \in \pi_i$. Let $\Delta_h = \|E(\boldsymbol{x}_{1j,h}) - E(\boldsymbol{x}_{2j',h})\|^2$. Then, $\Delta_h = \Delta - (\boldsymbol{h}^T_{i(1)}\boldsymbol{\mu})^2$. From Theorem 1, DBDA based on the transformed data has the classification consistency (7) under

$$\boldsymbol{\mu}^T\boldsymbol{\Sigma}_{i,h}\boldsymbol{\mu}/\Delta_h^2 = o(1), \ \ i = 1, 2, \ \text{ and } \ p/(n_{\min}\Delta_h^2) = o(1)$$

because $\mathrm{tr}(\boldsymbol{\Sigma}^2_{i,h}) = O(p)$. We note that $\liminf_{p\to\infty} \Delta_h/\Delta > 0$ in a natural situation where $\limsup_{p\to\infty} |\boldsymbol{h}^T_{i(1)}\boldsymbol{\mu}/\Delta^{1/2}| < 1$. In that sense, "$\boldsymbol{\mu}^T\boldsymbol{\Sigma}_{i,h}\boldsymbol{\mu}/\Delta_h^2 = o(1)$" is milder than (AY-i). Also, "$p/(n_{\min}\Delta_h^2) = o(1)$" is much milder than "$p^2/(n_{\min}\Delta^2) = o(1)$" which is equivalent to (AY-ii). Therefore, the above data transformation probably improves the classification consistency (7). This is a reason why we consider such a

data transformation. In Sect. 3.1, we give a general data transformation for the model (M-i).

3.1 Data transformation

Recall that $\boldsymbol{h}_{i(r)}$ is the r-th eigenvector of $\boldsymbol{\Sigma}_i$. Let

$$\boldsymbol{A}_i = \boldsymbol{I}_p - \sum_{r=1}^{k_i} \boldsymbol{h}_{i(r)}\boldsymbol{h}_{i(r)}^T = \sum_{r=k_i+1}^{p} \boldsymbol{h}_{i(r)}\boldsymbol{h}_{i(r)}^T \quad \text{and} \quad \boldsymbol{x}_{ij,A} = \boldsymbol{A}_i\boldsymbol{x}_{ij}$$

for $j = 1, ..., n_i;\ i = 1, 2$. Note that $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ for $i = 1, 2$. Let us write that $\boldsymbol{\mu}_{i,A} = \boldsymbol{A}_i\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_{i,A} = \boldsymbol{A}_i\boldsymbol{\Sigma}_i\boldsymbol{A}_i = \sum_{r=k_i+1}^{p} \lambda_{i(r)}\boldsymbol{h}_{i(r)}\boldsymbol{h}_{i(r)}^T$, $i = 1, 2$, $\boldsymbol{\mu}_A = \boldsymbol{\mu}_{1,A} - \boldsymbol{\mu}_{2,A}$ and $\Delta_A = \|\boldsymbol{\mu}_A\|^2$. Note that $E(\boldsymbol{x}_{ij,A}) = \boldsymbol{\mu}_{i,A}$ and $\mathrm{Var}(\boldsymbol{x}_{ij,A}) = \boldsymbol{\Sigma}_{i,A}$ for all i, j. Thus the transformed data, $\boldsymbol{x}_{ij,A}$, have the NSSE model in the sense that

$$\{\lambda_{\max}(\boldsymbol{\Sigma}_{i,A})\}^2/\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2) = \lambda_{i(k_i+1)}^2/\Psi_{i(k_i+1)} \to 0 \ \text{ as } p \to \infty,$$

where $\lambda_{\max}(\boldsymbol{M})$ denotes the largest eigenvalue of any positive-semidefinite matrix, $\boldsymbol{M}$. Hence, we can say that a classifier based on the transformed data satisfies the classification consistency (7) under mild conditions and provided (M-i) is satisfied. In addition, one can claim the asymptotic normality of the classifier even when the SSE condition (4) is met.

Now, we propose the classifier by using the transformed data. Let us write that $\boldsymbol{A}_* = (\boldsymbol{A}_1 + \boldsymbol{A}_2)/2$, $\boldsymbol{x}_{0,A*} = \boldsymbol{A}_*\boldsymbol{x}_0$ and $\overline{\boldsymbol{x}}_{i,A} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij,A}/n_i = \boldsymbol{A}_i\overline{\boldsymbol{x}}_i$ for $i = 1, 2$. We consider the following classifier:

$$\begin{aligned} W_A(\boldsymbol{x}_0) &= \left(\boldsymbol{x}_{0,A*} - \frac{\overline{\boldsymbol{x}}_{1,A} + \overline{\boldsymbol{x}}_{2,A}}{2}\right)^T (\overline{\boldsymbol{x}}_{2,A} - \overline{\boldsymbol{x}}_{1,A}) - \frac{\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{S}_1)}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{A}_2\boldsymbol{S}_2)}{2n_2} \\ &= \boldsymbol{x}_{0,A*}^T(\overline{\boldsymbol{x}}_{2,A} - \overline{\boldsymbol{x}}_{1,A}) + \sum_{j<j'}^{n_1} \frac{\boldsymbol{x}_{1j,A}^T\boldsymbol{x}_{1j',A}}{n_1(n_1-1)} - \sum_{j<j'}^{n_2} \frac{\boldsymbol{x}_{2j,A}^T\boldsymbol{x}_{2j',A}}{n_2(n_2-1)}. \end{aligned} \tag{11}$$

Then, one classifies $\boldsymbol{x}_0$ into π_1 if $W_A(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. Let $\boldsymbol{A}_{1,2} = \boldsymbol{A}_1 - \boldsymbol{A}_2$. Here, let us write that $\boldsymbol{\Sigma}_{i,A*} = \boldsymbol{A}_*\boldsymbol{\Sigma}_i\boldsymbol{A}_*$,

$$\delta_{oi,A} = \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{i,A})}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{i',A})}{n_{i'}} + \sum_{l=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)}{2n_l(n_l-1)} \right\}^{1/2};$$

$$\text{and } \delta_{i,A} = \left\{ \delta_{oi,A}^2 + \boldsymbol{\mu}_A^T\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\mu}_A + \boldsymbol{\mu}_i^T\boldsymbol{A}_{1,2}\boldsymbol{\Sigma}_{i,A}\boldsymbol{A}_{1,2}\boldsymbol{\mu}_i/(4n_i) \right.$$
$$\left. + (\boldsymbol{\mu}_A - \boldsymbol{A}_{1,2}\boldsymbol{\mu}_i/2)^T\boldsymbol{\Sigma}_{i',A}(\boldsymbol{\mu}_A - \boldsymbol{A}_{1,2}\boldsymbol{\mu}_i/2)/n_{i'} \right\}^{1/2}$$

for $i = 1, 2;\ i' \neq i$. Then, we claim that when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$,

$$E\{W_A(\boldsymbol{x}_0)\} = (-1)^i \frac{\Delta_A}{2} - (-1)^i \frac{\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2} \boldsymbol{\mu}_A}{2} \ \text{ and } \ \mathrm{Var}\{W_A(\boldsymbol{x}_0)\} = \delta_{i,A}^2. \tag{12}$$

Remark 4 In general, $\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2} \boldsymbol{\mu}_A$ in (12) is not sufficiently large because of $\mathrm{rank}(\boldsymbol{A}_{1,2}) \leq k_1 + k_2\ (< \infty)$. If $\boldsymbol{A}_1 = \boldsymbol{A}_2$, it holds that $E\{W_A(\boldsymbol{x}_0)\} = (-1)^i \Delta_A/2$ and

$$\mathrm{Var}\{W_A(\boldsymbol{x}_0)\} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A} \boldsymbol{\Sigma}_{2,A})}{n_{i'}} + \sum_{l=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)}{2n_l(n_l - 1)} + \boldsymbol{\mu}_A^T (\boldsymbol{\Sigma}_{i,A} + \boldsymbol{\Sigma}_{i',A}/n_{i'}) \boldsymbol{\mu}_A$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2;\ i' \neq i$.

In Sects. 3.2 and 3.3, we give consistency properties and an asymptotic normality of $W_A(\boldsymbol{x}_0)$. We assume the following conditions as necessary:

(C-i) $\dfrac{\boldsymbol{\mu}_A^T (\boldsymbol{\Sigma}_{i,A*} + \boldsymbol{\Sigma}_{i',A}/n_{i'}) \boldsymbol{\mu}_A}{\Delta_A^2} \to 0$ as $p \to \infty$ for $i = 1, 2;\ i' \neq i$;

(C-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*} \boldsymbol{\Sigma}_{l,A})}{n_l \Delta_A^2} \to 0$ as $m \to \infty$ for $i, l = 1, 2$;

(C-iii) $\dfrac{\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2} \boldsymbol{\mu}_A}{\Delta_A} \to 0$ as $p \to \infty$ and $\limsup_{m \to \infty} \dfrac{\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2}^2 \boldsymbol{\mu}_i}{n_{\min}^{1/2} \Delta_A} < \infty$ for $i = 1, 2$;

(C-iv) $\dfrac{\boldsymbol{\mu}_A^T (\boldsymbol{\Sigma}_{i,A*} + \boldsymbol{\Sigma}_{i',A}/n_{i'}) \boldsymbol{\mu}_A}{\delta_{oi,A}^2} \to 0$ as $m \to \infty$, $\liminf_{p \to \infty} \dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A} \boldsymbol{\Sigma}_{2,A})}{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2)} > 0$ for $i = 1, 2\ (i' \neq i)$, and $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A}^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_{2,A}^2)} \in (0, \infty)$ as $p \to \infty$;

(C-v) $\dfrac{\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2} \boldsymbol{\mu}_A}{\delta_{oi,A}} \to 0$ as $m \to \infty$, $\limsup_{m \to \infty} \dfrac{\boldsymbol{\mu}_i^T \boldsymbol{A}_{1,2}^2 \boldsymbol{\mu}_i}{n_{\min}^{1/2} \delta_{oi,A}} < \infty$, and $\dfrac{\lambda_{\max}(\boldsymbol{\Sigma}_{i,A*}^{1/2} \boldsymbol{\Sigma}_{l,A} \boldsymbol{\Sigma}_{i,A*}^{1/2})}{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*} \boldsymbol{\Sigma}_{l,A})} \to 0$ as $p \to \infty$ for $i, l = 1, 2$.

3.2 Consistency of the classifier (11)

We consider consistency properties of $W_A(\boldsymbol{x}_0)$. We note that $\delta_{i,A}^2/\Delta_A^2 \to 0$ as $m \to \infty$ under (C-i) to (C-iii). See Sect. 6.1. Then, we have the following results.

Theorem 3 *Assume (M-i). Assume also (C-i) to (C-iii). Then, it holds that as* $m \to \infty$

$$\frac{W_A(\boldsymbol{x}_0)}{\Delta_A} = \frac{(-1)^i}{2} + o_P(1) \ \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

For the classification rule (11)*, we have the classification consistency* (7) *as* $m \to \infty$.

Corollary 1 *If $\boldsymbol{A}_1 = \boldsymbol{A}_2$, for the classification rule (11), we have the classification consistency (7) as $m \to \infty$ under (M-i) and the following conditions:*

$$\frac{\boldsymbol{\mu}_A^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_A}{\Delta_A^2} \to 0 \ \text{as } p \to \infty \ \text{and} \ \frac{tr(\boldsymbol{\Sigma}_{i,A}^2)}{n_{\min} \Delta_A^2} \to 0 \ \text{as } m \to \infty \ \text{for } i = 1, 2.$$

Remark 5 For the classification rule (11), under (M-i) and (C-i) to (C-iii), one may write (7) as

$$e(i) = O(\delta_{i,A}^2 / \Delta_A^2) \ \text{ for } i = 1, 2.$$

Now, we consider the sufficient condition (C-ii) in Theorem 3. When $\lambda_{i(1)}^2 / \mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2) \to \infty$ as $p \to \infty$ for $i = 1, 2$, it holds that

$$\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*} \boldsymbol{\Sigma}_{l,A}) \le \{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}^2) \mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)\}^{1/2} = o\left[\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \mathrm{tr}(\boldsymbol{\Sigma}_l^2)\}^{1/2}\right]$$

for $i, l = 1, 2$, from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}^2) \le \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$. Then, (C-ii) is milder than (AY-ii) if Δ and Δ_A are of the same order.

3.3 Asymptotic normality of the classifier (11)

We consider the asymptotic normality of $W_A(\boldsymbol{x}_0)$. We have the following results.

Theorem 4 *Assume (A-i) and (M-i). Assume also (C-iv) and (C-v). Then, it holds that as $m \to \infty$*

$$\frac{W_A(\boldsymbol{x}_0) - (-1)^i \Delta_A / 2}{\delta_{oi,A}} \Rightarrow N(0, 1) \ \text{ when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2.$$

Furthermore, for the classification rule (11), it holds that as $m \to \infty$

$$e(i) - \Phi\left(\frac{-\Delta_A}{2\delta_{oi,A}}\right) = o(1) \ \text{ when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2. \tag{13}$$

Corollary 2 *If $\boldsymbol{A}_1 = \boldsymbol{A}_2$, for the classification rule (11), (13) holds as $m \to \infty$ under (A-i), (M-i) and the following conditions:*

$$\frac{\boldsymbol{\mu}_A^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_A}{\delta_{oi,A}^2} \to 0 \ \text{as } m \to \infty, \ \liminf_{p \to \infty} \frac{tr(\boldsymbol{\Sigma}_{1,A} \boldsymbol{\Sigma}_{2,A})}{tr(\boldsymbol{\Sigma}_{i,A}^2)} > 0 \ \text{for } i = 1, 2;$$
$$\text{and} \ \frac{tr(\boldsymbol{\Sigma}_{1,A}^2)}{tr(\boldsymbol{\Sigma}_{2,A}^2)} \in (0, \infty) \ \text{as } p \to \infty.$$

Remark 6 From (30) in Sect. 6, we note that $\delta_{oi,A}/\delta_{i,A} \to 1$ as $m \to \infty$ under (C-iv) and (C-v). Hence, one may write (13) as

$$e(i) - \Phi\{-\Delta_A/(2\delta_{i,A})\} = o(1) \ \text{ when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2.$$

Now, let us show an easy example to check the performance of DBDA and the classifier (11) for the SSE model. We considered the following setting:

(S-i) We set $p = 2^s$, $s = 5, ..., 13$, and $n_1 = \lceil p^{2/5} \rceil$ and $n_2 = 2n_1$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. 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 $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1)^T$ whose last $\lceil p^{1/2} \rceil$ elements are 1, $\boldsymbol{\Sigma}_1 = \mathrm{diag}(p^{2/3}, p^{1/2}, 1, ..., 1)$ and $\boldsymbol{\Sigma}_2 = 2\boldsymbol{\Sigma}_1$.

We note that (A-i), (M-i), (AY-i) to (AY-iii) and (C-i) to (C-v) are met for (S-i) from the facts that $\Delta = \Delta_A = \lceil p^{1/2} \rceil$ and $\boldsymbol{A}_1 = \boldsymbol{A}_2$ with $k_1 = k_2 = 2$, so that Theorems 1, 3 and 4 hold. However, the NSSE condition (2) is not met, so that Theorem 2 does not hold. In general, $\boldsymbol{A}_i$s are unknown in (11). Hence, we considered a naive estimator of $\boldsymbol{A}_i$ as $\widehat{\boldsymbol{A}}_i = \boldsymbol{I}_p - \sum_{r=1}^{k_i} \hat{\boldsymbol{h}}_{i(r)} \hat{\boldsymbol{h}}_{i(r)}^T$ and checked the performance of the classifier given by

$$\begin{aligned} \widehat{W}_A(\boldsymbol{x}_0) = &- \left\{\widehat{\boldsymbol{A}}_1(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{x}_0) + \widehat{\boldsymbol{A}}_2(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{x}_0)\right\}^T \left(\widehat{\boldsymbol{A}}_2\overline{\boldsymbol{x}}_2 - \widehat{\boldsymbol{A}}_1\overline{\boldsymbol{x}}_1\right)/2 \\ &- \mathrm{tr}(\widehat{\boldsymbol{A}}_1\boldsymbol{S}_1)/(2n_1) + \mathrm{tr}(\widehat{\boldsymbol{A}}_2\boldsymbol{S}_2)/(2n_2). \end{aligned} \tag{14}$$

Here, $\hat{\boldsymbol{h}}_{i(r)}$ denotes the r-th (unit) eigenvector of $\boldsymbol{S}_i$ for each i, r. Then, one classifies $\boldsymbol{x}_0$ into π_1 if $\widehat{W}_A(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. On the other hand, by using a bias-corrected estimator of the eigenstructures, we create a new distance-based classifier given by (21) in Sect. 4. We also checked the performance of the new classification rule (21). We call the classification rule (21) the "transformed distance-based discriminant analysis (T-DBDA)". We also describe the classification rule (11) as "T-DBDA before estimation (T-DBDA(b))" and the classification rule (14) as "T-DBDA by the naive estimator (T-DBDA(n))". For $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$ we calculated each classifier 2000 times to confirm if 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 the error rates, $\overline{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Their standard deviations are less than 0.011. In Fig. 3, we plotted $\overline{e}(1)$ and $\overline{e}(2)$ for DBDA, T-DBDA(n), T-DBDA(b) and T-DBDA. From Theorems 2 and 4 in view of Remarks 3 and 6, we also plotted the asymptotic error rates, $\Phi\{-\Delta/(2\delta_i)\}$ $(= \dot{e}(i),$ say) and $\Phi\{-\Delta_A/(2\delta_{i,A})\}$ $(= \dot{e}_A(i),$ say), in Fig. 3.

We observed that $\overline{e}(i)$ by T-DBDA(b) behaves very close to the asymptotic error rate, $\Phi\{-\Delta_A/(2\delta_{i,A})\}$, as expected theoretically. However, $\overline{e}(i)$ by DBDA does not converge to $\Phi\{-\Delta/(2\delta_i)\}$. This is because the classifier does not claim the asymptotic normality in Theorem 2 for the SSE model. Both DBDA and T-DBDA(b) have the classification consistency (7). However, T-DBDA(b) gave a much better performance than DBDA. This is probably due to the convergence rates. For the sufficient conditions in Theorems 1 and 3, we note that

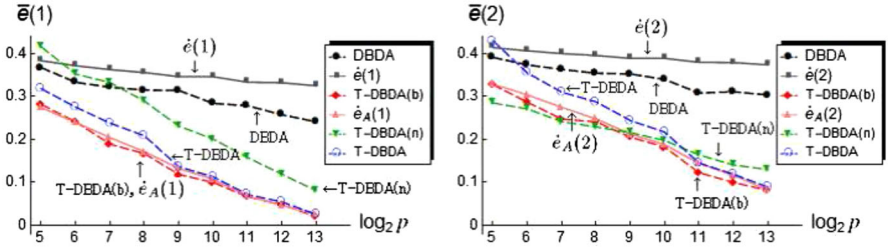


Fig. 3 The left panel displays $\bar{e}(1)$ and the right panel displays $\bar{e}(2)$. The error rates (dashed lines) of DBDA (the classifier (3)), T-DBDA(b) (the classifier (11)), T-DBDA(n) (the classifier (14)) and T-DBDA (the classifier (21)). The asymptotic error rates (solid lines) by $\dot{e}(i)$ $(= \Phi\{-\Delta/(2\delta_i)\})$ and $\dot{e}_A(i)$ $(= \Phi\{-\Delta_A/(2\delta_{i,A})\})$

$$\max_{i=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/(n_{\min}\Delta^2) = O(p^{1/3}/n_{\min}) = O(p^{-1/15}) \text{ in (AY-ii);}$$
$$\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{l,A})/(n_l\Delta_A^2) = O(n_l^{-1}) = O(p^{-2/5}) \text{ for } i,l = 1,2, \text{ in (C-ii).}$$

Hence, the error rates of T-DBDA(b) were smaller than those of DBDA. The T-DBDA(n) gave a worse performance than T-DBDA(b). This is probably because of the bias caused by the naive estimator, $\widehat{\boldsymbol{A}}_i$. See Sect. 4.1 for the details. Hence, we will consider a bias-correction of the naive estimator in Sect. 4. On the other hand, the performances of T-DBDA and T-DBDA(b) became similar to each other when p is large. We will discuss T-DBDA in Sect. 4.2.

In Sect. 4, we discuss estimation of the unknown parameters in (11). We create T-DBDA by the bias-corrected estimator of the parameters.

4 Distance-based classifier by estimating eigenstructures

In this section, we assume (A-i) and (M-i). Let $x_{0,i(r)} = \boldsymbol{x}_0^T\boldsymbol{h}_{i(r)}$ and

$$x_{ij(r)} = \boldsymbol{x}_{ij}^T\boldsymbol{h}_{i(r)} = \lambda_{i(r)}^{1/2}z_{ij(r)} + \mu_{i(r)} \text{ for all } i,j,r, \text{ where } \mu_{i(r)} = \boldsymbol{\mu}_i^T\boldsymbol{h}_{i(r)}.$$

Let us write that $\bar{x}_{i(r)} = \sum_{j=1}^{n_i} x_{ij(r)}/n_i$ for all i,r. Then, one can write (11) as follows:

$$\begin{aligned} W_A(\boldsymbol{x}_0) =& W(\boldsymbol{x}_0) + \sum_{r=1}^{k_1} x_{0,1(r)} \left\{ \bar{x}_{1(r)} - \frac{1}{2}\boldsymbol{h}_{1(r)}^T \left(\overline{\boldsymbol{x}}_2 - \sum_{s=1}^{k_2} \bar{x}_{2(s)}\boldsymbol{h}_{2(s)} \right) \right\} \\ & - \sum_{r=1}^{k_2} x_{0,2(r)} \left\{ \bar{x}_{2(r)} - \frac{1}{2}\boldsymbol{h}_{2(r)}^T \left(\overline{\boldsymbol{x}}_1 - \sum_{s=1}^{k_1} \bar{x}_{1(s)}\boldsymbol{h}_{1(s)} \right) \right\} \\ & - \sum_{r=1}^{k_1} \frac{\sum_{j<j'}^{n_1} x_{1j(r)}x_{1j'(r)}}{n_1(n_1-1)} + \sum_{r=1}^{k_2} \frac{\sum_{j<j'}^{n_2} x_{2j(r)}x_{2j'(r)}}{n_2(n_2-1)}. \end{aligned} \tag{15}$$

In order to use $W_A(\boldsymbol{x}_0)$, it is necessary to estimate $\boldsymbol{h}_{i(r)}$s, $x_{0,i(r)}$s, $x_{ij(r)}$s and k_is.

Let $\delta_{o\min,A} = \min\{\delta_{o1,A}, \delta_{o2,A}\}$. In this section, we assume the following conditions as necessary:

(C-vi) $\limsup_{p\to\infty} \left(\sum_{r=1}^{k_i} \frac{\boldsymbol{h}_{i(r)}^T \boldsymbol{\Sigma}_{i'} \boldsymbol{h}_{i(r)}}{\lambda_{i(r)}} \right) < \infty$ for $i = 1, 2$ $(i' \neq i)$;

(C-vii) $\limsup_{m\to\infty} \left(\sum_{r=1}^{k_i} \frac{n_i\{\mu_{i(r)}^2 + (\boldsymbol{\mu}_{i'}^T \boldsymbol{h}_{i(r)})^2\}}{\lambda_{i(r)}} \right) < \infty$, $\limsup_{m\to\infty} \frac{\lambda_{l(1)}}{n_i \lambda_{i(k_i)}} < \infty$,
and $\limsup_{m\to\infty} \left(\frac{\boldsymbol{\mu}_{l,A}^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_{l,A}}{\lambda_{i(k_i)}^2} \right) < \infty$ for $i, l = 1, 2$ $(i' \neq i)$;

(C-viii) $\frac{\lambda_{i(1)}}{n_{\min} \Delta_A} \to 0$ and $\frac{\boldsymbol{\mu}_{i,A}^T (\boldsymbol{\Sigma}_{i,A}/n_i + \boldsymbol{\Sigma}_{i',A}/n_{i'}) \boldsymbol{\mu}_{i,A}}{\Delta_A^2} \to 0$ as $m \to \infty$
for $i = 1, 2$ $(i' \neq i)$;

(C-ix) $\frac{\lambda_{i(1)}}{n_{\min} \delta_{o\min,A}} \to 0$ and $\frac{\boldsymbol{\mu}_{i,A}^T (\boldsymbol{\Sigma}_{i,A}/n_i + \boldsymbol{\Sigma}_{i',A}/n_{i'}) \boldsymbol{\mu}_{i,A}}{\delta_{o\min,A}^2} \to 0$ as $m \to \infty$
for $i = 1, 2$ $(i' \neq i)$.

4.1 Estimation of $\boldsymbol{h}_{i(r)}$s, $\boldsymbol{x}_{0,i(r)}$s and $\boldsymbol{x}_{ij(r)}$s

Let $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in}]$, $\overline{\boldsymbol{X}}_i = [\overline{\boldsymbol{x}}_i, ..., \overline{\boldsymbol{x}}_i]$ and $\boldsymbol{P}_{n_i} = \boldsymbol{I}_{n_i} - \boldsymbol{1}_{n_i} \boldsymbol{1}_{n_i}^T / n_i$ for $i = 1, 2$, where $\boldsymbol{1}_{n_i} = (1, ..., 1)^T$. Note that $\boldsymbol{S}_i = \boldsymbol{X}_i \boldsymbol{P}_{n_i} \boldsymbol{X}_i^T / (n_i - 1) = (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T / (n_i - 1)$. We define the $n_i \times n_i$ dual sample covariance matrix by

$$\boldsymbol{S}_{iD} = \boldsymbol{P}_{n_i} \boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{P}_{n_i} / (n_i - 1) = (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i) / (n_i - 1) \quad \text{for } i = 1, 2.$$

Note that $\boldsymbol{S}_i$ and $\boldsymbol{S}_{iD}$ share nonzero eigenvalues. Let us write the eigen-decomposition of $\boldsymbol{S}_i$ and $\boldsymbol{S}_{iD}$ as

$$\boldsymbol{S}_i = \sum_{r=1}^{p} \hat{\lambda}_{i(r)} \hat{\boldsymbol{h}}_{i(r)} \hat{\boldsymbol{h}}_{i(r)}^T \quad \text{and} \quad \boldsymbol{S}_{iD} = \sum_{r=1}^{n_i - 1} \hat{\lambda}_{i(r)} \hat{\boldsymbol{u}}_{i(r)} \hat{\boldsymbol{u}}_{i(r)}^T \quad \text{for } i = 1, 2,$$

where $\hat{\boldsymbol{h}}_{i(r)}$ and $\hat{\boldsymbol{u}}_{i(r)}$ denote unit eigenvectors corresponding to $\hat{\lambda}_{i(r)}$. We assume $\boldsymbol{h}_{i(r)}^T \hat{\boldsymbol{h}}_{i(r)} \geq 0$ w.p.1 for all i, r without loss of generality. Note that $\hat{\boldsymbol{h}}_{i(r)}$ can be calculated by $\hat{\boldsymbol{h}}_{i(r)} = \{(n_i - 1)\hat{\lambda}_{i(r)}\}^{-1/2} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i) \hat{\boldsymbol{u}}_{i(r)}$. However, as observed in Sect. 3.2, the classifier by $\hat{\boldsymbol{h}}_{i(r)}$s gave an inadequate performance.

Yata and Aoshima (2012) proposed a bias-corrected eigenvalue estimation called the noise-reduction (NR) methodology, which was brought about by a geometric representation of $\boldsymbol{S}_{iD}$. If one applies the NR methodology, the $\lambda_{i(r)}$s are estimated by

$$\tilde{\lambda}_{i(r)} = \hat{\lambda}_{i(r)} - \frac{\mathrm{tr}(\boldsymbol{S}_{iD}) - \sum_{s=1}^{r} \hat{\lambda}_{i(s)}}{n_i - 1 - r} \quad (r = 1, ..., n_i - 2;\ i = 1, 2). \tag{16}$$

Note that $\tilde{\lambda}_{i(r)} \geq 0$ w.p.1 for $r = 1, ..., n_i - 2$ and the second term in (16) is an estimator of $\sum_{r=k_i+1}^{p} \lambda_{i(r)}/(n_i - 1)$ $(= \kappa_i,$ say). When applying the NR methodology to the PC direction vector, one obtains

$$\tilde{\boldsymbol{h}}_{i(r)} = \{(n_i - 1)\tilde{\lambda}_{i(r)}\}^{-1/2}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)\hat{\boldsymbol{u}}_{i(r)} \text{ for } r = 1, ..., n_i - 2;\ i = 1, 2. \quad (17)$$

For $(\hat{\lambda}_{i(r)}, \hat{\boldsymbol{h}}_{i(r)})$s and $(\tilde{\lambda}_{i(r)}, \tilde{\boldsymbol{h}}_{i(r)})$s, Aoshima and Yata (2018) gave the following results.

Proposition 1 (Aoshima and Yata 2018) *Assume (A-i) and (M-i). It holds as* $m \to \infty$

$$\frac{\hat{\lambda}_{i(r)}}{\lambda_{i(r)}} = 1 + \frac{\kappa_i}{\lambda_{i(r)}} + O_P(n_i^{-1/2}),\ \ (\boldsymbol{h}_{i(r)}^T \hat{\boldsymbol{h}}_{i(r)})^2 = \left(1 + \frac{\kappa_i}{\lambda_{i(r)}}\right)^{-1} + O_P(n_i^{-1/2}),$$
$$\frac{\tilde{\lambda}_{i(r)}}{\lambda_{i(r)}} = 1 + O_P(n_i^{-1/2}) \ \textit{and}\ \ (\boldsymbol{h}_{i(r)}^T \tilde{\boldsymbol{h}}_{i(r)})^2 = 1 + O_P(n_i^{-1})$$

for $r = 1, ..., k_i;\ i = 1, 2$.

If $\kappa_i/\lambda_{i(r)} \to \infty$ as $m \to \infty$, $\hat{\lambda}_{i(r)}$ and $\hat{\boldsymbol{h}}_{i(r)}$ are strongly inconsistent in the sense that $\lambda_{i(r)}/\hat{\lambda}_{i(r)} = o_P(1)$ and $\boldsymbol{h}_{i(r)}^T \hat{\boldsymbol{h}}_{i(r)} = o_P(1)$. For example, in (S-i), $\kappa_i/\lambda_{i(2)} \to \infty$ as $m \to \infty$, so that $\boldsymbol{h}_{i(2)}^T \hat{\boldsymbol{h}}_{i(2)} = o_P(1)$. This is the main reason why the classifier by (14) gave an inadequate performance in Fig. 3. On the other hand, $\tilde{\lambda}_{i(r)}$ and $\tilde{\boldsymbol{h}}_{i(r)}$ are consistent estimators even when $\kappa_i/\lambda_{i(r)} \to \infty$ as $m \to \infty$. We note that $\mathrm{tr}(\boldsymbol{S}_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)\{1 + o_P(1)\}$ as $m \to \infty$ for $i = 1, 2$, under (A-i) and (M-i) from the fact that $\mathrm{Var}\{\mathrm{tr}(\boldsymbol{S}_i)\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i\} = o\{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2\}$ under (A-i) and (M-i). Hence, from Proposition 1 we claim that as $m \to \infty$

$$\hat{\varepsilon}_{i(r)} = \varepsilon_{i(r)}\{1 + o_P(1)\} \text{ for } r = 1, ..., k_i;\ i = 1, 2, \quad (18)$$

under (A-i) and (M-i).

Next, we consider an estimation of $x_{0,i(r)}$. Let

$$\tilde{x}_{0,i(r)} = \boldsymbol{x}_0^T \tilde{\boldsymbol{h}}_{i(r)} \text{ for all } i, r. \quad (19)$$

Note that $\mathrm{Var}(x_{0,i(r)}) = O(\lambda_{i(r)})$ as $p \to \infty$ under (C-vi) when $\boldsymbol{x}_0 \in \pi_{i'}$ for $r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i$. Then, we have the following results.

Proposition 2 *Assume (A-i), (M-i) and (C-vi). Assume also* $\limsup_{p\to\infty}[\{tr(\boldsymbol{\Sigma}_{i,A}\boldsymbol{\Sigma}_{i'}) + \max_{l=1,2} \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_l\}/\lambda_{i(k_i)}^2] < \infty$ *and* $\limsup_{p\to\infty}(\sum_{r=1}^{k_i}\{\mu_{i(r)}^2 + (\boldsymbol{\mu}_{i'}^T \boldsymbol{h}_{i(r)})^2\}/\lambda_{i(r)}) < \infty$ *for* $i = 1, 2;\ i' \neq i$. *Then, it holds as* $m \to \infty$

$$\boldsymbol{x}_0^T \hat{\boldsymbol{h}}_{i(r)} = \frac{x_{0,i(r)}}{(1 + \kappa_i/\lambda_{i(r)})^{1/2}} + O_P\left\{(\lambda_{i(r)}/n_i)^{1/2}\right\}$$
$$\textit{and}\ \ \tilde{x}_{0,i(r)} = x_{0,i(r)} + O_P\left\{(\lambda_{i(r)}/n_i)^{1/2}\right\}$$

when $\boldsymbol{x}_0 \in \pi_l$ *for* $r = 1, \ldots, k_i;\ i, l = 1, 2$.

Thus one can estimate $x_{0,i(r)}$ by $\tilde{x}_{0,i(r)}$ even when $\kappa_i/\lambda_{i(r)} \to \infty$ as $m \to \infty$.

Finally, we consider estimating $x_{ij(r)}$. We note that $\boldsymbol{x}_{ij}^T\tilde{\boldsymbol{h}}_{i(r)}$ is biased for high-dimensional data. This is because $\boldsymbol{x}_{ij}^T\tilde{\boldsymbol{h}}_{i(r)}$ includes $\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2$ which is very biased for high-dimensional data. Now, we explain the main reason why the inner products involve the large bias terms. We note that $\mathbf{1}_{n_i}^T\hat{\boldsymbol{u}}_{i(r)}=0$ and $\boldsymbol{P}_{n_i}\hat{\boldsymbol{u}}_{i(r)}=\hat{\boldsymbol{u}}_{i(r)}$ when $\hat{\lambda}_{i(r)}>0$ since $\mathbf{1}_{n_i}^T\boldsymbol{S}_{iD}\mathbf{1}_{n_i}=0$. Also, note that

$$\{(n_i-1)\tilde{\lambda}_{i(r)}\}^{1/2}\tilde{\boldsymbol{h}}_{i(r)}=\boldsymbol{X}_{o,i}\boldsymbol{P}_{n_i}\hat{\boldsymbol{u}}_{i(r)}=\boldsymbol{X}_{o,i}\hat{\boldsymbol{u}}_{i(r)} \ \text{ when } \hat{\lambda}_{i(r)}>0,$$

where $\boldsymbol{X}_{o,i}=\boldsymbol{X}_i-\boldsymbol{\mu}_i\mathbf{1}_{n_i}^T$. Let us write that $\hat{\boldsymbol{u}}_{i(r)}=(\hat{u}_{i1(r)},\ldots,\hat{u}_{in_i(r)})^T$ for all i,r. Then, it holds that $\{(n_i-1)\tilde{\lambda}_{i(r)}\}^{1/2}\tilde{\boldsymbol{h}}_{i(r)}^T(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)=\hat{\boldsymbol{u}}_{i(r)}^T\boldsymbol{X}_{o,i}^T(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)=\hat{u}_{ij(r)}\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2+\sum_{l=1(\neq j)}^{n_i}\hat{u}_{il(r)}(\boldsymbol{x}_l-\boldsymbol{\mu}_i)^T(\boldsymbol{x}_j-\boldsymbol{\mu}_i)$, so that $\hat{u}_{ij(r)}\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2$ is strongly biased since $E(\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2)/(n_i-1)\geq\kappa_i$. In fact, $\kappa_i^{-1}=O(n_i/p)=o(1)$ as $m\to\infty$ for the spiked model in (6) under $n_i/p\to 0$. Hence, one should not apply the $\tilde{\boldsymbol{h}}_{i(r)}$s (or the $\hat{\boldsymbol{h}}_{i(r)}$s) to the estimation of $x_{ij(r)}$. See Section 5.1 in Aoshima and Yata (2018) for more details. We consider a bias-reduced estimation of $x_{ij(r)}$. We modify $\hat{\boldsymbol{u}}_{i(r)}$ as

$$\hat{\boldsymbol{u}}_{ij(r)}=(\hat{u}_{i1(r)},\ldots,\hat{u}_{ij-1(r)},-\hat{u}_{ij(r)}/(n_i-1),\hat{u}_{ij+1(r)},\ldots,\hat{u}_{in_i(r)})^T$$

whose j-th element is $-\hat{u}_{ij(r)}/(n_i-1)$ for all i,j,r. Note that $\sum_{j=1}^{n_i}\hat{\boldsymbol{u}}_{ij(r)}/n_i=\{(n_i-2)/(n_i-1)\}\hat{\boldsymbol{u}}_{i(r)}$. Let

$$\tilde{\boldsymbol{h}}_{ij(r)}=\frac{(n_i-1)^{1/2}(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)\hat{\boldsymbol{u}}_{ij(r)}}{(n_i-2)\tilde{\lambda}_{i(r)}^{1/2}} \ \text{ for all } i,j,r.$$

Then, it holds that $\sum_{j=1}^{n_i}\tilde{\boldsymbol{h}}_{ij(r)}/n_i=\tilde{\boldsymbol{h}}_{i(r)}$ and

$$\begin{aligned}&(n_i-2)\{\tilde{\lambda}_{i(r)}/(n_i-1)\}^{1/2}\tilde{\boldsymbol{h}}_{ij(r)}^T(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)\\&\quad=(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)^T\boldsymbol{X}_{o,i}\boldsymbol{P}_{n_i}\hat{\boldsymbol{u}}_{ij(r)}=\sum_{l=1(\neq j)}^{n_i}\Big(\hat{u}_{il(r)}+\frac{\hat{u}_{ij(r)}}{n_i-1}\Big)(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}_{il}-\boldsymbol{\mu}_i)\end{aligned}$$

when $\hat{\lambda}_{i(r)}>0$ from the fact that

$$\boldsymbol{P}_{n_i}\hat{\boldsymbol{u}}_{ij(r)}=(\hat{u}_{i1(r)},\ldots,\hat{u}_{ij-1(r)},0,\hat{u}_{ij+1(r)},\ldots,\hat{u}_{in_i(r)})^T+(n_i-1)^{-1}\hat{u}_{ij(r)}\mathbf{1}_{n_i(j)},$$

where $\mathbf{1}_{n_i(j)}=(1,\ldots,1,0,1,\ldots,1)^T$ whose j-th element is 0. Thus the large biased term, $\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2$, is removed. Let

$$\tilde{x}_{ij(r)}=\boldsymbol{x}_{ij}^T\tilde{\boldsymbol{h}}_{ij(r)} \ \text{ for all } i,j,r. \tag{20}$$

See Section 5.1 in Aoshima and Yata (2018) for theoretical comparisons between $\boldsymbol{x}_{ij}^T\hat{\boldsymbol{h}}_{i(r)}$, $\boldsymbol{x}_{ij}^T\tilde{\boldsymbol{h}}_{i(r)}$ and $\tilde{x}_{ij(r)}$.

4.2 Distance-based classifier by the NR methodology

Let $\overline{\tilde{x}}_{i(r)} = \sum_{j=1}^{n_i}\tilde{x}_{ij(r)}/n_i$ for all i, r. By combining (15) with (17), (19) and (20), we propose the following classifier:

$$\begin{aligned}\widetilde{W}_A(\boldsymbol{x}_0) = W(\boldsymbol{x}_0) &+ \sum_{r=1}^{k_1}\tilde{x}_{0,1(r)}\left\{\overline{\tilde{x}}_{1(r)} - \frac{1}{2}\tilde{\boldsymbol{h}}_{1(r)}^T\left(\overline{\boldsymbol{x}}_2 - \sum_{s=1}^{k_2}\overline{\tilde{x}}_{2(s)}\tilde{\boldsymbol{h}}_{2(s)}\right)\right\} \\ &- \sum_{r=1}^{k_2}\tilde{x}_{0,2(r)}\left\{\overline{\tilde{x}}_{2(r)} - \frac{1}{2}\tilde{\boldsymbol{h}}_{2(r)}^T\left(\overline{\boldsymbol{x}}_1 - \sum_{s=1}^{k_1}\overline{\tilde{x}}_{1(s)}\tilde{\boldsymbol{h}}_{1(s)}\right)\right\} \\ &- \sum_{r=1}^{k_1}\frac{\sum_{j<j'}^{n_1}\tilde{x}_{1j(r)}\tilde{x}_{1j'(r)}}{n_1(n_1-1)} + \sum_{r=1}^{k_2}\frac{\sum_{j<j'}^{n_2}\tilde{x}_{2j(r)}\tilde{x}_{2j'(r)}}{n_2(n_2-1)}. \end{aligned} \tag{21}$$

Then, one classifies $\boldsymbol{x}_0$ into π_1 if $\widetilde{W}_A(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. In general, k_is are unknown in $\widetilde{W}_A(\boldsymbol{x}_0)$. See Sect. 4.3 for estimation of k_is. We call the classification rule (21) the "transformed distance-based discriminant analysis (T-DBDA)".

Now, we give asymptotic properties of T-DBDA. We have the following results.

Theorem 5 *Assume (A-i) and (M-i). Assume also (C-i) to (C-iii) and (C-vi) to (C-viii). Then, it holds that as* $m \to \infty$

$$\frac{\widetilde{W}_A(\boldsymbol{x}_0)}{\Delta_A} = \frac{(-1)^i}{2} + o_P(1) \ \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

For T-DBDA, we have the classification consistency (7) *as* $m \to \infty$.

Theorem 6 *Assume (A-i) and (M-i). Assume also (C-iv) to (C-vii) and (C-ix). Then, it holds that as* $m \to \infty$

$$\frac{\widetilde{W}_A(\boldsymbol{x}_0) - (-1)^i\Delta_A/2}{\delta_{oi,A}} \Rightarrow N(0,1) \ \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Furthermore, for T-DBDA, (13) *holds as* $m \to \infty$.

Remark 7 From (C-viii) or (C-ix) T-DBDA depends on the scale of $\boldsymbol{\mu}_i$s in the sense that $\boldsymbol{\mu}_{i,A}^T\boldsymbol{\Sigma}_{l,A}\boldsymbol{\mu}_{i,A}$ for $i, l = 1, 2$. Hence, we recommend that one should apply the classifier to a mean-centered data in actual data analyses. See Sect. 5.2 for example.

In Fig. 3, as expected theoretically, we observed that $\overline{e}(i)$ for T-DBDA becomes close to that for T-DBDA(b) when p and n are large.

4.3 Estimation of k_is

In this section, we introduce an estimation of k_i given by Aoshima and Yata (2018).

Let $n_{i1} = \lceil n_i/2 \rceil$ and $n_{i2} = n_i - n_{i1}$. Let $\boldsymbol{X}_{i1} = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_{i1}}]$ and $\boldsymbol{X}_{i2} = [\boldsymbol{x}_{in_{i1}+1}, ..., \boldsymbol{x}_{in_i}]$ for $i = 1, 2$. We define

$$\boldsymbol{S}_{iD(1)} = \{(n_{i1}-1)(n_{i2}-1)\}^{-1/2}(\boldsymbol{X}_{i1} - \overline{\boldsymbol{X}}_{i1})^T(\boldsymbol{X}_{i2} - \overline{\boldsymbol{X}}_{i2}) \ \text{ for } i = 1, 2,$$

where $\overline{\boldsymbol{X}}_{il} = [\overline{\boldsymbol{x}}_{il}, ..., \overline{\boldsymbol{x}}_{il}]$ with $\overline{\boldsymbol{x}}_{i1} = \sum_{j=1}^{n_{i1}} \boldsymbol{x}_{ij}/n_{i1}$ and $\overline{\boldsymbol{x}}_{i2} = \sum_{j=n_{i1}+1}^{n_i} \boldsymbol{x}_{ij}/n_{i2}$. Note that $\mathrm{rank}(\boldsymbol{S}_{iD(1)}) \leq n_{i2}-1$. By using the cross-data-matrix (CDM) methodology by Yata and Aoshima (2010), we estimate $\lambda_{i(r)}$ by the r-th singular value, $\acute{\lambda}_{i(r)}$, of $\boldsymbol{S}_{iD(1)}$, where $\acute{\lambda}_{i(1)} \geq \cdots \geq \acute{\lambda}_{i(n_{i2}-1)} \geq 0$. Yata and Aoshima (2010, 2013) showed that $\acute{\lambda}_{i(r)}$ has several consistency properties for high-dimensional non-Gaussian data. Aoshima and Yata (2011) applied the CDM methodology to obtain an unbiased estimator of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ as $\mathrm{tr}(\boldsymbol{S}_{iD(1)}\boldsymbol{S}_{iD(1)}^T)$, $i = 1, 2$. Note that $E\{\mathrm{tr}(\boldsymbol{S}_{iD(1)}\boldsymbol{S}_{iD(1)}^T)\} = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$. Also, note that $\acute{\lambda}_{i(r)}^2$ is the r-th eigenvalue of $\boldsymbol{S}_{iD(1)}\boldsymbol{S}_{iD(1)}^T$. By using the CDM methodology, we consider an estimation of $\Psi_{i(r)}$ as $\widehat{\Psi}_{i(1)} = \mathrm{tr}(\boldsymbol{S}_{iD(1)}\boldsymbol{S}_{iD(1)}^T)$ and

$$\widehat{\Psi}_{i(r)} = \mathrm{tr}(\boldsymbol{S}_{iD(1)}\boldsymbol{S}_{iD(1)}^T) - \sum_{s=1}^{r-1} \acute{\lambda}_{i(s)}^2 \ \text{ for } r = 2, ..., n_{i2};\ i = 1, 2. \tag{22}$$

Note that $\widehat{\Psi}_{i(r)} \geq 0$ w.p.1 for $r = 1, ..., n_{i2}$, and $\hat{\eta}_{i(r)} \in (0, 1]$ for $\acute{\lambda}_{i(r)} > 0$. Then, Aoshima and Yata (2018) gave the following result.

Lemma 1 (Aoshima and Yata 2018) *Assume (A-i) and (M-i). Then, it holds that* $\widehat{\Psi}_{i(r)}/\Psi_{i(r)} = 1 + o_P(1)$ *as* $m \to \infty$ *for* $r = 1, ..., k_i + 1;\ i = 1, 2$.

From (S7.1) in Appendix C of Aoshima and Yata (2018), it holds that $\acute{\lambda}_{i(r)}/\lambda_{i(r)} = 1 + o_P(1)$ as $m \to \infty$ for $r = 1, ..., k_i;\ i = 1, 2$, under (A-i) and (M-i). From Lemma 1 we claim under (A-i) and (M-i) that as $m \to \infty$

$$\hat{\eta}_{i(r)} = \eta_{i(r)}\{1 + o_P(1)\} \ \text{ for } r = 1, ..., k_i;\ i = 1, 2. \tag{23}$$

Let $\hat{\tau}_{i(r)} = \widehat{\Psi}_{i(r+1)}/\widehat{\Psi}_{i(r)}$ $(= 1 - \acute{\lambda}_{i(r)}^2/\widehat{\Psi}_{i(r)})$ for all i, r. Note that $1 - \hat{\tau}_{i(1)} = \hat{\eta}_{i(1)}$ and $\hat{\tau}_{i(r)} \in [0, 1)$ for $\acute{\lambda}_{i(r)} > 0$. Then, Aoshima and Yata (2018) gave the following result.

Proposition 3 (Aoshima and Yata 2018) *Assume (A-i) and (M-i). It holds for* $i = 1, 2$, *that as* $m \to \infty$

$$P(\hat{\tau}_{i(r)} < 1 - c_r) \to 1 \ \textit{with some fixed constant } c_r \in (0, 1) \textit{ for } r = 1, ..., k_i;$$
$$\hat{\tau}_{i(k_i+1)} = 1 + o_P(1).$$

Table 2 Estimates of k_i by $\hat{k}_i$ for the six well-known microarray data sets

	(D-i)	(D-ii)	(D-iii)	(D-iv)	(D-v)	(D-vi)
$\hat{k}_1$	2	3	2	2	1	2
$\hat{k}_2$	4	2	2	2	2	3

From Proposition 3, one may choose k_i as the first integer r such that $1-\hat{\tau}_{i(r+1)}$ is sufficiently small. In addition, Aoshima and Yata (2018) gave the following result for $\hat{\tau}_{i(k_i+1)}$.

Proposition 4 (Aoshima and Yata 2018) *Assume (A-i) and (M-i). Assume also* $\lambda^2_{i(1)}/\Psi_{i(k_i+1)} = o(n_i)$ *and* $\lambda^2_{i(k_i+1)}/\Psi_{i(k_i+1)} = O(n_i^{-c})$ *as* $m \to \infty$ *with some fixed constant* $c > 1/2$ *for* $i = 1, 2$. *It holds for* $i = 1, 2$ *that as* $m \to \infty$

$$P\Big(\hat{\tau}_{i(k_i+1)} > \{1 + (k_i+1)\gamma(n_i)\}^{-1}\Big) \to 1,$$

where $\gamma(n_i)$ *is a function such that* $\gamma(n_i) \to 0$ *and* $n_i^{1/2}\gamma(n_i) \to \infty$ *as* $n_i \to \infty$.

From Propositions 3 and 4, if one can assume the conditions in Proposition 4, one may consider k_i as the first integer r $(= \hat{k}_{oi},$ say) such that

$$\hat{\tau}_{i(r+1)}\{1 + (r+1)\gamma(n_i)\} > 1 \quad (r \geq 0). \tag{24}$$

Then, it holds that $P(\hat{k}_{oi} = k_i) \to 1$ as $m \to \infty$. Note that $\widehat{\Psi}_{i(n_{i2})} = 0$ from the fact that $\mathrm{rank}(\boldsymbol{S}_{iD(1)}) \leq n_{i2}-1$. Thus one may choose k_i as $\hat{k}_i = \min\{\hat{k}_{oi}, n_{i2}-2\}$ in actual data analyses. Aoshima and Yata (2018) recommended to use $\gamma(n_i) = (n_i^{-1}\log n_i)^{1/2}$. Hence, in this paper, we use $\gamma(n_i) = (n_i^{-1}\log n_i)^{1/2}$ in (24). If $\hat{k}_i = 0$ (that is, (24) holds when $r = 0$) for some i, one may consider the classifier by (21) with $\boldsymbol{A}_i = \boldsymbol{I}_p$. In addition, if $\hat{k}_i = 0$ for $i = 1, 2$, we recommend to use DBDA (the classifier by (3)) because one may assume the NSSE model when $\hat{k}_i = 0$ for $i = 1, 2$. We summarized $\hat{k}_i$s in Table 2 for the six well-known microarray data sets, (D-i) to (D-vi).

5 Performances of the new classifier for the SSE model

In this section, we discuss the performance of T-DBDA in numerical simulations and actual data analyses.

5.1 Simulation

We compared the performance of T-DBDA with other classifiers in complex settings. In general, k_is are unknown in (21). Hence, we estimated k_i by $\hat{k}_i$, where $\hat{k}_i$ is given in Sect. 4.3. Hereafter, we describe the classification rule (21) with $\hat{k}_i$ instead of k_i as "T-DBDA(∗)". We set $\gamma(n_i) = (n_i^{-1}\log n_i)^{1/2}$ in (24). We set $p = 2^s$, $s = 6, ..., 11$, $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1, -1...., -1)^T$ whose last $2\lceil p^{3/5}/2\rceil$ elements are

not 0. The last $\lceil p^{3/5}/2 \rceil$ elements are -1 and the previous $\lceil p^{3/5}/2 \rceil$ elements are 1. Note that $\Delta = p^{3/5}\{1 + o(1)\}$ as $p \to \infty$.

First, we considered an intraclass correlation model given by

$$\boldsymbol{\Gamma}_t = (\boldsymbol{I}_t + \mathbf{1}_t \mathbf{1}_t^T)/2.$$

Note that $\lambda_{\max}(\boldsymbol{\Gamma}_t) = (t+1)/2$ and the other eigenvalues are $1/2$. Let $\boldsymbol{\Omega}_t(\rho) = \boldsymbol{B}(\rho^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B} = \mathrm{diag}[\{0.5 + 1/(t+1)\}^{1/2}, ..., \{0.5 + t/(t+1)\}^{1/2}]$. Also, note that $[\lambda_{\max}\{\boldsymbol{\Omega}_t(\rho)\}]^2/\mathrm{tr}[\{\boldsymbol{\Omega}_t(\rho)\}^2] = o(1)$ as $t \to \infty$ for $|\rho| < 1$. We set $n_1 = \lceil p^{1/2} \rceil$, $n_2 = 2n_1$ and

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Gamma}_{p_{i(1)}} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Gamma}_{p_{i(2)}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & c_i \boldsymbol{\Omega}_{p_{i(3)}}(\rho) \end{pmatrix}, \quad i = 1, 2, \tag{25}$$

where $\rho = 0.3$, $p = p_{i(1)} + p_{i(2)} + p_{i(3)}$ and $(c_1, c_2) = (1, 1.3)$. We considered the following settings:

(S-ii) We generated $\boldsymbol{x}_{ij}$, $j = 1, 2, ...$ $(i = 1, 2)$ independently from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We set $(p_{1(1)}, p_{1(2)}) = (\lceil p^{2/3} \rceil, \lceil p^{1/2} \rceil)$ and $(p_{1(1)}, p_{1(2)}) = (2\lceil p^{2/3} \rceil, 2\lceil p^{1/2} \rceil)$;

(S-iii) We generated $\boldsymbol{x}_{ij}$, $j = 1, 2, ...$ $(i = 1, 2)$ independently from $z_{ij(r)} = (y_{ij(r)} - 1)/2^{1/2}$ $(r = 1, ..., p)$ in which $y_{ij(r)}$s are i.i.d. as the chi-squared distribution with 1 degree of freedom. We set $(p_{1(1)}, p_{1(2)}) = (\lceil p/3 \rceil, \lceil p/9 \rceil)$ and $(p_{1(1)}, p_{1(2)}) = (2\lceil p/3 \rceil, 2\lceil p/9 \rceil)$.

For (S-ii) and (S-iii) we note that $\Delta_A = \Delta$ and $\lambda_{i(r)} = (p_{i(r)} + 1)/2$, $i, r = 1, 2$, for sufficiently large p, so that (M-i) with $k_1 = k_2 = 2$ is met. In particular, the SSSE model (given by (5)) holds for (S-iii). Also, we note that (A-i), (AY-i), (C-i) to (C-iii) and (C-vi) to (C-viii) are met both for (S-ii) and (S-iii), and (AY-ii) is met for (S-ii). However, (AY-ii) is not met for (S-iii).

Next, we considered a Gaussian mixture model whose probability density function is given by

$$f_i(\boldsymbol{y}) = \frac{1}{3}\sum_{l=1}^{3} g(\boldsymbol{y};\ \boldsymbol{\mu}_{il(y)}, \boldsymbol{\Sigma}_{i(y)}), \quad i = 1, 2, \tag{26}$$

where $g(\boldsymbol{y};\ \boldsymbol{\mu}_{il(y)}, \boldsymbol{\Sigma}_{i(y)})$ is the probability density function of $N_p(\boldsymbol{\mu}_{il(y)}, \boldsymbol{\Sigma}_{i(y)})$. We set $\boldsymbol{\Sigma}_{1(y)} = \boldsymbol{\Omega}_p(0.3)$ and $\boldsymbol{\Sigma}_{2(y)} = \boldsymbol{\Omega}_p(0.5)$. Let $q_{1(1)} = \lceil p^{2/3} \rceil$, $q_{2(1)} = 2\lceil p^{2/3} \rceil$, $q_{1(2)} = 2\lceil p^{1/2} \rceil$ and $q_{2(2)} = \lceil p^{1/2} \rceil$. We set $\boldsymbol{\mu}_{i1(y)} = (3^{1/2}, ..., 3^{1/2}, 0, ..., 0)^T$ whose first $q_{i(1)}$ elements are $3^{1/2}$, $\boldsymbol{\mu}_{i2(y)} = (0, ..., 0, 3^{1/2}, ..., 3^{1/2}, 0, ..., 0)^T$ whose $(q_{i(1)} + 1)$-th to $(q_{i(1)} + q_{i(2)})$-th elements are $3^{1/2}$ and $\boldsymbol{\mu}_{i3(y)} = \mathbf{0}$. We generated $\boldsymbol{y}_{ij}$, $j = 1, 2, ...$ $(i = 1, 2)$ independently from (26). Note that $E(\boldsymbol{y}_{ij}) = \sum_{l=1}^{3} \boldsymbol{\mu}_{il(y)}/3$ for $i = 1, 2$. We set $\boldsymbol{x}_{ij} = \boldsymbol{y}_{ij} - \sum_{l=1}^{3} \boldsymbol{\mu}_{il(y)}/3 + \boldsymbol{\mu}_i$ for all i, j. Note that $\boldsymbol{\Sigma}_i = \mathrm{Var}(\boldsymbol{y}_{ij})$ for $i = 1, 2$, where

$$\mathrm{Var}(\boldsymbol{y}_{ij}) = \frac{1}{9}\sum_{l<l'}^{3}(\boldsymbol{\mu}_{il(y)} - \boldsymbol{\mu}_{il'(y)})(\boldsymbol{\mu}_{il(y)} - \boldsymbol{\mu}_{il'(y)})^T + \boldsymbol{\Sigma}_{i(y)}.$$

We note that $\lambda_{i(1)} = (2/3)q_{i(1)}\{1+o(1)\}$ and $\lambda_{i(2)} = (1/2)q_{i(2)}\{1+o(1)\}$ as $p \to \infty$ for $i = 1, 2$, so that (M-i) with $k_1 = k_2 = 2$ is met. See Corollary 2 in Yata and Aoshima (2015) for the details of the eigenvalues. Also, note that $\Delta_A = \Delta$ for sufficiently large p and (A-i) is not met. We considered the following settings:

(S-iv) $n_1 = \lceil p^{2/5} \rceil$ and $n_2 = 2n_1$;
(S-v) $n_1 = \lceil p^{3/5} \rceil$ and $n_2 = 2n_1$.

We note that (AY-i), (AY-ii), (C-i) to (C-iii) and (C-vi) to (C-viii) are met both for (S-iv) and (S-v).

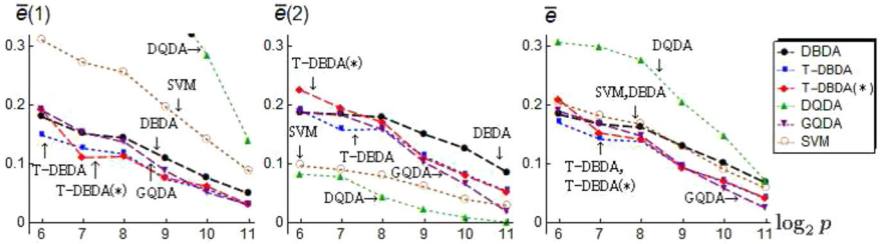
We considered DBDA (the classifier (3)), T-DBDA (the classifier (21)) and T-DBDA(∗) (the classifier (21) with $\hat{k}_i$ instead of k_i). We also considered the following three classifiers: Diagonal quadratic discriminant analysis (DQDA) given by Dudoit et al. (2002), Geometrical quadratic discriminant analysis (GQDA) given by Aoshima and Yata (2011, 2014), and Support vector machine (SVM). The rule of GQDA is given by (6) in Aoshima and Yata (2014). SVM is the hard-margin linear rule. Similar to Fig. 3, we calculated the error rates, $\overline{e}(1)$ and $\overline{e}(2)$, by 2000 replications. Also, we calculated the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. Their standard deviations are less than 0.011. In Fig. 4, we plotted the results for (S-ii) to (S-v).

We observed that GQDA gives a better performance than DBDA, DQDA and SVM for (S-ii). This is probably because $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$. DQDA performs better than DBDA, GQDA and SVM for (S-v). This is probably because n_is are relatively large and the diagonal elements of the two covariance matrices are not common. See Sections 2 to 4 in Aoshima and Yata (2015b) for the details of DQDA and GQDA. For SVM, $\overline{e}(1)$ and $\overline{e}(2)$ were unbalanced. The main reason must be due to a bias term in SVM. See Section 2 in Nakayama et al. (2017) for the details. On the other hand, DBDA gave a moderate performance for (S-iii). This is probably because DBDA is quite robust for non-Gaussian HDLSS data. See Aoshima and Yata (2014) for the details. On the whole, T-DBDA and T-DBDA(∗) gave adequate performances. In particular, T-DBDA(∗) (or T-DBDA) gave a much better performance than the other classifiers both for (S-iii), in which (5) holds, and (S-iv), in which n_is are relatively small. This is probably due to the sufficient conditions of the consistency properties. See Sect. 3.3 for the details. The performances of T-DBDA and T-DBDA(∗) became quite similar to each other in almost all the cases. Hence, we recommend to use "the classifier (21) with $\hat{k}_i$ instead of k_i" when the SSE condition (4) or the SSSE condition (5) holds.

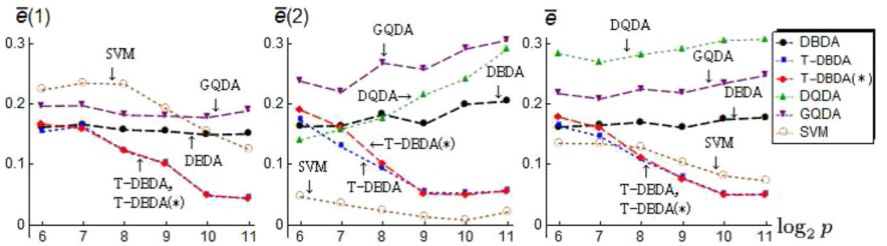
5.2 Example

In this section, we check the performance of T-DBDA(∗) by using the six well-known microarray data sets in Table 1.

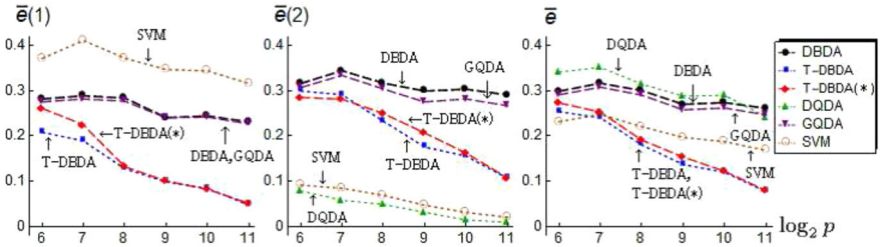
First, we used (D-v): myeloma data ($p = 12625$). We defined $n_1 = 36$ samples from π_1 and $n_2 = 136$ (the first 136) samples from π_2 as the training data, and the last (the 137-th) sample of π_2 as the test data. We centered each sample by $\boldsymbol{x}_{ij} -$



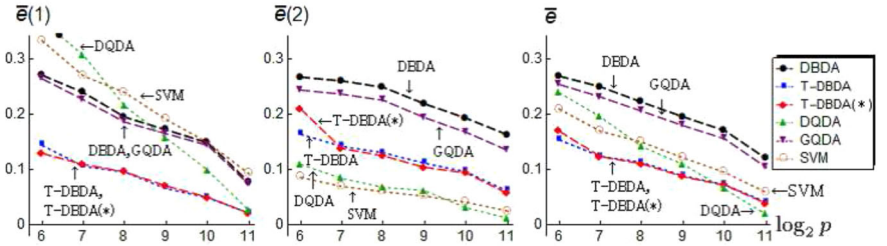
(S-ii): $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $(\lambda_{1(1)}, \lambda_{1(2)}) \approx (p^{2/3}/2, p^{1/2}/2)$ and $(\lambda_{2(1)}, \lambda_{2(2)}) \approx (p^{2/3}, p^{1/2})$.



(S-iii): $z_{ij(r)} = (y_{ij(r)} - 1)/2^{1/2}$ $(r = 1, ..., p)$ in which $y_{ij(r)}$s are i.i.d. as the chi-squared distribution with 1 degree of freedom, $(\lambda_{1(1)}, \lambda_{1(2)}) \approx (p/6, p/18)$ and $(\lambda_{2(1)}, \lambda_{2(2)}) \approx (p/3, p/9)$.



(S-iv): The mixture model given by (26) and $(n_1, n_2) = (\lceil p^{2/5} \rceil, 2\lceil p^{2/5} \rceil)$.



(S-v): The mixture model given by (26) and $(n_1, n_2) = (\lceil p^{3/5} \rceil, 2\lceil p^{3/5} \rceil)$.

Fig. 4 The left panel displays $\bar{e}(1)$, the middle panel displays $\bar{e}(2)$ and the right panel displays $\bar{e}$. The error rates of the classifiers, DBDA, T-DBDA, T-DBDA(∗), DQDA, GQDA, SVM. In the left panels, $\bar{e}(1)$s for DQDA are not described because the error rates were too high

$(\sum_{i'=1}^{2}\sum_{j'=1}^{n_{i'}} \boldsymbol{x}_{i'j'})/(n_1+n_2)$ for all i, j, and $\boldsymbol{x}_0 - (\sum_{i'=1}^{2}\sum_{j'=1}^{n_{i'}} \boldsymbol{x}_{i'j'})/(n_1+n_2)$, so that $\sum_{i=1}^{2}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} = \boldsymbol{0}$. We set $\gamma(n_i) = (n_i^{-1}\log n_i)^{1/2}$ in (24). Let $\tilde{\tau}_{i(r)} = \hat{\tau}_{i(r)}\{1 + r\gamma(n_i)\}$ for all i, r. We calculated that $(\tilde{\tau}_{1(1)}, \tilde{\tau}_{1(2)}) = (0.943, 1.046)$ and $(\tilde{\tau}_{2(1)}, \tilde{\tau}_{2(2)}, \tilde{\tau}_{2(3)}) = (0.878, 0.986, 1.168)$, so that $\hat{k}_1 = 1$ and $\hat{k}_2 = 2$. Thus, we chose $k_1 = 1$ and $k_2 = 2$. We calculated that $\widetilde{W}_A(\boldsymbol{x}_0) = 305.439$, so that we classified $\boldsymbol{x}_0$ into π_2 (the true class).

Similarly, we checked the accuracy of T-DBDA(∗) by the leave-one-out cross-validation (LOOCV) for (D-i) to (D-vi). Also, we checked the accuracy of the classifiers, DBDA, DQDA, GQDA, SVM, by the LOOCV for (D-i) to (D-vi). In addition, we checked the accuracy of the well-known classifiers, Diagonal linear discriminant analysis (DLDA) given by Dudoit et al. (2002) and distance weighted discrimination (DWD) given by Marron et al. (2007). For DWD, we calculated the normal vector by the SOCP solver in Marron et al. (2007) and set the intercept term as 0 since we used the mean-centered data.

We summarized misclassification rates, $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$, in Table 3.

We observed that T-DBDA(∗) gives adequate performances. In particular, the new classifier gave a much better performance than the other classifiers (except SVM) for (D-iv). This is probably because (D-iv) is close to the SSSE asymptotic domain (5). See Table 1 or Fig. 1. The other classifiers were probably affected by the strongly spiked eigenvalues directly. On the other hand, the new classifier is not directly affected by such eigenvalues. See Theorems 3 and 5 for the details. This is the reason why the new classifier gave a good performance for (D-iv). On the other hand, (D-i) is close to the SSSE asymptotic domain (5). However, the several classifiers gave adequate performances for (D-i). This is probably because n_is are relativity large compared to p.

6 Proofs

6.1 Proof of Theorem 3

We note that for $i, l = 1, 2;\ i' \neq i$

$$\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{l,A}) = \{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}\boldsymbol{\Sigma}_{l,A}) + 2\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}\boldsymbol{\Sigma}_{l,A}\boldsymbol{A}_{i'}) + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{A}_{i'}\boldsymbol{\Sigma}_{l,A}\boldsymbol{A}_{i'})\}/4. \quad (27)$$

From the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{A}_{i'}\boldsymbol{\Sigma}_{i,A}\boldsymbol{A}_{i'}) = \mathrm{tr}(\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{A}_{i'}\boldsymbol{\Sigma}_{i,A}\boldsymbol{A}_{i'}\boldsymbol{\Sigma}_i^{1/2}) \geq 0$ $(i' \neq i)$, under (C-ii), it holds that $\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2)/(n_i\Delta_A^2) \to 0$ as $m \to \infty$ for $i = 1, 2$. Thus we claim that $\delta_{oi,A}^2/\Delta_A^2 = o(1)$ for $i = 1, 2$, under (C-ii). Note that for $i = 1, 2$,

$$\begin{aligned}
\boldsymbol{\mu}_i^T\boldsymbol{A}_{1,2}\boldsymbol{\Sigma}_{l,A}\boldsymbol{A}_{1,2}\boldsymbol{\mu}_i/n_l &\leq \boldsymbol{\mu}_i^T\boldsymbol{A}_{1,2}^2\boldsymbol{\mu}_i\lambda_{\max}(\boldsymbol{\Sigma}_{l,A})/n_l \\
&= (\boldsymbol{\mu}_i^T\boldsymbol{A}_{1,2}^2\boldsymbol{\mu}_i/n_l^{1/2})(\lambda_{l(k_l+1)}/n_l^{1/2}),\ \ l = 1, 2;\ \text{ and} \\
|\boldsymbol{\mu}_A^T\boldsymbol{\Sigma}_{i',A}\boldsymbol{A}_{1,2}\boldsymbol{\mu}_i| &\leq \{(\boldsymbol{\mu}_A^T\boldsymbol{\Sigma}_{i',A}\boldsymbol{\mu}_A)(\boldsymbol{\mu}_i^T\boldsymbol{A}_{1,2}\boldsymbol{\Sigma}_{i',A}\boldsymbol{A}_{1,2}\boldsymbol{\mu}_i)\}^{1/2},\ \ i' \neq i. \quad (28)
\end{aligned}$$

Table 3 Error rates of the classifiers by the LOOCV for samples from (D-i) to (D-vi)

Classifier	T-DBDA(∗)	DBDA	DLDA	DQDA	GQDA	SVM	DWD
Error rates							
π_1: 104 samples and π_2: 113 samples in (D-i)							
$\bar{e}_1$	0.0	0.183	0.163	0.0	0.0	0.0	0.0
$\bar{e}_2$	0.009	0.009	0.009	0.018	0.044	0.0	0.009
$\bar{e}$	0.004	0.096	0.086	0.009	0.022	0.0	0.004
π_1: 40 samples and π_2: 22 samples in (D-ii)							
$\bar{e}_1$	0.15	0.15	0.15	0.15	0.15	0.15	0.15
$\bar{e}_2$	0.136	0.136	0.136	0.182	0.136	0.227	0.091
$\bar{e}$	0.143	0.143	0.143	0.166	0.143	0.189	0.12
π_1 : 111 samples and π_2 : 57 samples in (D-iii)							
$\bar{e}_1$	0.198	0.243	0.162	0.216	0.198	0.135	0.243
$\bar{e}_2$	0.281	0.316	0.368	0.456	0.404	0.439	0.246
$\bar{e}$	0.239	0.28	0.265	0.336	0.301	0.287	0.244
π_1 : 58 samples and π_2 : 19 samples in (D-iv)							
$\bar{e}_1$	0.034	0.172	0.19	0.155	0.172	0.017	0.224
$\bar{e}_2$	0.0	0.158	0.211	0.421	0.158	0.0	0.0
$\bar{e}$	0.017	0.165	0.2	0.288	0.165	0.009	0.112
π_1 : 36 samples and π_2 : 137 samples in (D-v)							
$\bar{e}_1$	0.25	0.278	0.528	0.639	0.278	0.75	0.222
$\bar{e}_2$	0.197	0.292	0.219	0.109	0.299	0.058	0.365
$\bar{e}$	0.224	0.285	0.373	0.374	0.289	0.404	0.294
π_1 : 84 samples and π_2 : 44 samples in (D-vi)							
$\bar{e}_1$	0.143	0.107	0.06	0.083	0.143	0.06	0.107
$\bar{e}_2$	0.182	0.25	0.318	0.227	0.227	0.25	0.205
$\bar{e}$	0.162	0.179	0.189	0.155	0.185	0.155	0.156

Thus by noting that $\lambda_{l(k_l+1)} = o\{\mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)^{1/2}\}$ under (M-i) and $\delta_{oi,A}^2/\Delta_A^2 = o(1)$ under (C-ii), we claim that $\delta_{i,A}^2/\Delta_A^2 = o(1)$ for $i = 1, 2$, under (M-i), (C-i) to (C-iii). From (12) and Chebyshev's inequality, we can conclude the results of Theorem 3. □

6.2 Proof of Corollary 1

By noting that $\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{l,A}) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}^2)\mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)\}^{1/2}$ for $i, l = 1, 2$, when $\boldsymbol{A}_1 = \boldsymbol{A}_2$, the result is obtained straightforwardly from Theorem 3. □

6.3 Proof of Theorem 4

We first consider the case when $\boldsymbol{x}_0 \in \pi_1$. Let $\omega_{i,A} = \{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{i,A})/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{i',A})/n_{i'}\}^{1/2}$ for $i = 1, 2;\ i' \neq i$. Then, from (27), under (C-iv), we have

that

$$\delta_{o1,A} = \omega_{1,A}\{1 + o(1)\} \tag{29}$$

and $\sum_{l=1}^{2} \mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)/n_l = O(\delta_{o1,A}^2)$ as $m \to \infty$. From (27), we note that $\lambda_{l(k_l+1)}/n_l^{1/2} = o[\{\mathrm{tr}(\boldsymbol{\Sigma}_{l,A}^2)/n_l\}^{1/2}] = o(\delta_{o1,A})$ for $l = 1, 2$, under (M-i) and (C-iv). Thus from (28) it holds that for $i = 1, 2$,

$$\delta_{1,A} = \delta_{o1,A}\{1 + o(1)\} \tag{30}$$

under (M-i), (C-iv) and (C-v). By combining (29) and (30), under (M-i), (C-iv) and (C-v), we have that $\delta_{1,A} = \omega_{1,A}\{1 + o(1)\}$ and

$$W_A(\boldsymbol{x}_0) + \frac{\Delta_A}{2} = (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{A}_*\{(\overline{\boldsymbol{x}}_{2,A} - \boldsymbol{\mu}_{2,A}) - (\overline{\boldsymbol{x}}_{1,A} - \boldsymbol{\mu}_{1,A})\} + o_P(\omega_{1,A}). \tag{31}$$

Let us write that

$$\begin{aligned} v_j &= -(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{A}_*(\boldsymbol{x}_{1j,A} - \boldsymbol{\mu}_{1,A})/(n_1\omega_{1,A}), \quad j = 1, ..., n_1; \\ v_{n_1+j} &= (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{A}_*(\boldsymbol{x}_{2j,A} - \boldsymbol{\mu}_{2,A})/(n_2\omega_{1,A}), \quad j = 1, ..., n_2. \end{aligned}$$

Note that $\sum_{j=1}^{n_1+n_2} E(v_j^2) = 1$ and $\sum_{j=1}^{n_1+n_2} v_j = (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{A}_*\{(\overline{\boldsymbol{x}}_{2,A} - \boldsymbol{\mu}_{2,A}) - (\overline{\boldsymbol{x}}_{1,A} - \boldsymbol{\mu}_{1,A})\}/\omega_{1,A}$. Then, it holds that $E(v_j | v_{j-1}, ..., v_1) = 0$ for $j = 2, ..., n_1 + n_2$. We consider applying the martingale central limit theorem given by McLeish (1974). In a way similar to the equations (23) and (24) in Aoshima and Yata (2014), we can evaluate that under (A-i)

$$(n_{l_j}\omega_1)^4 E(v_j^4) = O[\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A})^2 + \mathrm{tr}\{(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A})^2\}] \text{ and} \tag{32}$$

$$\begin{aligned} &(n_{l_j}n_{l_{j'}})^2\omega_1^4 E(v_j^2 v_{j'}^2) \\ &\quad = \mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A})\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_{j'},A}) + O\{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A}\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_{j'},A})\} \\ &\qquad + O[\{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A}\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A})\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_{j'},A}\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_{j'},A})\}^{1/2}] \end{aligned} \tag{33}$$

for $j \neq j'$, where $l_j = 1$ for $j \in [1, ..., n_1]$ and $l_j = 2$ for $j \in [n_1 + 1, ..., n_1 + n_2]$. For any $\tau > 0$ we note that $\sum_{j=1}^{n_1+n_2} E\{v_j^2 I(v_j^2 \geq \tau)\} \leq \sum_{j=1}^{n_1+n_2} E(v_j^4)/\tau$ from Chebyshev's inequality and Schwarz's inequality, where $I(\cdot)$ is the indicator function. Also, note that $\mathrm{tr}\{(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l,A})^2\} \leq \mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l,A})^2$ for $l = 1, 2$. Then, from (32), under (A-i), it holds that for Lindeberg's condition

$$\sum_{j=1}^{n_1+n_2} E\{v_j^2 I(v_j^2 \geq \tau)\} = O\left[\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{1,A})^2/n_1^3 + \mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{2,A})^2/n_2^3}{\omega_{1,A}^4}\right] = o(1)$$

for any $\tau > 0$. Note that for $l, l' = 1, 2$,

$$\begin{aligned}\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l,A}\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l',A}) &= \mathrm{tr}\{(\boldsymbol{\Sigma}_{1,A*}^{1/2}\boldsymbol{\Sigma}_{l,A}\boldsymbol{\Sigma}_{1,A*}^{1/2})(\boldsymbol{\Sigma}_{1,A*}^{1/2}\boldsymbol{\Sigma}_{l',A}\boldsymbol{\Sigma}_{1,A*}^{1/2})\} \\ &\leq \lambda_{\max}(\boldsymbol{\Sigma}_{1,A*}^{1/2}\boldsymbol{\Sigma}_{l,A}\boldsymbol{\Sigma}_{1,A*}^{1/2})\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}^{1/2}\boldsymbol{\Sigma}_{l',A}\boldsymbol{\Sigma}_{1,A*}^{1/2}) \\ &= o\{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l,A})\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l',A})\}\end{aligned}$$

under (C-v), so that $(n_{l_j}n_{l_{j'}})^2\omega_1^4 E(v_j^2v_{j'}^2) = \mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_j,A})\mathrm{tr}(\boldsymbol{\Sigma}_{1,A*}\boldsymbol{\Sigma}_{l_{j'},A})\{1+o(1)\}$ for $j \neq j'$. Hence, by using Chebyshev's inequality, from (32) and (33), under (A-i) and (C-v), it holds that for any $\tau > 0$

$$P\left(\Big|\sum_{j=1}^{n_1+n_2} v_j^2 - 1\Big| \geq \tau\right) \leq \frac{E\Big[\sum_{j,j'=1}^{n_1+n_2}\{v_j^2 - E(v_j^2)\}\{v_{j'}^2 - E(v_{j'}^2)\}\Big]}{\tau^2} = o(1),$$

so that $\sum_{j=1}^{n_1+n_2} v_j^2 = 1 + o_P(1)$. Hence, by using the martingale central limit theorem, we obtain that $\sum_{j=1}^{n_1+n_2} v_j \Rightarrow N(0, 1)$ under (A-i) and (C-v). Thus from (31) we conclude the result when $\boldsymbol{x}_0 \in \pi_1$. When $\boldsymbol{x}_0 \in \pi_2$, we can conclude the result similarly. The proof is completed. □

6.4 Proof of Corollary 2

When $\boldsymbol{A}_1 = \boldsymbol{A}_2$, we note that $\lambda_{\max}(\boldsymbol{\Sigma}_{i,A*}^{1/2}\boldsymbol{\Sigma}_{l,A}\boldsymbol{\Sigma}_{i,A*}^{1/2}) \leq \lambda_{i(k_i+1)}\lambda_{l(k_l+1)}$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{i,A*}\boldsymbol{\Sigma}_{l,A}) = \mathrm{tr}(\boldsymbol{\Sigma}_{i,A}\boldsymbol{\Sigma}_{l,A})$ for $i, l = 1, 2$. On the other hand, when $\boldsymbol{A}_1 = \boldsymbol{A}_2$, it holds that $\boldsymbol{\mu}_A^T\boldsymbol{\Sigma}_{i',A}\boldsymbol{\mu}_A/(n_{i'}\delta_{oi,A}^2) = o(1)$ as $m \to \infty$ for $i = 1, 2$; $i' \neq i$, under $\boldsymbol{\mu}_A^T\boldsymbol{\Sigma}_{i',A}\boldsymbol{\mu}_A/(\delta_{oi',A}^2) = o(1)$ as $m \to \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{1,A}^2)/\mathrm{tr}(\boldsymbol{\Sigma}_{2,A}^2) \in (0, \infty)$ as $p \to \infty$. Hence, from Theorem 4 we can conclude the results. □

6.5 Proof of Proposition 2

We assume (A-i) and (M-i). Let $\boldsymbol{u}_{i(r)} = (z_{i1(r)}, ..., z_{in_i(r)})/(n_i - 1)^{1/2}$ and $\dot{\boldsymbol{u}}_{i(r)} = \|\boldsymbol{u}_{i(r)}\|^{-1}\boldsymbol{u}_{i(r)}$ for all i, j. Then, from (S6.1) to (S6.3) and (S6.5) in Appendix B of Aoshima and Yata (2018), we can claim that as $m \to \infty$ for $i = 1, 2$,

$$\begin{aligned}&\tilde{\lambda}_{i(r)}/\lambda_{i(r)} = ||\boldsymbol{u}_{i(r)}||^2 + O_P(n_i^{-1}) = 1 + O_P(n_i^{-1/2}) \\ &\quad \text{and } \hat{\boldsymbol{u}}_{i(r)}^T\dot{\boldsymbol{u}}_{i(r)} = 1 + O_P(n_i^{-1}) \text{ for } r = 1, ..., k_i; \qquad (34) \\ &\hat{\boldsymbol{u}}_{i(s)}^T\boldsymbol{u}_{i(r)} = O_P(n_i^{-1/2}\lambda_{i(s)}/\lambda_{i(r)}) \\ &\quad \text{and } \hat{\boldsymbol{u}}_{i(r)}^T\boldsymbol{u}_{i(s)} = O_P(n_i^{-1/2}) \text{ for } r < s \leq k_i. \qquad (35)\end{aligned}$$

From (34) there exists a unit random vector $\boldsymbol{\zeta}_{i(r)}$ such that $\dot{\boldsymbol{u}}_{i(r)}^T\boldsymbol{\zeta}_{i(r)} = 0$ and

$$\hat{\boldsymbol{u}}_{i(r)} = \{1 + O_P(n_i^{-1})\}\dot{\boldsymbol{u}}_{i(r)} + \boldsymbol{\zeta}_{i(r)} \times O_P(n_i^{-1/2}) \qquad (36)$$

for $r = 1, ..., k_i;\ i = 1, 2$. We note that $\mathbf{1}_n^T \hat{\boldsymbol{u}}_{i(r)} = 0$ and $\boldsymbol{P}_{n_i} \hat{\boldsymbol{u}}_{i(r)} = \hat{\boldsymbol{u}}_{i(r)}$ when $\hat{\lambda}_{i(r)} > 0$ since $\mathbf{1}_{n_i}^T \boldsymbol{S}_{iD} \mathbf{1}_{n_i} = 0$. Also, when $\hat{\lambda}_{i(r)} > 0$, note that

$$\tilde{\boldsymbol{h}}_{i(r)} = \frac{(\boldsymbol{X}_i - \boldsymbol{\mu}_i \mathbf{1}_{n_i}^T)\boldsymbol{P}_{n_i} \hat{\boldsymbol{u}}_{i(r)}}{\{(n_i - 1)\tilde{\lambda}_{i(r)}\}^{1/2}} = \frac{\sum_{s=1}^{p} \lambda_{i(s)}^{1/2} \boldsymbol{h}_{i(s)} \boldsymbol{u}_{i(s)}^T \hat{\boldsymbol{u}}_{i(r)}}{\tilde{\lambda}_{i(r)}^{1/2}},$$

so that $\boldsymbol{x}_0^T \tilde{\boldsymbol{h}}_{i(r)} = \sum_{s=1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}^T \hat{\boldsymbol{u}}_{i(r)} / \tilde{\lambda}_{i(r)}^{1/2}$. Here, we claim that when $\boldsymbol{x}_0 \in \pi_l,\ l = 1, 2$,

$$E\left\{\left(\frac{\sum_{s=k_i+1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}^T \boldsymbol{u}_{i(r)}}{\lambda_{i(r)}^{1/2}}\right)^2\right\} = O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_l \boldsymbol{\Sigma}_{i,A}) + \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_l}{n_i \lambda_{i(r)}}\right\};$$

$$E\left\{\left\|\frac{\sum_{s=k_i+1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}}{\lambda_{i(r)}^{1/2}}\right\|^2\right\} = O\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_l \boldsymbol{\Sigma}_{i,A}) + \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_l}{\lambda_{i(r)}}\right\}$$

for $r = 1, ..., k_i;\ i = 1, 2$. Then, from (34) and (36), it holds that when $\boldsymbol{x}_0 \in \pi_l,\ l = 1, 2$,

$$\frac{\sum_{s=k_i+1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}^T \hat{\boldsymbol{u}}_{i(r)}}{\tilde{\lambda}_{i(r)}^{1/2}} = O_P\left\{\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_l \boldsymbol{\Sigma}_{i,A}) + \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_l}{n_i \lambda_{i(r)}}\right)^{1/2}\right\} \tag{37}$$

for $r = 1, ..., k_i;\ i = 1, 2$, from the fact that $\sum_{s=k_i+1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}^T \boldsymbol{\zeta}_{i(r)} / \lambda_{i(r)}^{1/2} \leq \|\lambda_{i(r)}^{-1/2} \sum_{s=k_i+1}^{p} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}\| \cdot \|\boldsymbol{\zeta}_{i(r)}\|$ and Markov's inequality. Note that $E(x_{0,i(s)}^2) = \boldsymbol{h}_{i(s)}^T (\boldsymbol{\Sigma}_l + \boldsymbol{\mu}_l \boldsymbol{\mu}_l^T) \boldsymbol{h}_{i(s)}$ when $\boldsymbol{x}_0 \in \pi_l$ $(l = 1, 2)$ for all i, s, so that $x_{0,i(s)} = O_P[\{\boldsymbol{h}_{i(s)}^T (\boldsymbol{\Sigma}_l + \boldsymbol{\mu}_l \boldsymbol{\mu}_l^T) \boldsymbol{h}_{i(s)}\}^{1/2}]$. Then, from (34) and (35), we have that when $\boldsymbol{x}_0 \in \pi_l,\ l = 1, 2$,

$$\begin{aligned} &\frac{\sum_{s=1}^{k_i} \lambda_{i(s)}^{1/2} x_{0,i(s)} \boldsymbol{u}_{i(s)}^T \hat{\boldsymbol{u}}_{i(r)}}{\tilde{\lambda}_{i(r)}^{1/2}} \\ &\quad = x_{0,i(r)} + O_P\left\{\left(\sum_{s=1}^{k_i} \frac{\lambda_{i(s)} \boldsymbol{h}_{i(s)}^T (\boldsymbol{\Sigma}_l + \boldsymbol{\mu}_l \boldsymbol{\mu}_l^T) \boldsymbol{h}_{i(s)}}{n_i \max\{\lambda_{i(s)}^2 / \lambda_{i(r)}, \lambda_{i(r)}\}}\right)^{1/2}\right\} \end{aligned} \tag{38}$$

for $r = 1, \ldots, k_i;\ i = 1, 2$. By combining (37) and (38), we can conclude the second result of Proposition 2. For the first result, from Proposition 1 and the second result, it concludes the result. □

6.6 Proofs of Theorems 5 and 6

Assume (A-i) and (M-i). We first consider the proof of Theorem 5. Let $\psi_{i(r)} = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/(n_i^2\lambda_{i(r)}) + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/(n_i\lambda_{i(r)})$ for $r = 1, ..., k_i;\ i = 1, 2$. Then, from Lemma B.1 and (S6.27) in Appendix B of Aoshima and Yata (2018), we claim that as $m \to \infty$

$$\overline{\tilde{x}}_{i(r)} = \bar{x}_{i(r)} + O_P(\psi_{i(r)}^{1/2}) \ \text{ and } \ \bar{x}_{i(r)} = \mu_{i(r)} + O_P\left\{(\lambda_{i(r)}/n_i)^{1/2}\right\} \tag{39}$$

for $r = 1, ..., k_i;\ i = 1, 2$. Note that under (C-vii)

$$\psi_{i(r)} = O\left(\frac{\lambda_{i(1)}^2 + n_i\boldsymbol{\mu}_{i,A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i,A}}{n_i^2\lambda_{i(r)}}\right) \ \text{ for } r = 1, ..., k_i;\ i = 1, 2. \tag{40}$$

Note that $\mathrm{tr}(\boldsymbol{\Sigma}_{i,A}\boldsymbol{\Sigma}_{i'}) = \mathrm{tr}(\boldsymbol{\Sigma}_{1,A}\boldsymbol{\Sigma}_{2,A}) + O(\lambda_{i(k_i)}\lambda_{i'(1)}) = O(\lambda_{i(k_i)}\lambda_{i'(1)})$ and $\boldsymbol{\mu}_{i'}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i'} = O(\boldsymbol{\mu}_{i',A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i',A} + \sum_{s=1}^{k_{i'}}\lambda_{i(k_i)}\mu_{i'(s)}^2)$ for $i = 1, 2;\ i' \neq i$ from the facts that $\mathrm{tr}(\boldsymbol{\Sigma}_{1,A}\boldsymbol{\Sigma}_{2,A}) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A}^2)\mathrm{tr}(\boldsymbol{\Sigma}_{2,A}^2)\}^{1/2} = O(\lambda_{1(k_1)}\lambda_{2(k_2)})$ and $\boldsymbol{\mu}_{i',A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{h}_{i'(s)}\mu_{i'(s)} = O(\boldsymbol{\mu}_{i',A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i',A} + \lambda_{i(k_i)}\mu_{i'(s)}^2)$ for $s = 1, ..., k_{i'}$. From (37) and (38) we have that when $\boldsymbol{x}_0 \in \pi_l,\ l = 1, 2$,

$$\tilde{x}_{0,i(r)} = x_{0,i(r)} + O_P\left\{\left(\frac{\lambda_{i(r)}^2 + \boldsymbol{\mu}_{l,A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{l,A}}{n_i\lambda_{i(r)}}\right)^{1/2}\right\} + O_P\{(\lambda_{l(1)}/n_i)^{1/2}\}$$
$$\text{and } \ x_{0,i(r)} = O_P(\lambda_{i(r)}^{1/2}) \ \text{ for } r = 1, ..., k_i;\ i = 1, 2 \tag{41}$$

under (C-vi) and (C-vii). Then, from (39) to (41), under (C-vi) to (C-viii), we have that when $\boldsymbol{x}_0 \in \pi_l,\ l = 1, 2$,

$$\begin{aligned}\tilde{x}_{0,i(r)}\overline{\tilde{x}}_{i(r)} - x_{0,i(r)}\bar{x}_{i(r)} &= (\tilde{x}_{0,i(r)} - x_{0,i(r)})\overline{\tilde{x}}_{i(r)} + x_{0,i(r)}(\overline{\tilde{x}}_{i(r)} - \bar{x}_{i(r)}) \\ &= o_P(\Delta_A) \ \text{ for } r = 1, ..., k_i;\ i = 1, 2.\end{aligned} \tag{42}$$

On the other hand, from (S6.29) in Appendix B of Aoshima and Yata (2018) we claim that for $r = 1, ..., k_1$ and $s = 1, ..., k_2$

$$\begin{aligned}&\tilde{\boldsymbol{h}}_{1(r)}^T\tilde{\boldsymbol{h}}_{2(s)} = \boldsymbol{h}_{1(r)}^T\boldsymbol{h}_{2(s)} + O_P(n_{\min}^{-1/2}),\ \ \tilde{\boldsymbol{h}}_{1(r)}^T(\tilde{\boldsymbol{h}}_{2(s)} - \boldsymbol{h}_{2(s)}) = O_P(n_2^{-1/2}), \\ &\tilde{\boldsymbol{h}}_{2(s)}^T(\tilde{\boldsymbol{h}}_{1(r)} - \boldsymbol{h}_{1(r)}) = O_P(n_1^{-1/2}) \\ &\text{and } \ (\tilde{\boldsymbol{h}}_{1(r)} - \boldsymbol{h}_{1(r)})^T(\tilde{\boldsymbol{h}}_{2(s)} - \boldsymbol{h}_{2(s)}) = O_P\{(n_1n_2)^{-1/2}\}.\end{aligned} \tag{43}$$

Note that $\bar{x}_{i(r)}\boldsymbol{h}_{i(r)} - \overline{\tilde{x}}_{i(r)}\tilde{\boldsymbol{h}}_{i(r)} = \bar{x}_{i(r)}(\boldsymbol{h}_{i(r)} - \tilde{\boldsymbol{h}}_{i(r)}) - (\overline{\tilde{x}}_{i(r)} - \bar{x}_{i(r)})\tilde{\boldsymbol{h}}_{i(r)}$ for all i, r. Then, from (39) and (43), we have that for $r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i$,

$$\tilde{\boldsymbol{h}}_{i(r)}^T \sum_{s=1}^{k_{i'}} (\bar{x}_{i'(s)} \boldsymbol{h}_{i'(s)} - \bar{\tilde{x}}_{i'(s)} \tilde{\boldsymbol{h}}_{i'(s)})$$
$$= O_P \left(\sum_{s=1}^{k_{i'}} (\psi_{i'(s)}^{1/2} (\boldsymbol{h}_{i(r)}^T \boldsymbol{h}_{i'(s)} + n_{\min}^{-1/2}) + \lambda_{i'(s)}^{1/2} / n_{i'} + \mu_{i'(s)} / n_{i'}^{1/2}) \right). \quad (44)$$

Similar to the proof of Proposition 2 and (41), under (C-vi) and (C-vii), we can claim that for $r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i$,

$$\tilde{\boldsymbol{h}}_{i(r)}^T \bar{\boldsymbol{x}}_{i',A} = \boldsymbol{h}_{i(r)}^T \bar{\boldsymbol{x}}_{i',A} + O_P \left\{ \left(\frac{\lambda_{i(r)}^2 / n_{\min} + \boldsymbol{\mu}_{i',A}^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_{i',A}}{n_i \lambda_{i(r)}} \right)^{1/2} \right\}$$
$$+ O_p[\{\lambda_{i'(1)} / (n_1 n_2)\}^{1/2}]. \quad (45)$$

Note that $\sum_{s=1}^{k_{i'}} (\boldsymbol{h}_{i(r)}^T \boldsymbol{h}_{i'(s)})^2 / \lambda_{i'(s)} = O(1/\lambda_{i(r)})$ under (C-vi) for $r = 1, ..., k_i; i = 1, 2;\ i' \neq i$. From (40), (44) and (45) we have that for $r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i$,

$$\tilde{\boldsymbol{h}}_{i(r)}^T \left(\bar{\boldsymbol{x}}_{i'} - \sum_{s=1}^{k_{i'}} \bar{\tilde{x}}_{i'(s)} \tilde{\boldsymbol{h}}_{i'(s)} \right) - \boldsymbol{h}_{i(r)}^T \left(\bar{\boldsymbol{x}}_{i'} - \sum_{s=1}^{k_{i'}} \bar{x}_{i'(s)} \boldsymbol{h}_{i'(s)} \right)$$
$$= O_P \left\{ \left(\frac{\boldsymbol{\mu}_{i',A}^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_{i',A}}{n_i \lambda_{i(r)}} + \frac{\boldsymbol{\mu}_{i',A}^T \boldsymbol{\Sigma}_{i',A} \boldsymbol{\mu}_{i',A}}{\min\{\lambda_{i(r)}, n_{\min} \lambda_{i'(k_{i'})}\} n_{i'}} + \frac{\lambda_{i(1)} + \lambda_{i'(1)}}{n_{\min}^2} + \frac{\lambda_{i'(1)}^2}{n_{\min}^2 \lambda_{i(r)}} \right)^{1/2} \right\} \quad (46)$$

under (C-vi) and (C-vii). Note that $\boldsymbol{h}_{i(r)}^T (\bar{\boldsymbol{x}}_{i'} - \sum_{s=1}^{k_{i'}} \bar{x}_{i'(s)} \boldsymbol{h}_{i'(s)}) = O_P\{(\lambda_{i(r)} / n_{i'})^{1/2}\}$ under (C-vi) and (C-vii) for $r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i$. Then, similar to (42), from (41) and (46), we have that

$$\tilde{x}_{0,i(r)} \tilde{\boldsymbol{h}}_{i(r)}^T \left(\bar{\boldsymbol{x}}_{i'} - \sum_{s=1}^{k_{i'}} \bar{\tilde{x}}_{i'(s)} \tilde{\boldsymbol{h}}_{i'(s)} \right) - x_{0,i(r)} \boldsymbol{h}_{i(r)}^T \left(\bar{\boldsymbol{x}}_{i'} - \sum_{s=1}^{k_{i'}} \bar{x}_{i'(s)} \boldsymbol{h}_{i'(s)} \right)$$
$$= o_P(\Delta_A) \quad \text{for } r = 1, ..., k_i;\ i = 1, 2;\ i' \neq i \quad (47)$$

under (C-vi) to (C-viii). Also, from (S6.28) in Appendix B of Aoshima and Yata (2018), we claim that for $r = 1, ..., k_i;\ i = 1, 2$,

$$\sum_{j<j'}^{n_i} \frac{\tilde{x}_{ij(r)} \tilde{x}_{ij'(r)} - x_{ij(r)} x_{ij'(r)}}{n_i (n_i - 1)} = O_P \left\{ \psi_{i(r)}^{1/2} (\psi_{i(r)}^{1/2} + \lambda_{i(r)}^{1/2} / n_i^{1/2} + \mu_{i(r)}) \right\}.$$

Note that under (C-vii) and (C-viii)

$$\sum_{r=1}^{k_i} \psi_{i(r)}^{1/2}(\psi_{i(r)}^{1/2} + \lambda_{i(r)}^{1/2}/n_i^{1/2} + \mu_{i(r)}) = O\left(\frac{\lambda_{i(1)}\lambda_{i(k_i)} + \boldsymbol{\mu}_{i,A}^T \boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i,A}}{n_i\lambda_{i(k_i)}} + \frac{(\lambda_{i(1)}^2 + n_i\boldsymbol{\mu}_{i,A}^T\boldsymbol{\Sigma}_{i,A}\boldsymbol{\mu}_{i,A})^{1/2}}{n_i^{3/2}}\right) = o_P(\Delta_A) \tag{48}$$

for $i = 1, 2$. By combining (42), (47) and (48), it holds that $\widetilde{W}_A(\boldsymbol{x}_0) = W_A(\boldsymbol{x}_0) + o_P(\Delta_A)$ when $\boldsymbol{x}_0 \in \pi_i,\ i = 1, 2$ under (C-vi) to (C-viii). It concludes the results of Theorem 5.

Similar to the proof of Theorem 5, it holds that $\widetilde{W}_A(\boldsymbol{x}_0) = W_A(\boldsymbol{x}_0) + o_P(\delta_{o\min,A})$ when $\boldsymbol{x}_0 \in \pi_i,\ i = 1, 2$ under (C-vi), (C-vii) and (C-ix). It concludes the results of Theorem 6. □

Acknowledgements We would like to thank two anonymous referees for their constructive comments.

References

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

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 6745–6750.

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

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

Aoshima, M., Yata, K. (2015a). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis*, *34*, 279–294.

Aoshima, M., Yata, K. (2015b). High-dimensional quadratic classifiers in non-sparse settings. arXiv preprint. arXiv:1503.04549.

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

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

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.

Cai, T. T., Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, *106*, 1566–1577.

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. *The Annals of Statistics*, *38*, 808–835.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., et al. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics*, *5*, e1000602.

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.

Fan, J., Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, *36*, 2605–2637.
Glaab, E., Bacardit, J., Garibaldi, J. M., Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS ONE*, *7*, e39932.
Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., et al. (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*, *49*, 1125–1134.
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.
Jeffery, I. B., Higgins, D. G., Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, *7*, 359.
Li, Q., Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, *25*, 457–473.
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. *The Annals of Probability*, *2*, 620–628.
Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., et al. (2007). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, *26*, 1507–1516.
Nakayama, Y., Yata, K., Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *Journal of Statistical Planning and Inference*, *191*, 88–100.
Ramey J. A. (2016). Datamicroarray: collection of data sets for classification. https://github.com/ramhiser/datamicroarray.
Shao, J., Wang, Y., Deng, X., Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, *39*, 1241–1265.
Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*, 68–74.
Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., et al. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *The New England Journal of Medicine*, *349*, 2483–2494.
Watanabe, H., Hyodo, M., Seo, T., Pavlenko, T. (2015). Asymptotic properties of the misclassification rates for Euclidean distance discriminant rule in high-dimensional data. *Journal of Multivariate Analysis*, *140*, 234–244.
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. (2012). 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. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis*, *122*, 334–354.
Yata, K., Aoshima, M. (2015). Principal component analysis based clustering for high-dimension, low-sample-size data. arXiv preprint. arXiv:1503.04525.