HIGH-DIMENSIONAL QUADRATIC CLASSIFIERS IN NON-SPARSE SETTINGS

BY MAKOTO AOSHIMA* AND KAZUYOSHI YATA†

University of Tsukuba

We consider high-dimensional quadratic classifiers in non-sparse settings. The target of classification rules is not Bayes error rates in the context. The classifier based on the Mahalanobis distance does not always give a preferable performance even when the sample sizes grow to infinity and the population distributions are assumed Gaussian, having known covariance matrices. The quadratic classifiers proposed in this paper draw information effectively about heteroscedasticity through the difference of parameters related to the expanding covariance matrices. We show that the quadratic classifiers hold consistency properties in which misclassification rates tend to zero as the dimension goes to infinity under non-sparse conditions. We verify that the quadratic classifiers are asymptotically distributed as a normal distribution when the dimension goes to infinity, also under certain conditions. We discuss feature selection and sparse inverse covariance matrix estimation for further evaluation of misclassification rates to give guidelines for the choice of the classifiers.

1. Introduction. Globally, there is an ever increasing need for fast, accurate and cost effective analysis of high-dimensional data in many fields, including academia, medicine and business. However, existing classifiers for high-dimensional data are often complex, time consuming and have high computational cost. In this paper we hope to provide better options. A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. This is the so-called “HDLSS” or “large p, small n” data situation 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, \( \pi_i, i = 1, 2 \), having an unknown mean vector \( \mu_i = (\mu_{i1}, ..., \mu_{ip})^T \) and unknown covariance matrix \( \Sigma_i (> O) \) for each \( i \). Let \( \mu_{12} = \mu_1 - \mu_2 = (\mu_{121}, ..., \mu_{12p})^T \) and \( \Sigma_{12} = \Sigma_1 - \Sigma_2 \). We assume that \( \limsup_{p \to \infty} |\mu_{12j}| < \infty \)

* Supported by Grants-in-Aid for Scientific Research (B) and Challenging Exploratory Research, JSPS, under Contract Numbers 22300094 and 26540010.
†Supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078.

AMS 2000 subject classifications: Primary 62H30; secondary 62H10

Keywords and phrases: Bayes error rate, Discriminant analysis, Feature selection, Heteroscedasticity, Large p small n
for all $j$. Note that $\limsup_{p \to \infty} \|\mu_2\|^2/p < \infty$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\sigma_{i(j)}$ be the $j$-th diagonal element of $\Sigma_i$ for $j = 1, \ldots, p$ ($i = 1, 2$). We assume that $\sigma_{i(j)} \in (0, \infty)$ as $p \to \infty$ for all $i, j$. Here, for a function, $f(\cdot)$, “$f(p) \in (0, \infty)$ as $p \to \infty$” implies that $\liminf_{p \to \infty} f(p) > 0$ and $\limsup_{p \to \infty} f(p) < \infty$. Then, it holds that $\text{tr}(\Sigma_i)/p \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$. We do not assume $\Sigma_1 = \Sigma_2$. The eigen-decomposition of $\Sigma_i$ is given by $\Sigma_i = H_i \Lambda_i H_i^T$, where $\Lambda_i = \text{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$, and $H_i = [h_{i1}, \ldots, h_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. We have independent and identically distributed (i.i.d.) observations, $x_{i1}, \ldots, x_{im}$, from each $\pi_i$, where $x_{ij} = (x_{i1j}, \ldots, x_{ipj})^T$, $j = 1, \ldots, n_i$. We assume $n_i \geq 2$, $i = 1, 2$. Let $n_{\min} = \min\{n_1, n_2\}$. We estimate $\mu_i$ and $\Sigma_i$ by $\overline{\mu}_i = (\overline{x}_{1m}, \ldots, \overline{x}_{pm})^T = \sum_{j=1}^{n_i} x_{ij}/n_i$ and $\overline{\Sigma}_i = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{im})(x_{ij} - \overline{x}_{im})^T/n_i - 1$. Let $\sigma_{in}(j)$ be the $j$-th diagonal element of $\overline{\Sigma}_{in}$ for $j = 1, \ldots, p$ ($i = 1, 2$).

In this paper, we consider high-dimensional quadratic classifiers in non-sparse settings. Let $x_0 = (x_{01}, \ldots, x_{0p})^T$ be an observation vector of an individual belonging to one of the two populations. Let $|M|$ be the determinant of a square matrix $M$. When $\pi_i$s are Gaussian, a Bayes optimal rule is given as follows: One classifies the individual into $\pi_1$ if

\begin{equation}
(x_0 - \mu_1)^T \Sigma_1^{-1} (x_0 - \mu_1) - \log |\Sigma_2 \Sigma_1^{-1}| < (x_0 - \mu_2)^T \Sigma_2^{-1} (x_0 - \mu_2)
\end{equation}

and into $\pi_2$ otherwise. Since $\mu_i$s and $\Sigma_i$s are unknown, one usually consider the following typical classifier: $(x_0 - \overline{x}_{1m})^T S_{1m}^{-1} (x_0 - \overline{x}_{1m}) - \log |S_{2m} S_{1m}^{-1}| < (x_0 - \overline{x}_{2m})^T S_{2m}^{-1} (x_0 - \overline{x}_{2m})$. The classifier usually converges to the Bayes optimal classifier when $n_{\min} \to \infty$ while $p$ is fixed or $n_{\min}/p \to \infty$. However, the inverse matrix of $S_{in}$ does not exist in the HDLSS context. When $\Sigma_1 = \Sigma_2$, Bickel and Levina [5] considered an inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. Fan and Fan [12] considered a classification after feature selection. Fan et al. [13] proposed a regularized optimal affine discriminant (ROAD). When $\Sigma_1 \neq \Sigma_2$, Dudoit et al. [11] considered an inverse matrix defined by only diagonal elements of $S_{in}$. Aoshima and Yata [1] considered substituting $\{\text{tr}(S_{in})/p\} I_p$ for $S_{in}$ by using a geometric representation of HDLSS data from each $\pi_i$ and gave a quadratic classifier whose misclassification rates are no more than specified thresholds. Hall et al. [15] and Marron et al. [18] considered distance weighted classifiers, and Aoshima and Yata [2] considered distance-based classifiers for multiclass, high-dimensional data.

Recently, Cai and Liu [9], Li and Shao [17] and Shao et al. [20] gave sparse linear or quadratic classification rules for high-dimensional data. They showed that their classification rules have Bayes error rates when $\pi_i$s are
Gaussian. They assumed that λ_{ij}s are bounded under some sparsity conditions such as \( \mu_{12}, \Sigma_{1}s \) and \( \Sigma_{12} \) (or \( \Sigma_{i}^{-1}s \) and \( \Sigma_{2}^{-1} - \Sigma_{1}^{-1} \)) are sparse. For example, when \( \Sigma_{1} = \Sigma_{2} (= \Sigma, \text{ say}) \), the error rate of their classification rules is given by \( \Phi(-\Delta_{MD}/2) + o(1) \) as \( p \to \infty \), where \( \Delta_{MD} = \mu_{12}^T \Sigma^{-1} \mu_{12} \) and \( \Phi(.) \) denotes the cumulative distribution function of the standard normal distribution. Here, \( \Phi(-\Delta_{MD}/2) \) is the Bayes error rate.

In this paper, we investigate quadratic classifiers from a different point of view. We do not assume that \( \mu_{12}, \Sigma_{1}s \) and \( \Sigma_{12} \) are sparse. In the context, the target of classification rules is not Bayes error rates as in \( \Phi(-\Delta_{MD}/2) + o(1) \) as \( p \to \infty \). We consider a consistency property such as misclassification rates tend to 0 as \( p \) increases, i.e.,

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

where \( e(i) \) denotes the error rate of misclassifying an individual from \( \pi_i \) into the other class. For example, when \( \pi_i \)'s are Gaussian having \( \Sigma_{1} = \Sigma_{2} \), the Bayes rule by (1.1) has such a consistency property under “\( \Delta_{MD} \to \infty \) as \( p \to \infty \)”. It is likely that “\( \Delta_{MD} \to \infty \) as \( p \to \infty \)” holds when \( \mu_{12} \) is non-sparse in the sense that \( ||\mu_{12}|| \to \infty \) as \( p \to \infty \). We emphasize that such non-sparse situations often occur in high-dimensional settings. See Hall et al. [15] or Section 6 for example. We will show that quadratic classifiers hold the consistency property when \( \mu_{12} \) or \( \Sigma_{12} \) is non-sparse such as \( ||\mu_{12}|| \to \infty \) or \( ||\Sigma_{12}||_{F} \to \infty \) as \( p \to \infty \), where \( || \cdot \||_{F} \) is the Frobenius norm.

In this paper, we consider the following function of \( A_{j} \) to discriminate \( \pi_j, \ j = 1, 2; \)

\[(1.2) \quad W_j(A_j) = (x_0 - \bar{x}_{jn})^T A_j^{-1} (x_0 - \bar{x}_{jn}) - \text{tr}(S_{jn} A_j^{-1})/n_j + \log |A_j|
\]

for \( A_{j} (> O) \) such that \( \text{tr}\{\Sigma_i(A_j^{-1} - A_i^{-1})\} = \text{tr}(A_iA_j^{-1}) - p \ (i \neq j) \). Here, \( \text{tr}(S_{jn} A_j^{-1})/n_j \) is a bias correction term. Then, we consider a quadratic classification rule in which one classifies the individual into \( \pi_1 \) if

\[(1.3) \quad W_1(A_1) - W_2(A_2) < 0
\]

and into \( \pi_2 \) otherwise. Note that (1.3) becomes a linear classifier when \( A_1 = A_2 \). Let

\[(1.4) \quad \Delta_i = \mu_{12}^T A_j^{-1} \mu_{12} + \text{tr}\{\Sigma_i(A_j^{-1} - A_i^{-1})\} + \log |A_jA_i^{-1}|
\]

for \( i = 1, 2 \) (\( j \neq i \)). Then, \( E\{W_j(A_j)\} - E\{W_i(A_i)\} = \Delta_i \) when \( x_0 \in \pi_i \).

**Proposition 1.1.** (i) \( \Delta_i \geq 0 \). (ii) \( \Delta_i > 0 \) when \( \mu_1 \neq \mu_2 \) or \( A_1 \neq A_2 \).
Remark 1. As for $k (\geq 3)$-class classification, one may consider a classification rule such as one classifies the individual into $\pi_i$ if
\[
\arg\min_{j=1, \ldots, k} W_j(A_j) = i.
\]

In this paper, we consider the following four typical $A_j$s:

(I) $A_j = I_p$, (II) $A_j = \frac{\text{tr}(\Sigma_j)}{p} I_p$, (III) $A_j = \Sigma_{j(d)}$, and (IV) $A_j = \Sigma_j$,

where $\Sigma_{j(d)} = \text{diag}(\sigma_{j(1)}, \ldots, \sigma_{j(p)})$. These four $A_j$s are specifically selected because they provide historical background of discriminant analysis and satisfy the condition $\text{tr}\{\Sigma_i(A_j^{-1} - A_i^{-1})\} = \text{tr}(A_iA_j^{-1}) - p$. Note that $||\Sigma_{12}||_F \geq ||A_1 - A_2||_F$ for these four $A_j$s. Also, under (I) to (IV), we note that $\Delta_i \rightarrow \infty$ as $p \rightarrow \infty$ when $\mu_{12}$ or $\Sigma_{12}$ is non-sparse. Practically, $A_j$s should be estimated except for (I). We will consider quadratic classifiers given by estimating $A_j$s in Section 4. Let us see an easy example to check the performance of (I) to (IV) in (1.3). We set $p = 2^s$, $s = 3, \ldots, 12$. Independent pseudo random observations were generated from $\pi_i : N_p(\mu_i, \Sigma_i)$, $i = 1, 2$. We set $\mu_1 = 0$ and $\Sigma_1 = B_1(0.3^{|i-j|^{1/3}})B_1$, where
\[
B_1 = \text{diag}[[0.5 + 1/(p + 1)]^{1/2}, \ldots, (0.5 + p/(p + 1))]^{1/2}.
\]

Note that $\text{tr}(\Sigma_1) = p$ and $\Sigma_{1(d)} = B_1^2$. When $\Sigma_1 = \Sigma_2$ and $(n_1, n_2) = (\log_2 p, 2 \log_2 p)$ that is $n_{\text{min}} \rightarrow \infty$ as $p \rightarrow \infty$, we considered two cases:

(a) $\mu_2 = (1, \ldots, 1, 0, \ldots, 0)^T$ whose first $[p^{2/3}]$ elements are 1, and
(b) $\mu_2 = (0, \ldots, 0, 1, \ldots, 1)^T$ whose last $[p^{2/3}]$ elements are 1.

Here, $[x]$ denotes the smallest integer $\geq x$. Next, when $\mu_2 = 0$ (i.e., $\mu_{12} = 0$) and $(n_1, n_2) = (5, 10)$ that is $n_i$s are fixed, we considered two cases:

(c) $\Sigma_2 = 1.5 \Sigma_1$ and (d) $\Sigma_2 = 1.2 I_p$.

Note that $\mu_{12}$ or $\Sigma_{12}$ is non-sparse for (a) to (d) because $||\mu_{12}|| \rightarrow \infty$ or $||\Sigma_{12}||_F \rightarrow \infty$ as $p \rightarrow \infty$. For $x_0 \in \pi_i$ ($i = 1, 2$) we repeated 2000 times to confirm if the classification rule by (1.3) with either of (I) to (IV) does (or does not) classify $x_0$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each $\pi_i$. We calculated the error rates, $\tau(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Also, we calculated the average error rate, $\overline{\tau} = \{\tau(1) + \tau(2)\}/2$. Their standard deviations are less than 0.011. In Fig. 1, we plotted $\tau$ for (a) and (b). Note that (I) is equivalent to (II) for (a) and (b). In Fig. 2, we plotted $\tau$ for (c) and (d). We observed that (IV) gives the worst performance in Fig. 1 contrary to expectations. In general, one would think that the classifier based on the
Mahalanobis distance such as (1.2) with (IV) is the best when $\pi_i$s are Gaussian and $n_{\text{min}} \to \infty$. We emphasize that it is not true for high-dimensional data. We will explain its theoretical reason in Section 3.2. We observed that (I) (or (II)) gives a better performance compared to (III) for (b) in Fig. 1. We will discuss the reasons in Section 3.4. In Fig. 2, the error rates of (I) are close to 0.5 because of $\mu_{12} = 0$. On the other hand, (II), (III) and (IV) gave good performances as $p$ increases by drawing information on heteroscedasticity in the classifiers. We will give their theoretical backgrounds in Sections 2.2 and 3.4.

In this paper, we pay special attention to the difference of covariance matrices in classification for high-dimensional data. We first consider the quadratic classifiers by (1.2) having either of (I) to (IV). In Section 2, we show that the classification rule by (1.2) holds the consistency property under non-sparse conditions. In Section 3, we verify that the quadratic classifiers are asymptotically distributed as a normal distribution under certain conditions when $p$ goes to infinity. In Section 4, we consider the estimation of $A_{ij}$s and give asymptotic properties of the estimated classifiers. In Section 5, we discuss quadratic classifiers by feature selection or by sparse inverse covariance matrices. In Section 6, we give examples by using leukemia data. Finally, in Section 7, we give concluding remarks of our study.
2. Consistency of the quadratic classifiers. In this section, we give
sufficient conditions for the quadratic classifiers given by (1.2) to hold the
consistency property in misclassification rates.

2.1. Preliminary. Similar to [2] and [4], we assume the following asump-
tion about population distributions as necessary:

(A-i) Let \( y_{ij}, j = 1, \ldots, n_i \), be i.i.d. random \( q_i \)-vectors having \( E(y_{ij}) = 0 \) and \( \text{Var}(y_{ij}) = I_{q_i} \) for each \( i = 1, 2 \), where \( q_i \geq p \). Let \( y_{ij} = (y_{i1j}, \ldots, y_{iq_ij})^T \) whose components satisfy that \( \lim \sup_{p \to \infty} E(y_{ij}^4) < \infty \) for all \( r \) and

\[
E(y_{irj}^2) = E(y_{irj}^2)(E(y_{irj}^2) = 1 \quad \text{and} \quad E(y_{irj}y_{isj}y_{itj}y_{iju}) = 0
\]

for all \( r \neq s, t, u \). Then, the observations, \( x_{ij} \), from each \( \pi_i \) \((i = 1, 2)\) are given by

\[
x_{ij} = \Gamma_i y_{ij} + \mu_i, \quad j = 1, \ldots, n_i,
\]

where \( \Gamma_i = [\gamma_{i1}, \ldots, \gamma_{iq_i}] \) is a \( p \times q_i \) matrix such that \( \Gamma_i \Gamma_i^T = \Sigma_i \).

Note that \( \Gamma_i \) includes the case that \( \Gamma_i = H_i A_i^{1/2} = [\lambda_i^{1/2} h_{i1}, \ldots, \lambda_i^{1/2} h_{ip}] \).

We assume the following assumption instead of (A-i) as necessary:

(A-ii) (A-i) by replacing (2.1) with the assumption that \( y_{isj}, s = 1, \ldots, q_i \), are independent for each \( i, j \) \((i = 1, 2; j = 1, \ldots, n_i)\).

Note that (A-ii) is a special case of (A-i). When \( \pi_i \) has \( N_p(\mu_i, \Sigma_i) \), (A-ii) naturally holds.

Now, we consider the following divergence condition for \( p \) and \( n_i \) s:

\[
(*) \quad p \to \infty \quad \text{either when} \quad n_i \text{ is fixed or} \quad n_i \to \infty \quad \text{for} \quad i = 1, 2.
\]

Let \( \Delta_iA = \mu_{i2} A_i^{-1} \Sigma_i A_i^{-1} \mu_{i2} \) for \( i = 1, 2 \) \((j \neq i)\). We consider the following conditions under \((*)\) for \( i = 1, 2 \) \((j \neq i)\):

(C-i) \( \frac{\text{tr}((\Sigma_i A_i^{-1})^2)}{n_i \Delta_i^2} = o(1) \) and \( \frac{\text{tr}(\Sigma_i A_i^{-1} \Sigma_j A_j^{-1}) + \text{tr}((\Sigma_j A_j^{-1})^2)/n_j}{n_i \Delta_i^2} = o(1) \)

(C-ii) \( \frac{\Delta_iA}{\Delta_i^2} = o(1) \), \( \text{and} \) (C-iii) \( \frac{\text{tr}((\Sigma_i (A_i^{-1} - A_j^{-1}))^2)}{\Delta_i^2} = o(1) \).

Then, we claim the consistency property of (1.2) in (1.3) as follows:

THEOREM 2.1. Assume (A-i). Assume also (C-i) to (C-iii). Then, we have that

\[
\frac{W_j(A_j)}{\Delta_i} - W_i(A_i) = 1 + o_P(1) \quad \text{under} \quad (*) \quad \text{when} \quad x_0 \in \pi_i \quad \text{for} \quad i = 1, 2 \quad (j \neq i).
\]
Furthermore, for the classification rule by (1.3) with (1.2), we have that
\[(2.2) \quad e(i) \to 0, \ i = 1, 2, \ under \ (*)\.

**Remark 2.** When \(A_1 = A_2\), we can claim Theorem 2.1 without (A-i) and (C-iii).

Let \(\lambda_{\min}(M)\) and \(\lambda_{\max}(M)\) be the smallest and the largest eigenvalues of any positive definite matrix, \(M\). We use the phrase “\(\lambda(M) \in (0, \infty)\) as \(p \to \infty\)” in the sense that \(\liminf_{p \to \infty} \lambda_{\min}(M) > 0\) and \(\limsup_{p \to \infty} \lambda_{\max}(M) < \infty\). We note that \(A_i\)s in (I) to (III) satisfy the condition “\(\lambda(A_i) \in (0, \infty)\) as \(p \to \infty\)”. Let \(\Delta_{\min} = \min\{\Delta_1, \Delta_2\}\), \(\lambda_{\max} = \max\{\lambda_{11}, \lambda_{21}\}\) and \(\text{tr}(\Sigma_{\max}^2) = \max\{\text{tr}(\Sigma_1^2), \text{tr}(\Sigma_2^2)\}\). Now, instead of (C-i) and (C-ii), we consider the following simpler conditions under (\(\star\)):

\[(C-i') \quad \frac{\text{tr}(\Sigma_{\max}^2)}{n_{\min} \Delta_{\min}^2} = o(1) \text{ and } (C-ii') \quad \frac{\lambda_{\max}}{\lambda_{\min}} = o(1)\.

**Proposition 2.1.** Assume that \(\liminf_{p \to \infty} \lambda_{\min}(A_i) > 0\) for \(i = 1, 2\). Then, (C-i') and (C-ii') imply (C-i) and (C-ii), respectively. Furthermore, if \(\lambda(A_i) \in (0, \infty)\) as \(p \to \infty\) for \(i = 1, 2\), and \(A_i, i = 1, 2\), are diagonal matrices such as in (I) to (III) in Section 1, (C-ii') implies (C-iii).

From the fact that \(\lambda_{i1} \leq \text{tr}(\Sigma_i^2)^{1/2}\) for \(i = 1, 2\), we note that (C-i') and (C-ii') hold even when \(n_{\min}\) is fixed under
\[(2.3) \quad \frac{\text{tr}(\Sigma_{\max}^2)}{\Delta_{\min}^2} \to 0 \text{ as } p \to \infty\.

2.2. Consistency for (I) to (IV). As mentioned in Section 1, four typical \(A_j\)s were specifically selected. Now, we consider (1.2) and (1.4). In (I), when \(A_j = I_p, j = 1, 2\), they are given by
\[(2.4) \quad W_j(I_p) = ||x_0 - \bar{x}_{j_n}||^2 - \frac{\text{tr}(S_{jn})}{n_j}
\text{and } \Delta_1 = \Delta_2 = ||\mu_12||^2 \text{ (hereafter called } \Delta_{(I)}\).

In (II), when \(A_j = \{\text{tr}(\Sigma_j)/p\}I_p, j = 1, 2\), they are given by
\[(2.5) \quad W_j(\{\text{tr}(\Sigma_j)/p\}I_p) = p||x_0 - \bar{x}_{j_n}||^2 - \frac{\text{tr}(S_{jn})}{n_j \text{tr}(\Sigma_j)} + p \log\{\text{tr}(\Sigma_j)/p\}
\text{and } \Delta_i = \frac{p\Delta_{(I)}}{\text{tr}(\Sigma_j)} + \frac{p \text{tr}(\Sigma_j)}{\text{tr}(\Sigma_j)} - p + p \log\{\frac{\text{tr}(\Sigma_j)}{\text{tr}(\Sigma_i)}\} \text{ (hereafter called } \Delta_{(II)}\).
In (III), when $A_{j} = \Sigma_{j(d)}$, $j = 1, 2$, they are given by
\begin{equation}
W_{j}(\Sigma_{j(d)}) = \sum_{r=1}^{p} \left( \frac{(x_{0r} - \overline{x}_{jrn})}{\sigma_{j(r)}} - \frac{s_{jrn}(r)}{n_{j} \sigma_{j(r)}} + \log \sigma_{j(r)} \right) \quad \text{(2.6)}
\end{equation}
and
\begin{equation}
\Delta_{i} = \sum_{s=1}^{p} \left\{ \frac{\mu_{12s}^2}{\sigma_{j(s)}} + \frac{\sigma_{i(s)}}{\sigma_{j(s)}} - 1 + \log \frac{\sigma_{j(s)}}{\sigma_{i(s)}} \right\} \quad \text{(hereafter called $\Delta_{i(III)}$)}
\end{equation}

In (IV), when $A_{j} = \Sigma_{j}$, $j = 1, 2$, they are given by
\begin{equation}
W_{j}(\Sigma_{j}) = (x_{0} - \overline{x}_{j})^{T} \Sigma_{j}^{-1} (x_{0} - \overline{x}_{j}) - \frac{\text{tr}(S_{jrn} \Sigma_{j}^{-1})}{n_{j}} + \sum_{s=1}^{p} \log \lambda_{js} \quad \text{(2.7)}
\end{equation}
and
\begin{equation}
\Delta_{i} = \mu_{12}^{T} \Sigma_{j}^{-1} \mu_{12} + \text{tr}(\Sigma_{i} \Sigma_{j}^{-1}) - p + \sum_{s=1}^{p} \log \frac{\lambda_{js}}{\lambda_{is}} \quad \text{(hereafter called $\Delta_{i(IV)}$)}
\end{equation}

We first consider the classifiers by (2.4) to (2.6). From Theorem 2.1 and Proposition 2.1, we have the following result.

**Corollary 2.1.** Assume (C-i') and (C-ii'). Then, for the classification rule by (1.3) with (2.4), we have (2.2). Furthermore, for the classification rule by (1.3) with (2.5) or (2.6), we have (2.2) under (A-i).

We note that the classifier by (2.4) is equivalent to the distance-based classifier by Aoshima and Yata [2]. They gave a partial result of Corollary 2.1 under different conditions. Hereafter, we call the classifier by (2.4) the “distance-based discriminant analysis (DBDA)”. From Corollary 2.1, under (2.3), the classification rule by (1.3) with (2.4), (2.5) or (2.6) has (2.2) even when $n_{i}$s are fixed. Note that DBDA has the consistency property without (A-i), so that DBDA is quite robust for non-Gaussian cases. See Aoshima and Yata [2] for details. When $\mu_{1} = \mu_{2}$, DBDA does not satisfy (C-i') and (C-ii'), however, the classifier by (2.5) or (2.6) still satisfies them.

Now, we consider the following condition for $\Sigma_{i}$, $i = 1, 2$:
\begin{equation}
\text{tr}(\Sigma_{i}^{2})/\text{tr}(\Sigma_{i})^{2} \to 0 \quad \text{as} \quad p \to \infty. \quad \text{(2.8)}
\end{equation}

We note that tr$(\Sigma_{i}^{2})/\text{tr}(\Sigma_{i})^{2}$ is a measure of sphericity. Also, note that (2.8) is equivalent to the condition that “$\lambda_{i1}/\text{tr}(\Sigma_{i}) \to 0 \quad \text{as} \quad p \to \infty$”. Under (A-i) and (2.8), from the fact that $\text{Var}(|x_{0} - \mu_{i}|^{2}) = O(\text{tr}(\Sigma_{i}^{2}))$ when $x_{0} \in \pi_{i}$, we have that as $p \to \infty$
\begin{equation}
||x_{0} - \mu_{i}|| = \text{tr}(\Sigma_{i})^{1/2} \{ 1 + o_{p}(1) \} \quad \text{when} \quad x_{0} \in \pi_{i}.
\end{equation}
Thus the centroid data lies near the surface of an expanding sphere. See Hall et al. [15] for details of the geometric representation. We emphasize that the classifier by (2.5) draws information about heteroscedasticity thorough the geometric representation having different radii, tr(Σ_i)^{1/2}s, of expanding two spheres. Note that tr(Σ_i^2) = o(p^2) under (2.8). Hence, for the classifier by (2.5), (2.3) holds under (2.8) and \( \liminf_{p \to \infty} \Delta_{\min(I)} / p > 0 \), where \( \Delta_{\min(I)} = \min\{\Delta_{1(I)}, \Delta_{2(I)}\} \). Note that \( \Delta_{\min(I)} > 0 \) when tr(Σ_i) \( \neq \) tr(Σ_2) in view of Proposition 1.1. If one can assume that \( \liminf_{p \to \infty} |tr(\Sigma_1) / tr(\Sigma_2) - 1| > 0 \), it follows \( \liminf_{p \to \infty} \Delta_{\min(I)} / p > 0 \), so that (2.3) holds under (2.8). Hence, for the classification rule by (1.3) with (2.5), we have (2.2) even when \( \mu_1 = \mu_2 \) and \( n_i \)s are fixed. See (II) in Fig. 2. The accuracy becomes higher as the difference between tr(Σ_j)s grows.

Similarly, for the classifier by (2.6), it follows that (2.3) holds under (2.8) and \( \liminf_{p \to \infty} \Delta_{\min(I)} / p > 0 \), where \( \Delta_{\min(I)} = \min\{\Delta_{1(I)}, \Delta_{2(I)}\} \). If one can assume that \( \liminf_{p \to \infty} \sum_{s=1}^{p} |\sigma_{1(s)} / \sigma_{2(s)} - 1| / p > 0 \), it follows \( \liminf_{p \to \infty} \Delta_{\min(I)} / p > 0 \), so that the classification rule by (1.3) with (2.6) has (2.2) even when \( \mu_1 = \mu_2 \) and \( n_i \)s are fixed. The classifier by (2.6) draws information about heteroscedasticity via the difference of diagonal elements between the two covariance matrices. The accuracy becomes higher as the difference of those diagonal elements grows. See (III) in Fig. 2.

Next, we consider the classifier by (2.7). From Theorem 2.1 and Proposition 2.1, we have the following result.

**Corollary 2.2.** Assume (A-i). Assume also \( \liminf_{p \to \infty} \lambda_{ij} > 0 \) for \( i = 1, 2 \). Then, for the classification rule by (1.3) with (2.7), we have (2.2) under (C-i'), (C-ii') and the condition that \( \lambda_{ij}^2 / \lambda_{ij} / \lambda_{jj} - 1 / p > 0 \) for \( i = 1, 2, j \neq i \), where \( \Delta_{\min(IV)} = \min\{\Delta_{1(IV)}, \Delta_{2(IV)}\} \).

When \( \Sigma_1 \neq \Sigma_2 \), note that \( \Delta_{\min(IV)} > 0 \) in view of Proposition 1.1. Then, we have the following result.

**Proposition 2.2.** When \( \liminf_{p \to \infty} |tr(\Sigma_i \Sigma_j^{-1}) / p - 1| > 0 \) or \( \liminf_{p \to \infty} \sum_{s=1}^{p} |\lambda_{is} / \lambda_{js} - 1| / p > 0 \) for \( i \neq j \), it follows that \( \liminf_{p \to \infty} \Delta_{i(IV)} / p > 0 \).

Note that \( tr((I_p - \Sigma_i \Sigma_j^{-1})^2) \leq p + tr((\Sigma_i \Sigma_j^{-1})^2) = p + O(tr(\Sigma_i^2)) = o(p^2) \) under (2.8) and \( \liminf_{p \to \infty} \lambda_{ip} > 0 \). Hence, from Corollary 2.2, for the classification rule by (1.3) with (2.7), it holds (2.2) under (A-i), (2.8), \( \liminf_{p \to \infty} \Delta_{\min(IV)} / p > 0 \) and \( \liminf_{p \to \infty} \lambda_{i} > 0 \) for \( i = 1, 2 \). Thus from Proposition 2.2, the accuracy becomes higher as the difference of eigenvalues or eigenvectors between the two covariance matrices grows. See (IV) in Fig. 2.
3. Asymptotic normality of the quadratic classifiers. In this section, we give sufficient conditions for the quadratic classifiers given by (1.2) to hold the asymptotic normality. With the help of the asymptotic normality, we discuss the Bayes error rates for high-dimensional data.

3.1. Preliminary. Let

\[ \delta_i = 2 \left\{ \frac{\text{tr}\{(\Sigma_i A_i^{-1})^2\}}{n_i} + \frac{\text{tr}\{(\Sigma_i A_j^{-1}) \Sigma_j A_j^{-1}\}}{n_j} + \Delta_{iA} \right\}^{1/2} \quad \text{for } i = 1, 2 \quad (j \neq i). \]

Note that \( \delta_i^2 = \text{Var}[2(x_0 - \mu_i)^T \{ A_i^{-1}(\xi_{in_i} - \mu_i) - A_j^{-1}(\xi_{jn_j} - \mu_j + (-1)^i \mu_{12})\}] \) for \( i = 1, 2 \quad (j \neq i) \). Let \( m = \min\{p, n_{\min}\} \). We assume the following conditions when \( m \to \infty \) for \( i = 1, 2 \quad (j \neq i) \):

\begin{align*}
\text{(C-iv)} & \quad \frac{\mu^T A_j^{-1} \Sigma_j A_j^{-1} \mu_{12}}{n_j \delta_i^2} + \frac{\text{tr}\{ (\Sigma_i A_j^{-1})^2 \}}{n_j \delta_i^2} = o(1), \quad \frac{\text{tr}\{ (\Sigma_i A_j^{-1})^4 \}}{n_i \delta_i^2} = o(1), \\
\text{(C-v)} & \quad \frac{\text{tr}\{ (\Sigma_i A_1^{-1} - A_2^{-1})^2 \}}{\delta_i^2} = o(1); \quad \text{and} \quad \text{(C-vi)} \frac{\Delta_{iA}}{\delta_i^2} = o(1).
\end{align*}

From (A.6) in Appendix, under (A-i), (C-iv) and (C-v), it holds that

\[ W_j(A_j) - W_i(A_i) - \Delta_i = 2(x_0 - \mu_i)^T \left\{ A_i^{-1}(\xi_{in_i} - \mu_i) - A_j^{-1}(\xi_{jn_j} - \mu_j + (-1)^i \mu_{12}) \right\} + o_P(\delta_i) \]

as \( m \to \infty \) when \( x_0 \in \pi_i \) for \( i = 1, 2 \quad (j \neq i) \). Under (C-vi), it holds that \( (x_0 - \mu_i)^T A_j^{-1} \mu_{12} = o_P(\delta_i) \) as \( m \to \infty \) when \( x_0 \in \pi_i \) for \( i = 1, 2 \quad (j \neq i) \). Then, we claim the asymptotic normality of (1.2) under (A-i) as follows:

**Theorem 3.1.** Assume (A-i). Assume also (C-iv) to (C-vi). Then, we have that

\[ W_j(A_j) - W_i(A_i) - \Delta_i = \frac{\Delta_i}{\delta_i} \Rightarrow N(0, 1) \quad \text{as } m \to \infty \]

when \( x_0 \in \pi_i \) for \( i = 1, 2 \quad (j \neq i) \),

where “\( \Rightarrow \)” denotes the convergence in distribution and \( N(0, 1) \) denotes a random variable distributed as the standard normal distribution. Furthermore, for the classification rule by (1.3) with (1.2), it holds that

\[ e(i) = \Phi \left( \frac{-\Delta_i}{\delta_i} \right) + o(1) \quad \text{as } m \to \infty \quad \text{for } i = 1, 2 \quad (j \neq i). \]
Let $\delta_{\min} = \min\{\delta_1, \delta_2\}$. Now, instead of (C-iv) to (C-vi), we consider the following conditions when $m \to \infty$:

$$
(C-iv') \quad \frac{||\mu_{12}||^2 \lambda_{\max} + \text{tr}(\Sigma_{\max}^2)/n_{\min}}{n_{\min} \delta_{\min}^2} = o(1) \quad \text{and} \quad \frac{\lambda_{\max}^2}{n_{\min} \delta_{\min}^2} = o(1),
$$

$$
(C-v) \quad \frac{\text{tr}\{(A_1 - A_2)^2\}/\lambda_{\max}}{\delta_{\min}^2} = o(1), \quad \text{and} \quad (C-v') \quad \frac{||\mu_{12}||^2 \lambda_{\max}}{\delta_{\min}^2} = o(1).
$$

**Proposition 3.1.** Assume that $\liminf_{p \to \infty} \lambda_{\min}(A_i) > 0$ for $i = 1, 2$. Then, (C-iv') and (C-v') imply (C-iv) and (C-v), respectively. Furthermore, if $\lambda(A_i) \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$, and $A_i$, $i = 1, 2$, are diagonal matrices such as in (I) to (III) in Section 1, (C-v') implies (C-v).

Next, we consider the asymptotic normality of (1.2) under (A-ii). We assume the following condition instead of (C-vi) when $m \to \infty$ for $i = 1, 2$ ($j \neq i$):

$$
(C-vii) \quad \frac{\sum_{s=1}^{q_i} \{\gamma_{is}^T A^{-1}_j \mu_{12}\}^4}{\delta_i^4} = o(1).
$$

Note that $\sum_{s=1}^{q_i} \{\gamma_{is}^T A^{-1}_j \mu_{12}\}^4 \leq \sum_{s=t=1}^{q_i} \{\gamma_{is}^T A^{-1}_j \mu_{12}\}^2 \{\gamma_{it}^T A^{-1}_j \mu_{12}\}^2 = \Delta_i^2 A$. Thus (C-vii) is milder than (C-vi). The condition (C-vii) can be reduced to eigenvalues and eigenvectors such as in the following remark.

**Remark 3.** If $\Gamma_i = H_i A_i^{1/2}$, $A_i = \Sigma_i$, $i = 1, 2$, and $\Sigma_1 = \Sigma_2$, it holds that $\sum_{s=1}^{q_i} \{\gamma_{is}^T A^{-1}_j \mu_{12}\}^4 = \sum_{s=1}^{p} \psi_s^2$ and $\Delta_i A = \mu_{12}^T \Sigma_i^{-1} \mu_{12} = \sum_{s=1}^{p} \psi_s$, where $\psi_s = (\mu_{12}^T h_{is})^2/\lambda_{is}$. Hence, when $\sum_{s=1}^{p} \psi_s^2/(\sum_{s=1}^{p} \psi_s)^2 \to 0$ as $p \to \infty$, (C-vii) is satisfied.

Now, we claim the asymptotic normality of (1.2) under (A-ii) as follows:

**Theorem 3.2.** Assume (A-ii). Assume also (C-iv), (C-v) and (C-vii). Then, we have (3.1). Furthermore, for the classification rule by (1.3) with (1.2), we have (3.2).

3.2. Bayes error rates. From Theorem 3.2, under the assumptions of Theorem 3.2 and

$$
(3.3) \quad \text{tr}\{(\Sigma_i A_i^{-1})^2\}/n_i + \text{tr}(\Sigma_i A_j^{-1} \Sigma_j A_j^{-1})/n_j = o(\Delta_i A) \quad \text{as} \quad m \to \infty
$$

for $i = 1, 2$ ($j \neq i$), it holds that

$$
e(i) = \Phi\{-\Delta_i/(2\Delta_i^{1/2})\} + o(1) \quad \text{for} \quad i = 1, 2 \quad (j \neq i).$$
Note that $\delta_i/(2\Delta_i A_i^{1/2}) = 1 + o(1)$ under (3.3). If $\Sigma_1 = \Sigma_2 (= \Sigma)$, the ratio $\Delta_i/\Delta_i A_i^{1/2}$ has a maximum when $A_1 = A_2 = \Sigma$. Then, the ratio becomes the Mahalanobis distance such as $\Delta_i/\Delta_i A_i^{1/2} = \Delta_{MD}$, so that the classification rule by (1.3) has an error rate converging to the Bayes error rate in the sense that $e(i) = \Phi(-\Delta_{MD}/2) + o(1)$ for $i = 1, 2$. On the other hand, if $\Sigma_1 \neq \Sigma_2$ and $\pi_i$s are Gaussian, under (C-iii) for (IV), the Bayes optimal classifier by (1.1) becomes as follows:

$$(1)^{i+1} 2(x_0 - \mu_i)^T \Sigma_i^{-1} \mu + o(\Delta_i A_i^{1/2}) < (1)^{i+1} 2(\Delta_i A_i^{1/2})$$

when $x_0 \in \pi_i (j \neq i)$. Note that $\text{Var}\{ (x_0 - \mu_i)^T \Sigma_i^{-1} \mu_2 \} = \mu_2^T \Sigma^{-1} \Sigma_i \Sigma_i^{-1} \mu_2$ (hereafter called $\Delta_i A_i^{1/2}$) when $x_0 \in \pi_i (j \neq i)$ and $\Delta_i A_i^{1/2}$ is the same as $\Delta_i A_i^{1/2}$ for (IV). Hence, $(x_0 - \mu_i)^T \Sigma_i^{-1} \mu_2/\Delta_i A_i^{1/2}$ is distributed as $N(0, 1)$ when $x_0 \in \pi_i : N_p(\mu_i, \Sigma_i)$. Then, the Bayes error rate becomes $e(i) = \Phi\{ -\Delta_i A_i^{1/2} / (2\Delta_i A_i^{1/2}) \} + o(1)$ for $i = 1, 2$, under some conditions.

In general, under the conditions of Theorem 3.2 and

$$(3.4) \quad p/n_i + \text{tr}(\Sigma_i \Sigma_i^{-1})/n_j = o(\Delta_i A_i^{1/2}) \quad \text{as} \quad m \to \infty$$

for $i = 1, 2 (j \neq i)$, the classification rule by (1.3) with (2.7) has the Bayes error rate asymptotically even when $\pi_i$s are non-Gaussian as long as (A-ii) holds. Note that (3.4) is equivalent to (3.3) for (IV) and (3.4) usually holds when $n_{\min} \to \infty$ while $p$ is fixed or $p \to \infty$ but $n_{\min}/p \to \infty$. If (3.4) is not met, the classifier by (2.7) is not optimal. We emphasize that (3.4) does not always hold for high-dimensional settings such as $n_{\min}/p \to 0$ or $n_{\min}/p \to c (> 0)$. For example, we consider the setup of Fig. 1. The condition “$p/n_i = o(\Delta_i A_i^{1/2})$” is not met from the facts that $\Delta_i A_i^{1/2} = O(p^{2/3})$ and $n_1 = n_2 = o(p^{1/3})$, so that (3.4) does not hold. On the other hand, (C-iv) to (C-vi) hold, so that one can claim the asymptotic normality in Theorem 3.1. Note that (3.4) does not hold under (C-vi) for (IV). Thus the error rate of the classifier based on the Mahalanobis distance does not converge to the Bayes error rate when Theorem 3.1 is claimed. Such situations frequently occur in HDLSS settings such as $n_{\min}/p \to 0$. This is the reason why the classifier based on the Mahalanobis distance does not always give a preferable performance for high-dimensional data even when $n_{\min} \to \infty$, $\Sigma_i$s are known and $\pi_i$s are Gaussian.

3.3. Asymptotic normality for (I) to (IV). We consider $\delta_i$s for $i \neq j$. In (I), when $A_{ij} = I_p$, $j = 1, 2$, they are given by

$$\delta_i = 2\left\{ \frac{\text{tr}(\Sigma_i^2)}{n_i} + \frac{\text{tr}(\Sigma_i \Sigma_j)}{n_j} + \mu_2^T \Sigma_i \mu_2 \right\}^{1/2} \quad \text{(hereafter called $\delta_i(I)$)}.$$
In (II), when $A_j = \{\text{tr}(\Sigma_j)/p\} I_p$, $j = 1, 2$, they are given by

$$\delta_i = \frac{2p}{\text{tr}(\Sigma_j)} \left\{ \frac{\delta_{i(I)}^2}{\delta_{i(I)}} + \frac{\text{tr}(\Sigma_j^2)}{\text{tr}(\Sigma_j)^2} \right\}^{1/2}$$

(hereafter called $\delta_{i(II)}$).

In (III), when $A_j = \Sigma_{j(d)}$, $j = 1, 2$, they are given by

$$\delta_i = 2\left\{ \frac{\text{tr}(\Sigma_i\Sigma_{i(d)}^{-1})^2}{n_i} + \frac{\text{tr}(\Sigma_i\Sigma_{j(d)}^{-1}\Sigma_j\Sigma_{j(d)}^{-1})}{n_j} + \frac{\mu_{12}^T \Sigma_{j(d)}^{-1} \Sigma_i \Sigma_{j(d)}^{-1} \mu_{12}}{n_j} \right\}^{1/2}$$

(hereafter called $\delta_{i(III)}$).

In (IV), when $A_j = \Sigma_i$, $j = 1, 2$, they are given by

$$\delta_i = 2\left\{ \frac{p}{n_i} + \frac{\text{tr}(\Sigma_i\Sigma_{i(d)}^{-1})}{n_j} + \frac{\mu_{12}^T \Sigma_{j(d)}^{-1} \Sigma_i \Sigma_{j(d)}^{-1} \mu_{12}}{n_j} \right\}^{1/2}$$

(hereafter called $\delta_{i(IV)}$).

From Theorems 3.1, 3.2 and Proposition 3.1, we have the following result for (I) to (III).

**Corollary 3.1.** Assume (C-iv'). Assume either (A-i) and (C-vi') or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (2.4), we have (3.2). Furthermore, under (C-v'), for the classification rule by (1.3) with (2.5) or (2.6), we have (3.2).

For DBDA, Aoshima and Yata [2] gave a partial result of Corollary 3.1 under different conditions. When comparing (3.2) of DBDA and the classifier by (2.5), we give the following remark.

**Remark 4.** When $\text{tr}(\Sigma_1)/\text{tr}(\Sigma_2) \rightarrow 1$ as $p \rightarrow \infty$, it holds $\{\delta_{i(I)}p/\text{tr}(\Sigma_j)\}/\delta_{i(II)} = 1 + o(1)$. Note that $\Delta_{i(I)}\text{tr}(\Sigma_j)/p \geq \Delta_{i(I)}$. Then, it follows that $\Delta_{i(I)}/\delta_{i(I)} \leq \Delta_{i(I)}/\delta_{i(II)}$ for sufficiently large $p$.

From Theorems 3.1 and 3.2 and Proposition 3.1, we have the following result for (IV).

**Corollary 3.2.** Assume that (C-iv'), $\lim_{p \rightarrow \infty} \lambda_{ip} > 0$ and $\text{tr}\{(I_p - \Sigma_i\Sigma_{j(i)}^{-1})^2\}/\delta_{\min(IV)}^2 = o(1)$ for $i = 1, 2$, $j \neq i$, where $\delta_{\min(IV)} = \min\{\delta_{1(IV)}, \delta_{2(IV)}\}$. Assume either (A-i) and (C-v'ii) or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (2.7), we have (3.2).
3.4. Comparisons of the classifiers. In this section, we investigate performances of the classifiers by (I) to (IV) in (1.2). We compare the amount of \( \Delta_i/\delta_i \) in (3.2) for (I) to (IV). We first examine (I) and (II). As mentioned in Section 2.2, the classifier by (2.5) gives a better performance compared to (2.4) when \( \text{tr}(\Sigma_1) \neq \text{tr}(\Sigma_2) \). However, if one cannot assume (A-i) or (C-iii), one may use the classifier by (2.4) free from the assumptions.

Next, we examine (I), (III) and (IV) in the setup of Fig. 1. We plotted the three asymptotic error rates, \( \Phi(-\Delta_{i(I)}/\delta_{i(I)}) \), \( \Phi(-\Delta_{i(III)}/\delta_{i(III)}) \) and \( \Phi(-\Delta_{i(IV)}/\delta_{i(IV)}) \) in Fig. 3. Also, we plotted the Bayes error rate, \( \Phi(-\Delta_{1/MD}/2) \). We laid the average error rate, \( \overline{e} = \{e(1) + e(2)/2 \} \), for (2.4), (2.6) and (2.7) by borrowing from Fig. 1. Note that (I), (III) and (IV) satisfy (C-iv) to (C-vi) from the facts that \( n_{\min} = o(p^{1/3}) \), \( \lambda_{i1} = O(1) \) and \( \Delta_{iA} = O(\|\mu_{12}\|) = O(p^{2/3}) \) for \( i = 1, 2 \). Thus, Theorem 3.1 is claimed for (I), (III) and (IV). From (3.2), we note that \( e(1) - e(2) = o(1) \) when \( \Sigma_1 = \Sigma_2 \). Thus, \( \overline{e} \) is regarded as an estimate of \( e(1) \). We observed that the classifier by (I) or (III) gives adequate performances, however, that by (IV) does not perform well. Note that the classifier based on the Mahalanobis distance does not converge to the Bayes error rate when Theorem 3.1 is claimed. See Section 3.2 for the details. As for (I) and (III), the difference of the performance depends on the configuration of \( \mu_{ij}s \) and \( \sigma_{ij}s \). When \( p \) is sufficiently large, we note that \( \Delta_{i(I)} = \sum_{s=1}^p \mu_{i2s}^2 < \Delta_{i(III)} = \sum_{s=1}^p \mu_{i2s}^2/\sigma_{2(s)} \) for (a) and \( \Delta_{i(I)} > \Delta_{i(III)} \) for (b) because \( \sigma_{2(s)} = 0.5 + s/(p+1) \), \( s = 1, \ldots, p \) for (a) and (b). When \( p \) is sufficiently large, it follows that \( \Delta_{i(I)}/\delta_{i(I)} < \Delta_{i(III)}/\delta_{i(III)} \) for (a) and \( \Delta_{i(I)}/\delta_{i(I)} > \Delta_{i(III)}/\delta_{i(III)} \) for (b). Thus for (a), the classifier by (III) is better than that by (I), however, they trade places for (b). See the dashed lines in Fig. 3.

When \( \Sigma_1 \neq \Sigma_2 \), the classifier by (II), (III) or (IV) draws information...
about heteroscedasticity through the difference of \( \text{tr}(\Sigma_j)s \), \( \Sigma_{j(d)}s \) or \( \Sigma_js \), respectively. We checked their performances in the setup of Fig. 2. For (c), we note that \( \Delta_{(I)} = 0 \) but \( \Delta_{(II)} = \Delta_{(III)} = \Delta_{(IV)} > cp \) for some constant \( c > 0 \), so that the three classifiers hold the consistency property even when \( n_i \)s are fixed because (C-i) to (C-iii) hold. We observed that the three classifiers gave preferable performances by using the difference of \( \text{tr}(\Sigma_j)s \), \( \Sigma_{j(d)}s \) or \( \Sigma_js \) as \( p \) increases. For (d), we note that the difference of \( \text{tr}(\Sigma_j)s \) is smaller than that for (c). We observed that the classifier by (II) gives a worse performance for (d) compared to (c). On the other hand, the classifier by (III) gave a better performance compared to (II) because \( \Delta_{(III)} \) is sufficiently larger than \( \Delta_{(II)} \) for (d) when \( p \) is large. The classifier by (IV) draws information about heteroscedasticity from the difference of the covariance matrices themselves, so that it gave the best performance. However, we note that it is quite difficult to estimate \( \Sigma_j^{-1}s \) for high-dimensional data in actual data analyses. See Section 5.2 for details.

4. Estimation of the quadratic classifiers. We denote an estimator of \( A_j \) by \( \hat{A}_j \). We assume that \( \hat{A}_j \)s are positive definite matrices w.p.1. We consider the classifier by \( W_j(\hat{A}_j) \).

4.1. Preliminary. Let \( \|M\| = \lambda_{\max}^{1/2}(M^TM) \) for any square matrix \( M \). Let \( \kappa \) be a constant such as \( \kappa = \Delta_{\min} \) or \( \kappa = \delta_{\min} \). We consider the following condition for \( \hat{A}_j \)s under (\( \star \)): 

\( (\text{C-viii}) \quad p\|\hat{A}_i^{-1} - A_i^{-1}\| = o_P(\kappa) \) for \( i = 1, 2 \).

Proposition 4.1. Assume (C-viii). Assume also that \( \lambda(A_i) \in (0, \infty) \) as \( p \to \infty \) for \( i = 1, 2 \). Then, we have that 

\( W_1(\hat{A}_1) - W_2(\hat{A}_2) = W_1(A_1) - W_2(A_2) + o_P(\kappa) \)

under (\( \star \)) when \( x_0 \in \pi_i \) for \( i = 1, 2 \).

When \( A_1 = A_2 (= A) \) and \( \hat{A}_1 = \hat{A}_2 (= \hat{A}) \), we consider the following condition for \( \hat{A} \) under (\( \star \)): 

\( (\text{C-ix}) \quad (p/n_{\min}^{1/2} + p^{1/2}\|\mu_{12}\|)\|\hat{A}^{-1} - A^{-1}\| = o_P(\kappa) \).

We have the following result.

Proposition 4.2. Assume (C-ix). Then, we have (4.1).

We note that (C-ix) is milder than (C-viii) from the fact that \( \|\mu_{12}\| = O(p^{1/2}) \). Hence, we recommend to use the quadratic classifiers when the
difference of covariance matrices is considerably large. Otherwise one is recommended to use the linear classifier such as (2.4) or (4.5). See Section 4.3 for the details.

4.2. Quadratic classifier by \( \hat{A}_j = \{ \text{tr}(S_j)/p \} I_p \), \( j = 1, 2 \). We consider the classifier by

\[
W_j(\{ \text{tr}(S_{jn_j})/p \} I_p) = \frac{p||x_0 - \overline{x}_{jn_j}||^2}{\text{tr}(S_{jn_j})} - \frac{p}{n_j} + p \log \{ \text{tr}(S_{jn_j})/p \}.
\]

(4.2) Note that \( \delta_i = \delta_{i(I)} \), \( \Delta_i = \Delta_{i(I)} \) and \( A_j = \{ \text{tr}(\Sigma_j)/p \} I_p \). By combining Corollary 2.1 with Proposition 4.1, we have the following result.

**Corollary 4.1.** Assume (A-i). Assume also (C-i’) and (C-ii’). Then, for the classification rule by (1.3) with (4.2), we have (2.2).

Aoshima and Yata [2] gave Corollary 4.1 under different conditions. The classifier by (4.2) depends on a geometric representation. See Section 2.2. Hereafter, we call the classifier by (4.2) the “geometrical quadratic discriminant analysis (GQDA)”. Similar to Section 2.2, we have (2.2) for GQDA under (A-i) and (2.3) even when \( n_{\min} \) is fixed. If one can assume that \( \liminf_{p \to \infty} |\text{tr}(\Sigma_1)/\text{tr}(\Sigma_2) - 1| > 0 \), we have (2.2) for GQDA under (A-i) and (2.8) even when \( n_{\min} \) is fixed and \( \mu_1 = \mu_2 \). As for the asymptotic normality, by combining Corollary 3.1 with Lemma B.3 given in Appendix B, we have the following result.

**Corollary 4.2.** Assume (C-iv’) and (C-v’). Assume either (A-i) and (C-vii’) or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (4.2), we have (3.2) under \( (\text{tr}(\Sigma_1)/\text{tr}(\Sigma_2) - 1)^2 \text{tr}(\Sigma_{\max}^2)/(n_{\min} \delta_{\min(I)}) = o(1) \) as \( m \to \infty \), where \( \delta_{\min(I)} = \min \{ \delta_{i(I)}, \delta_{2(I)} \} \).

Now, we compare DBDA with GQDA. We have that

\[
\hat{\Delta}_i = ||\overline{x}_{1n_1} - \overline{x}_{2n_2}||^2 - \text{tr}(S_{1n_1})/n_1 - \text{tr}(S_{2n_2})/n_2 \quad \text{and}
\]

\[
\hat{\Delta}_{i(I)} = \frac{p}{\text{tr}(S_{jn_j})} [\hat{\Delta}_i + \text{tr}(S_{jn_i}) - \text{tr}(S_{jn_j}) + \text{tr}(S_{jn_j}) \log \{ \text{tr}(S_{jn_j}) \}]
\]

for \( i = 1, 2 \) (\( j \neq i \)). Note that \( E(\hat{\Delta}_i) = \Delta_i \). From (3.2) and Remark 4, if \( \hat{\Delta}_{i(I)} \text{tr}(S_{jn_i})/p \) is sufficiently larger than \( \hat{\Delta}_i \) for some \( i \), we recommend to use GQDA. Otherwise one may use DBDA free from (A-i). See Corollary 2.1 for the details.
4.3. **Quadratic classifier by** \( \hat{A}_j = S_{jn_j(d)} \), \( j = 1, 2 \). Let \( S_{jn_j(d)} = \text{diag}(s_{jn_j(1)}, \ldots, s_{jn_j(p)}) \). We consider the classifier by

\[
W_j(S_{jn_j(d)}) = \sum_{r=1}^{p} \left( \frac{(x_{0r} - \mu_{jr})^2}{s_{jn_j(r)}} - \frac{1}{n_j} + \log s_{jn_j(r)} \right).
\]

Note that \( \delta_i = \delta_i(III) \), \( \Delta_i = \Delta_i(III) \) and \( A_j = \Sigma_j(d) \). Dudoit et al. [11] considered the quadratic classifier without the bias correction term. That was called the diagonal quadratic discriminant analysis (DQDA). Hereafter, we call the classifier by (4.3) “DQDA-bc”. Let \( \eta_i(s) = \text{Var}\{ (x_{isj} - \mu_{is})^2 \} \) for \( i = 1, 2, \) and \( s = 1, \ldots, p \) \( (j = 1, \ldots, n_i) \). We consider the following assumption:

\( \text{(A-iii)} \) \( \eta_i(s) \in (0, \infty) \) as \( p \to \infty \) and \( \limsup_{p \to \infty} E\{ \exp(t_{is} |x_{isj} - \mu_{is}|^2/\eta_i^{1/2}(s)) \} < \infty \) for some \( t_{is} > 0 \), \( i = 1, 2 \), and \( s = 1, \ldots, p \) \( (j = 1, \ldots, n_i) \).

Note that (A-iii) is satisfied when \( \pi_i \) has \( N_p(\mu_i, \Sigma_i) \) for \( i = 1, 2 \). By combining Corollary 2.1 with Proposition 4.1, we have the following result.

**Corollary 4.3.** Assume (A-i) and (A-iii). Assume also (C-ii'). Then, for the classification rule by (1.3) with (4.3), we have (2.2) under the condition that

\[
p^2 \log p/n_{\min} \Delta_{\min}^2(III) = o(1).
\]

Note that (C-i') holds under (4.4). From the fact that \( \Delta_{\min}(III) = O(p) \), it follows that \( n_{\min}^{-1} \log p = o(1) \) under (4.4). Similar to Section 2.2, if one can assume \( \liminf_{p \to \infty} ||\mu_{12}||^2/p > 0 \) or \( \liminf_{p \to \infty} \sum_{s=1}^{p} |\sigma_{1(s)}/\sigma_{2(s)} - 1|/p > 0 \), DQDA-bc holds (2.2) under (A-i), (A-iii), (2.8) and \( n_{\min}^{-1} \log p = o(1) \). When \( \Delta_{\min}(III) \) is not sufficiently large, say \( \Delta_{\min}(III) = O(p^{1/2}) \), we can claim the conclusion of Corollary 4.3 in high-dimension, large-sample-size settings such as \( n_{\min}/p \to \infty \). In Section 5, we give a DQDA type classifier by feature selection which holds the consistency property even when \( n_{\min}/p \to 0 \) and \( \Delta_{\min}(III) \) is not sufficiently large.

As for the asymptotic normality, by combining Corollary 3.1 with Proposition 4.1, we have the following result.

**Corollary 4.4.** Assume (A-ii) and (A-iii). Assume also (C-v') and (C-vii). Then, for the classification rule by (1.3) with (4.3), we have (3.2) under \( p^2 \log p/(n_{\min} \delta_{\min}^2(III)) = o(1) \), where \( \delta_{\min}(III) = \min\{\delta_1(III), \delta_2(III)\} \).
Under (C-vi), we note that $\frac{\delta^2(t_{III})}{\Delta_{III}} = O(p^2/n_{\min})$, so that the condition $\frac{p^2 \log p}{(n_{\min} \delta^2_{III})} = o(1)$ does not hold. Thus, one cannot claim the asymptotic normality for DQDA-bc under (C-vi).

Next, we consider the pooled sample diagonal matrix, $S_{n(d)} = \sum_{i=1}^{2} (n_i - 1)S_{in_i(d(i)}/(\sum_{i=1}^{2} n_i - 2)$. Note that $E(S_{n(d)}) = \sum_{i=1}^{2} (n_i - 1)\Sigma_i(d(i)/\sum_{i=1}^{2} n_i - 2)$ (hereafter called $\Sigma(d)$). When $\Sigma_1(d) = \Sigma_2(d)$, it follows that $\Sigma(d) = \Sigma_i(d)$, $i = 1, 2$. Let us write $S_{n(d)} = \text{diag}(s_{n(1)}, \ldots, s_{n(p)})$ and $\Sigma(d) = \text{diag}(\sigma_1, \ldots, \sigma_p)$. We consider the classifier by

$$W_j(S_{n(d)}) = \sum_{r=1}^{p} \left( \frac{(x_{0r} - \bar{x}_{jrn_j})^2}{s_{n(r)}} - \frac{s_{jn_j(r)}}{n_j s_{n(r)}} \right).$$

We note that the classification rule by (1.3) with (4.5) becomes a linear classifier. Bickel and Levina [5] and Dudoit et al. [11] considered the linear classifier without the bias correction term. That was called the diagonal linear discriminant analysis (DLDA). Hereafter, we call the classifier by (4.5) “DLDA-bc”. Although Huang et al. [16] gave bias corrected versions of DLDA and DQDA, they considered a bias correction only when $\pi_i$s are Gaussian. We note that $\Delta_1 = \Delta_2 = \sum_{s=1}^{p} \mu_{12s}/\sigma(s)$ (hereafter called $\Delta_{III'}$) and $A_1 = A_2 = \Sigma(d)$. Then, by combining Theorem 2.1 with Propositions 2.1 and 4.2, we have the following result.

**Corollary 4.5.** Assume (A-iii). Assume also (C-i') and (C-ii'). Then, for the classification rule by (1.3) with (4.5), we have (2.2) under the condition that

$$\frac{p \log p}{n_{\min} \Delta_{III'}} = o(1).$$

Under $n_{\min} \log p = o(1)$, one may claim that (4.6) is milder than (4.4) if $\Delta_{n_{\min}(III)}$ and $\Delta_{(III')}$ are of the same order. Hence, we recommend to use DQDA-bc when $\Delta_{n_{\min}(III)}$ is considerably larger than $\Delta_{(III')$. Otherwise one may use DLDA-bc even when $\Sigma_i(d)$s are not common.

4.4. **Quadratic classifier by $\hat{A}_j = S_{jn_j}$, $j = 1, 2$.** In this section, we consider high-dimension, large-sample-size settings such as $n_{\min}/p \to \infty$ as $p \to \infty$. We consider the classifier by

$$W_j(S_{jn_j}) = (x_j - \bar{x}_{jn_j})^T S_{jn_j}^{-1}(x_j - \bar{x}_{jn_j}) - p/n_j + \log |S_{jn_j}|.$$

Note that $\delta_i = \delta_{i(IV)}$, $\Delta_i = \Delta_{i(IV)}$ and $A_j = \Sigma_j$. Let $\eta_{i(r,s)} = \text{Var}\{(x_{irl} - \mu_{irl})(x_{isl} - \mu_{is})\}$ for $r, s = 1, \ldots, p$ $(i = 1, 2)$. By combining Theorem 2.1 with Proposition 4.1, we have the following result.
Corollary 4.6. Assume (A-i) and (A-iii). Assume also \( \lambda(\Sigma_i) \in (0, \infty) \) as \( p \to \infty \) and \( \liminf_{p \to \infty} n_{i(r,s)} > 0 \) for all \( r, s; i = 1, 2 \). Then, for the classification rule by (1.3) with (4.7), we have (2.2) under the conditions that \( p^{1/2}/\Delta_{\min(IV)} = o(1) \) and

\[
\frac{p^4 \log p}{n_{\min} \Delta^2_{\min(IV)}} = o(1).
\]

From the fact that \( \Delta_{ij(IV)} = O(p) \) when \( \lambda(\Sigma_i) \in (0, \infty) \) as \( p \to \infty \) for \( i = 1, 2 \), it follows that \( n_{\min} p^2 \log p = o(1) \) under (4.8). Thus, we recommend to use the classification rule by (1.3) with (4.7) only when \( n_{\min} p^2 \log p = o(1) \). However, the condition “\( n_{\min} p^2 \log p = o(1) \)” is quite strict for high-dimensional data. Hence, we consider estimating sparse inverse covariance matrices when \( n_{\min}/p \to 0 \). In Section 5, we give a classifier by sparse inverse covariance matrix estimation.

5. Quadratic classifiers by feature selection and sparse inverse covariance matrix estimation. In this section, we find some alternative quadratic classifiers for (4.3) and (4.7).

5.1. Quadratic classifier after feature selection. We consider applying a variable selection procedure to classification. Fan and Fan [12] proposed feature annealed independent rules based on the difference of mean vectors. However, we consider the difference of the classes not only for mean vectors but also for covariance matrices. We have that

\[
\Delta_{1(III)} + \Delta_{2(III)} = \sum_{r=1}^{p} \sum_{i \neq j}^{2} \left( \frac{\mu_{12r}^2 + \sigma_{i(r)}}{\sigma_{j(r)}} - 1 \right).
\]

Let \( \theta_r = \sum_{i \neq j}^{2} \{\mu_{2r}^2 + \sigma_{i(r)}\}/(2\sigma_{j(r)}) - 1 \) for \( r = 1, \ldots, p \). Note that \( \Delta_{1(III)} + \Delta_{2(III)} = 2 \sum_{r=1}^{p} \theta_r \). Also, note that \( \theta_r > 0 \) when \( \mu_{1r} \neq \mu_{2r} \) or \( \sigma_{1(r)} \neq \sigma_{2(r)} \).

Now, we give an estimator of \( \theta_r \) by

\[
\hat{\theta}_r = \sum_{i \neq j}^{2} \frac{(x_{1rn_1} - x_{2rn_2})^2 + s_{in_i(r)}}{2s_{jn_j(r)}} - 1.
\]

Then, we have the following result.

Theorem 5.1. Assume (A-iii). Assume also \( n_{\min}^{-1} \log p = o(1) \). Then, we have that as \( p \to \infty \)

\[
\max_{r=1, \ldots, p} |\hat{\theta}_r - \theta_r| = O_P\{n_{\min}^{-1} \log p \}^{1/2}.
\]
Let \( D = \{ r \mid \theta_r > 0 \text{ for } r = 1, \ldots, p \} \) and \( p_* = \#D \), where \( \#S \) denotes the number of elements in a set \( S \). Let \( \xi = (n_{\text{min}}^{-1} \log p)^{1/2} \). We select a set of significant variables by

\[
(5.1) \quad \hat{D} = \{ r \mid \hat{\theta}_r > \xi \gamma \text{ for } r = 1, \ldots, p \},
\]

where \( \gamma \in (0, 1) \) is a chosen constant. Then, from Theorem 5.1, we have the following result.

**Corollary 5.1.** Assume (A-iii) and \( n_{\text{min}}^{-1} \log p = o(1) \). Assume also \( \liminf_{p \to \infty} \theta_r > 0 \text{ for all } r \in D \). Then, we have that \( P(D = \hat{D}) \to 1 \) as \( p \to \infty \).

**Remark 5.** As for \( k (\geq 3) \)-class classification, one may consider \( \hat{\theta}_r \) such as

\[
\hat{\theta}_r = \sum_{i \neq j} k (k-1) s_{jn_j(r)} \bigg/ \{k(k-1) s_{jn_j(r)} \} - 1 \text{ for } r = 1, \ldots, p.
\]

Now, we consider a classifier using only the variables in \( \hat{D} \). We define the classifier by

\[
(5.2) \quad W_j(S_{j(d)}) = \sum_{r \in \hat{D}} \left( \frac{\langle x_{0r} - \bar{x}_{jr} \rangle^2}{s_{jn_j(r)}} - \frac{1}{n_j} + \log s_{jn_j(r)} \right)
\]

for \( j = 1, 2 \). We consider the classification rule by (1.2) with (5.2). We call this feature selected DQDA “FS-DQDA”. Let us write that \( x_{i*} = (x_{ir_1}, \ldots, x_{ir_p})^T \) for all \( i, j \), where \( D \) = \{\( r_1, \ldots, r_p \}\}. Let \( \Sigma_{i*} = \text{Var}(x_{i*}) \) for \( i = 1, 2 \text{ (} j = 1, \ldots, n_i \text{).} \) Then, from Theorem 2.1 and Corollary 5.1, we have the following result.

**Corollary 5.2.** Assume (A-i) and (A-iii). Assume also \( \lambda_{\text{max}}(\Sigma_{i*}) = o(p_*) \) for \( i = 1, 2 \), and \( \liminf_{p \to \infty} \theta_j > 0 \text{ for all } j \in D \). Then, for the classification rule by (1.3) with (5.2), we have (2.2) under \( n_{\text{min}}^{-1} \log p = o(1) \).

By comparing Corollary 5.2 with 4.3, note that the condition “\( n_{\text{min}}^{-1} \log p = o(1) \)” is much milder than (4.4). Thus we recommend FS-DQDA more than DQDA-bc (or the original DQDA). For a choice of \( \gamma \in (0, 1) \) in (5.1), we recommend applying cross-validation procedures or choosing a constant such as \( \gamma = 0.5 \) because Corollary 5.2 is claimed for any \( \gamma \in (0, 1) \). In addition, we emphasize that the computational cost of FS-DQDA is quite low even when \( p \geq 10,000 \).
5.2. Quadratic classifier by sparse inverse covariance matrix estimation. We consider applying a sparse estimation of inverse covariance matrices to classification. Bickel and Levina [6] considered a sparse estimator of $\Sigma^{-1}$ and gave the following result: Let $\sigma_{ij}$ be the $(s,t)$ element of $\Sigma$ for $s,t = 1, \ldots, p$ ($i = 1, 2$). A sparsity measure of $\Sigma$ ($i = 1, 2$) is given by $c_{p,h_i} = \max_{1 \leq i \leq p} \sum_{k=1}^{p} |\sigma_{i}(st)|^{h_i}$ for $0 \leq h_i < 1$. Note that $c_{p,0} = O(p)$. If $c_{p,h_i}$ is much smaller than $p$ for a constant $h_i \in [0, 1)$, $\Sigma$ is considered as sparse in the sense that many elements of $\Sigma$ are very small. See Section 3 in [20] for the details. Let $I(\cdot)$ be the indicator function. A thresholding operator is defined by $T_t(M) = [m_{ij}I(|m_{ij}| \geq t)]$ for any $t > 0$ and any symmetric matrix $M = [m_{ij}]$. Let $t_{n_i} = M'(n_i^{-1} \log p)^{1/2}$ for some constant $M' > 0$.

**Theorem 5.2** (see [6]). Assume (A-iii), $n_i^{-1} \log p = o(1)$ and $\lambda(\Sigma_i) \in (0, \infty)$ as $p \to \infty$ for a sufficiently large $M'(>0)$, it holds that as $p \to \infty$

$$||\{T_{t_{n_i}}(S_{m_i})\}^{-1} - \Sigma_i^{-1}|| = O_P\left(c_{p,h_i}(n_i^{-1} \log p)^{(1-h_i)/2}\right).$$

**Remark 6.** Theorem 5.2 is obtained by Theorem 1 and Section 2.3 in [6].

By combining Theorem 5.2 and Proposition 4.1, if it holds that $\lambda(\Sigma_i) \in (0, \infty)$ as $p \to \infty$ and

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

for $W_j(T_{t_{n_i}}(S_{m_i})), \lambda = \Delta_{\min(IV)}$. Thus one can apply $W_j(T_{t_{n_i}}(S_{m_i}))$ to the classification rule by (1.3). If $\liminf_{p \to \infty} \Delta_{\min(IV)} / p > 0$ and $\Sigma_i$s are sparse such as $c_{p,h_i} = O(1)$ for some $h_i$, $i = 1, 2$, (5.3) holds in HDLSS settings such as $n_i^{-1} \log p = o(1)$. Li and Shao [17] and Shao et al. [20] considered a linear and a quadratic classifier by the sparse estimation of $\Sigma_i^{-1}$s under some sparsity conditions. On the other hand, Cai et al. [8] gave a constrained $\ell_1$-minimization for inverse matrix estimation (CLIME). One may apply the CLIME to the classification rule by (1.3). However, one should note that the computational cost for the sparse estimation of $\Sigma_i^{-1}$s is extremely high even when $p \approx 1000$. It is quite unrealistic to apply the estimation to classification when $p$ is very high such as $p \geq 10,000$.

5.3. Simulation. We used computer simulations to compare the performance of the classifiers: DBDA by (2.4), GQDA by (4.2), DLDA-bc by (4.5), DQDA-bc by (4.3) and FS-DQDA by (5.2). We did not compare the classifiers with the one given by sparse estimation of $\Sigma_i^{-1}$s such as $W_j(T_{t_{n_i}}(S_{m_i}))$. 

HIGH-DIMENSIONAL QUADRATIC CLASSIFIERS

21
in Section 5.2 because the computational cost of the sparse estimation is very high when \( p \) is large. Thus we considered the classifier by (2.7) instead of using the sparse estimation, provided that \( \Sigma \)'s were known. We set \( \gamma = 0.5 \) in (5.1). We considered \( p_* = \lceil p^{1/2} \rceil \). We generated \( x_{ij} \) independently from (i) \( N_p(0, \Sigma_i) \) or (ii) a \( p \)-variate \( t \)-distribution, \( t_p(0, \Sigma_i, \nu) \) with mean zero, covariance matrix \( \Sigma_i \) and degrees of freedom \( \nu \). We set \( p = 2^s, \ s = 3, \ldots, 10 \) for (i), and \( p = 500 \) and \( \nu = 4^s, \ s = 1, \ldots, 8 \) for (ii). We set \( \mu_1 = 0, \mu_2 = (0, \ldots, 0, 1, \ldots, 1)^T \) whose last \( p_+ \) elements are 1 and \( \Sigma = \Sigma_1 = B_1(0,3^{(i-1)/3})B_1 \), where \( B_1 \) is defined by (1.5). Let \( B_2 = \text{diag}(1, \ldots, 1, 2^{1/2}, \ldots, 2^{1/2}) \) whose last \( p_+ \) diagonal elements are \( 2^{1/2} \). We considered four cases:

(a) \( n_1 = 10, \ n_2 = 20 \) and \( \Sigma_2 = \Sigma_1 \) for (i) \( N_p(0, \Sigma_i) \);
(b) \( n_1 = \lceil (\log p)^2 \rceil, \ n_2 = 2n_1 \) and \( \Sigma_2 = \Sigma_1 \) for (i) \( N_p(0, \Sigma_i) \);
(c) \( n_1 = \lceil (\log p)^2 \rceil, \ n_2 = 2n_1 \) and \( \Sigma_2 = B_2 \Sigma_1 B_2 \) for (i) \( N_p(0, \Sigma_i) \); and
(d) \( n_1 = \lceil (\log p)^2 \rceil, \ n_2 = 2n_1 \) and \( \Sigma_2 = B_2 \Sigma_1 B_2 \) for (ii) \( t_p(0, \Sigma_i, \nu) \).

It holds that \( n_{\min}^{-1} \log p = o(1) \) for (b), (c) and (d), \( \liminf_{p \to \infty} \Delta_{\min} / p_* > 0 \) for (a) to (d), and \( \liminf_{p \to \infty} |\text{tr}(\Sigma_1) - \text{tr}(\Sigma_2)| / p_* > 0 \) for (c) and (d). Similar to Section 1, we calculated the average error rate, \( \overline{e} \), by 2000 replications and plotted the results in Fig. 4 (a) to (d).

We observed from (a) in Fig. 4 that DBDA and GQDA give preferable performances when \( n_i \)'s are fixed. DLDA-bc gave a moderate performance because \( \Sigma_1 = \Sigma_2 \). However, the other classifiers did not give preferable performances when \( p \) is large. This is probably due to the consistency property of those classifiers (except (2.7)) which is claimed under at least \( n_{\min}^{-1} \log p = o(1) \). Actually, as for (b), the other classifiers gave moderate performances because \( n_{\min}^{-1} \log p = o(1) \). Thus we do not recommend to use quadratic classifiers including all the elements (or the diagonal elements) of sample covariance matrices, such as DQDA-bc and FS-DQDA, when the condition “\( n_{\min}^{-1} \log p = o(1) \)” is not satisfied. When \( n_{\min}^{-1} \log p \neq o(1) \) or \( n_i \)'s are fixed, we recommend to use DBDA and GQDA. On the other hand, FS-DQDA gave a good performance for (c) because the difference of the covariance matrices becomes large as \( p \) increases. We note that from Corollary 5.2 FS-DQDA holds the consistency property for (c). However, DQDA-bc did not give a preferable performance because \( \Delta_{\min}(III) = O(p^{1/2}) \), so that DQDA-bc does not hold the consistency property from Corollary 4.3. We note that \( \Sigma_1 \neq \Sigma_2 \) but \( \Delta_i(II) / \delta_i(II) \approx \Delta_i(III) / \delta_i(III) \) for (c). Thus GQDA gave a similar performance to DBDA for (c). As for (d), DBDA gave a preferable performance even when \( \nu \) is small because DBDA holds the consistency property without (A-i). The other classifiers did not give preferable perfor-
mances when \( \nu \) is small. However, these classifiers gave moderate performances when \( \nu \) becomes large because \( t_p(0, \Sigma, \nu) \Rightarrow N_p(0, \Sigma) \) as \( \nu \to \infty \). Especially, FS-DQDA gave a good performance when \( \nu \) is not small. This is probably because the classifier by (5.2) has smaller variance due to feature selection, such as \( p_* / p \to 0 \), compared to the other classifiers.

Throughout the simulations, the classifier by (2.7) did not give preferable performances in spite that \( \Sigma_i \)'s are known. See Section 3.2 for theoretical reasons. Therefore, it is likely that the classifier by \( W_j(T_{n_j} (S_{jn_j})) \) gives poor performances for these high-dimensional settings.

6. Example: Leukemia data sets. We first analyzed gene expression data given by Golub et al. [14] in which the data set consists of 7129 (= \( p \)) genes and 72 samples. We had 2 classes of leukemia subtypes, that is, \( \pi_1 \): acute lymphoblastic leukemia (ALL) (47 samples) and \( \pi_2 \): acute myeloid leukemia (AML) (25 samples). The data set consisted of two sets as 38 training samples (ALL: 27 samples and AML: 11 samples) and 34 test samples (ALL: 20 samples and AML: 14 samples). Note that \( S_{1n_1(d)} = S_{2n_2(d)} \) if each sample has unit variance. Thus we did not standardize each sample so as to have unit variance.

First, we checked several sparsity conditions. We standardized each sam-
ple by \( x_{ij}/(\sum_{i=1}^{2} \text{tr}(S_{in_i})/(2p))^{1/2} \) for all \( i \), \( j \), so that \( \text{tr}(S_{1n_1})/2 + \text{tr}(S_{2n_2})/2 = p \). By using all the samples (i.e., 72 samples), we calculated \( \hat{\Delta}_{(I)} = 2060 \) (\( \approx 0.29p \)), where \( \hat{\Delta}_{(I)} \) is given in Section 4.2. From this observation, we concluded that \( \mu_{12} \) is non-sparse. Next, we considered an estimator of \( ||\Sigma_{12}||_{F}^{2} = \sum_{i=1}^{2} \text{tr}(\Sigma_{i}^{2}) - 2 \text{tr}(\Sigma_{1} \Sigma_{2}) \) by \( \hat{\Delta}_{\Sigma} = \sum_{i=1}^{2} W_{in_i} - 2 \text{tr}(S_{1n_1}S_{2n_2}) \) having \( W_{in_i} \) defined by (16) in Aoshima and Yata [2]. Here, \( W_{in_i} \) is an unbiased estimator of \( \text{tr}(\Sigma_{i}^{2}) \), so that \( E(\hat{\Delta}_{\Sigma}) = ||\Sigma_{12}||_{F}^{2} \). We calculated \( \hat{\Delta}_{\Sigma} = 9.77 \times 10^{5} \) (\( \approx 137p \)). From this observation, we concluded that \( \Sigma_{12} \) is non-sparse. Therefore, the Bayes error rates of this data set are probably close to 0. Also, we calculated \( (\hat{\lambda}_{11}, \hat{\lambda}_{12}) = (1223, 1457) \) (\( \approx (0.172p, 0.204p) \)), where \( \hat{\lambda}_{11} \) is an estimate of the largest eigenvalue due to the noise-reduction methodology by Yata and Aoshima [23]. We concluded that “\( \lambda(\Sigma_{i}) \in (0, \infty) \) as \( p \to \infty \)” does not hold and \( \Sigma_{i} \)'s are non-sparse because \( \lambda_{1i} \) are very large. Therefore, we do not recommend to apply the classifier by the sparse estimation of \( \Sigma_{i}^{-1} \), such as \( W_{j}(T_{in_j}(S_{jn_j})) \). Actually, we did not use any classifiers by sparse estimation of \( \Sigma_{i}^{-1} \) in this section. Also, note that the computational cost for the sparse estimation of \( \Sigma_{i}^{-1} \) is very high when \( p \) is large.

We constructed the classifiers: DBDA, GQDA, DQDA-bc, DLDA-bc and FS-DQDA, by using the training samples having \( n_1 = 27 \) and \( n_2 = 11 \), and checked the accuracy by using the test samples from each \( \pi_i \). Throughout this section, we set \( \gamma = 0.5 \) in (5.1) for FS-DQDA. We compared the classifiers with the hard-margin linear support vector machine (HM-LSVM) given by Vapnic [22]. Note that the data sets are linearly separable by a hyperplane because \( p > n_1 + n_2 \). We emphasize that the computational cost of DBDA, GQDA, DQDA-bc, DLDA-bc or FS-DQDA is as low as HM-LSVM even when \( p \geq 10,000 \). We summarized misclassification rates in the first block of Table 1. We note that \( n_{\min}^{-1} \log p = 0.81 \), so that “\( n_{\min}^{-1} \log p = o(1) \)” does not hold. That is probably the reason why DQDA-bc, DLDA-bc and FS-DQDA seem to lose the consistency property. See Sections 4 and 5 for the details. On the other hand, DBDA and GQDA gave reasonable performances even when \( n_i \)’s are small and seem to hold the consistency property. We calculated \( \text{tr}(S_{1n_1})/\text{tr}(S_{2n_2}) = 0.989 \) and \( (\hat{\Delta}_{ij(I)}/\text{tr}(S_{jn_j})/p)/\hat{\Delta}_{(I)} \approx 1 \) for \( i \neq j \). The difference of the trace of the covariance matrices is small and this is probably the reason why DBDA gave a preferable performance. See Section 4.2 for the details. In addition, HM-LSVM also gave a preferable performance. See Hall et al. [15] for the consistency property of HM-LSVM. For this data set, Cai and Liu [9] summarized misclassification rates for several other classifiers including a sparse linear classifier called LPD. See Table 6 in [9] for the performances of the other classifiers. Note that LPD has the Bayes error rates asymptotically
Table 1: Error rates of the classifiers for samples from [14].

| Classifier | DBDA | GQDA | DLDA-bc | DQDA-bc | FS-DQDA | HM-LSVM |
|------------|------|------|---------|---------|---------|---------|
| **Test samples (ALL: 20 and AML: 14)** | | | | | | |
| Error rate | 1/34 | 1/34 | 5/34 | 2/34 | 3/34 | 1/34 |
| **LOOCV of samples (ALL: 47 and AML: 25)** | | | | | | |
| Error rate | 3/72 | 6/72 | 11/72 | 1/72 | 0/72 | 2/72 |

under several sparsity conditions. However, we observed that DBDA and GQDA gave the same performance as LPD. This is probably because the Bayes error rates are close to 0 or the sparsity conditions do not hold for this data set.

Next, by using all the samples (i.e., 72 samples), we checked the accuracy of the classifiers by using the Leave-One-Out Cross-Validation (LOOCV). We summarized misclassification rates in the second block of Table 1. We note that $n_{\text{min}} = 24$ and $n - 1\log p = 0.37$ or $n_{\text{min}} = 25$ and $n - 1\log p = 0.35$ in this case, so that $n_{\text{min}}^{-1}\log p$ is a little small. We observed that DQDA-bc and FS-DQDA give preferable performances. On the other hand, DLDA-bc gave a poor performance because it does not draw information about heteroscedasticity. For other classifiers, Tan et al. [21] summarized results of the LOOCV for this data set.

Finally, we analyzed gene expression data given by Armstrong et al. [3] in which the data set consists of 12582 (= $p$) genes and 72 samples. We had 3 classes of leukemia subtypes: acute lymphoblastic leukemia (ALL: 24 samples), mixed-lineage leukemia (MLL: 20 samples), and acute myeloid leukemia (AML: 28 samples). We considered three cases: (a) ALL and MLL, (b) ALL and AML, and (c) MLL and AML. We standardized each sample by $x_{ij}/\{\sum_{i=1}^{3}\text{tr}(S_{in})/(3p)\}^{1/2}$ for all $i, j$, as before. Then, we calculated $(\hat{\Delta}_I, \hat{\Delta}_\Sigma)$ as $(4076, 1.12 \times 10^8)$ for (a), $(15050, 5.49 \times 10^6)$ for (b), and $(8546, 1.16 \times 10^8)$ for (c). From this observation, we concluded that $\mu_{12}$ and $\Sigma_{12}$ are non-sparse for (a) to (c). Also, by using $\hat{\lambda}_{11}$, we estimated the largest eigenvalues as 1896, 3206 and 2101 for ALL, MLL and AML, respectively.

From this observation, we concluded that $\Sigma_{i*}$ are non-sparse. We considered estimators of $\text{tr}(\Sigma_{\text{max}}^2)/(n_{\text{min}}\Delta(I))$ and $\lambda_{\text{max}}/\hat{\Delta}(I)$ in (C-i’) and (C-ii’) by $C_i = \max\{W_{1n1}, W_{2n2}\}/(n_{\text{min}}\Delta_{I(i)}^2)$ and $C_{ii} = \max\{\hat{\lambda}_{11}, \hat{\lambda}_{21}\}/\hat{\Delta}(I)$, respectively. Then, we calculated $(C_i, C_{ii})$ as (0.362, 0.787) for (a), (0.001, 0.14) for (b), and (0.082, 0.375) for (c). Note that $\liminf_{p \to \infty} \Delta_{\text{min}(Ii)}/\Delta(I) > 0$ and $\liminf_{p \to \infty} \Delta_{\text{min}(III)}/\Delta(I) > 0$. From these observations, it is likely that the classifiers by (I) to (III) satisfy (C-i’) and (C-ii’) especially for (b) and...
Table 2

| Classifier | DBDA | GQDA | DLDA-bc | DQDA-bc | FS-DQDA | HM-LSVM |
|------------|------|------|---------|---------|---------|---------|
| Error rate | LOOCV of samples from (a) ALL: 24 and MLL: 20 | 1/44 | 2/44 | 6/44 | 1/44 | 0/44 | 0/44 |
| Error rate | LOOCV of samples from (b) ALL: 24 and AML: 28 | 1/52 | 1/52 | 1/52 | 0/52 | 0/52 | 0/52 |
| Error rate | LOOCV of samples from (c) MLL: 20 and AML: 28 | 4/48 | 4/48 | 1/48 | 3/48 | 3/48 | 3/48 |
| Error rate | LOOCV of samples from ALL: 24, MLL: 20 and AML: 28 | 5/72 | 6/72 | 7/72 | 4/72 | 2/72 | 3/72 |

hold the consistency property in (2.2) from Proposition 2.1.

Based on all the samples, we checked the accuracy of the classifiers by using the LOOCV for (a) to (c). We checked the accuracy for 3-class classification as well by using the multiclass classification rule given in Remark 1. In the 3-class classification, we used $\hat{\theta}_r$ given in Remark 5 for FS-DQDA and the one-versus-one approach for HM-LSVM. We summarized misclassification rates in Table 2. We observed that FS-DQDA gives excellent performances. HM-LSVM also gave reasonable performances, however, it does not draw information about the difference of the covariance matrices. See Section 2.2 in Aoshima and Yata [2] for such an example. As for (b), all the classifiers gave preferable performances. This is probably because the classifiers by (I) to (III) satisfy (C-i'}) and (C-ii'}) for (b).

7. Concluding remarks. In this paper, we considered high-dimensional quadratic classifiers in non-sparse settings. The classifier based on the Mahalanobis distance does not always give a preferable performance even when $n_{\min} \to \infty$ and $\pi_i$s are assumed Gaussian, having known covariance matrices. See Sections 1 and 3. We emphasize that the quadratic classifiers proposed in this paper draw information effectively about heteroscedasticity through the difference of parameters related to the expanding covariance matrices. See Section 3.4 for the details. If the difference is not sufficiently large, we recommend to use the linear classifiers, DBDA and DLDA-bc (or the original DLDA). They are quite flexible about the conditions to claim the consistency property. See Sections 4.2 and 4.3 for the details. We emphasize that DLDA-bc, DQDA-bc and FS-DQDA can hold the consistency property under at least $n_{\min}^{-1}\log p = o(1)$. Thus we do not recommend to use the classifiers when $n_{\min}^{-1}\log p \neq o(1)$. In such cases, one should use DBDA and GQDA because they hold the consistency property even when
$n_i$s are fixed. See Section 4.2 about the choice between DBDA and GQDA. When $n_{\min}^{-1} \log p = o(1)$, we recommend DQDA-bc and FS-DQDA. Especially, FS-DQDA can claim the consistency property even when $n_{\min}/p \to 0$ and $\Delta_{\min}$ is not sufficiently large. See Section 5.1 for the details. For a choice of $\gamma \in (0,1)$ in (5.1), one may apply cross-validation procedures or simply choose as $\gamma = 0.5$. Actually, FS-DQDA with $\gamma = 0.5$ gave preferable performances throughout our simulations and real data analyses. On the other hand, even when $n_{\min}^{-1} \log p = o(1)$, we do not recommend to use classifiers by the sparse estimation of $\Sigma_i^{-1}$ unless (1) the eigenvalues are bounded in the sense that $\lambda(\Sigma_i) \in (0,\infty)$ as $p \to \infty$, and (2) $\Sigma_i$s are sparse in the sense that many elements of $\Sigma_i$s are very small. We emphasize that “$\lambda_{i1}$s are bounded” is a strict condition since the eigenvalues should depend on $p$ and it is probable that $\lambda_{ij} \to \infty$ as $p \to \infty$ for the first several $j$s. See Yata and Aoshima [23] for the details. Also, the computational cost of the classifiers by the sparse estimation is terribly high.

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

APPENDIX A

We give proofs of the theorems. For proofs of the corollaries and the propositions, see Appendix B.

Proof of Theorem 2.1. We consider the case when $x_0 \in \pi_1$. Under (C-i) and (C-ii), it holds that for $i = 1,2$

$$\text{Var}\{(x_0 - \mu_1)^T A_i^{-1}(\pi_{i,n} - \mu_i)^T\} = \text{tr}(\Sigma_i A_i^{-1} \Sigma_1 A_i^{-1})/n_i = o(\Delta_i^2)$$
(A.1) and

$$\text{Var}\{(x_0 - \mu_1 - \bar{x}_{2n_2} + \mu_2)^T A_2^{-1} \mu_1\} = \mu_{12} T A_2^{-1}(\Sigma_1 + \Sigma_2/n_2) A_2^{-1} \mu_1 = o(\Delta_i^2)$$

from the fact that $\mu_{12} T A_2^{-1} \Sigma_2 A_2^{-1} \mu_1 \leq \mu_{12} T A_2^{-1} \mu_1 \leq \lambda_{\max}(A_2^{-1/2} \Sigma_2 A_2^{-1/2}) \leq \Delta_1 \text{tr}((\Sigma_2 A_2^{-1})^2)^{1/2} = o(n_2 \Delta_i^2)$ under (C-i). Note that $(\pi_{i,n} - \mu_i)^T A_i^{-1} (\pi_{i,n} - \mu_i) - \text{tr}(A_i^{-1} S_{i,n})/n_i = \sum_{i \neq i'} (\pi_{i,n} - \mu_i)^T A_i^{-1} (\pi_{i,n} - \mu_i)/\{n_i(n_i-1)\}$. Then, under (C-i), it follows that for $i = 1,2$

$$\text{Var}\{(\pi_{i,n} - \mu_i)^T A_i^{-1} (\pi_{i,n} - \mu_i) - \text{tr}(A_i^{-1} S_{i,n})/n_i\} = 2\text{tr}((\Sigma_i A_i^{-1})^2)/\{n_i(n_i-1)\} = o(\Delta_i^2).$$
(A.2)

Then, by using Chebyshev’s inequality, from (A.1) and (A.2), we find that

$$W_2(A_2) - W_1(A_1)$$
(A.3) \[ = \text{tr}[(x_0 - \mu_1)(x_0 - \mu_1)^T - \Sigma_1][(A_2^{-1} - A_1^{-1})] + \Delta_1 + o_P(\Delta_1), \]
Here, under (A-i) and (C-iii), it follows that
\[
\text{Var}(\text{tr}[(x_0 - \mu_i)(x_0 - \mu_i)^T - \Sigma_i](A_i^{-1} - A_i^{-1}))
\]
\[(A.4)\]
\[= O(\text{tr}[(\Sigma_i(A_i^{-1} - A_i^{-1}))^2]) = o(\Delta^2_i).
\]
Thus by combining (A.3) with (A.4), under (A-i) and (C-i) to (C-iii), we obtain that \(\{W_2(A_2) - W_1(A_1)/\Delta_1 = 1 + o_P(1)\), so that \(P\{W_2(A_2) - W_1(A_1) > 0\} \rightarrow 1\). When \(x_0 \in \pi_2\), we have the same arguments. The proof is completed.

**Proof of Theorem 3.1.** Note that \(\text{tr}\{(\Sigma_iA_i^{-1})^2)/n_i^2 = o(\delta^2_i), i = 1, 2\). Then, similar to (A.1) to (A.4), under (A-i) and (C-iv) to (C-vi), we have that as \(m \rightarrow \infty\)
\[
W_j(A_j) - W_i(A_i) - \Delta_i = 2(x_0 - \mu_i)^T(A_i^{-1}(\pi_{in_i} - \mu_i) - A_j^{-1}(\pi_{jn_j} - \mu_j) + o_P(\delta_i)
\]
(A.5)
when \(x_0 \in \pi_i\ (j \neq i)\). Note that \(2\omega_i/\delta_i \rightarrow 1\) as \(m \rightarrow \infty\) for \(i = 1, 2\), under (C-vi), where \(\omega_i = \{\text{tr}\{(\Sigma_iA_i^{-1})^2)/n_i + \text{tr}(\Sigma_iA_i^{-1}\Sigma_jA_j^{-1})/n_j\}^{1/2} (j \neq i)\) in view of Lemma B.1 of Appendix B. Then, by combining Lemma B.1 with (A.5), we conclude the results.

**Proof of Theorem 3.2.** Similar to (A.5), under (A-i), (C-iv), (C-v) and (C-vii), we have the following as \(m \rightarrow \infty\):
\[
W_j(A_j) - W_i(A_i) - \Delta_i = 2(x_0 - \mu_i)^T(A_i^{-1}(\pi_{in_i} - \mu_i) - A_j^{-1}(\pi_{jn_j} - \mu_j + (-1)^{j_2}(-1)^{j_2})) + o_P(\delta_i)
\]
(A.6)
when \(x_0 \in \pi_i\ (j \neq i)\). Then, by combining Lemma B.2 of Appendix B with (A.6), we conclude the results.

**Proof of Theorem 5.1.** By using (B.17) and (B.18) in Appendix B, we claim the result. □

**APPENDIX B**

Throughout, we consider the eigen-decomposition of \(A_i\ (i = 1, 2)\) such as \(A_i = H_i(A)\Lambda_i(A)H_i^T(A)\), where \(\Lambda_i(A)\) is a diagonal matrix of eigenvalues, \(\lambda_{i1}(A) \geq \cdots \geq \lambda_{ip(A)} > 0\), and \(H_i(A) = [h_{i1}(A), \ldots, h_{ip(A)}]\) is an orthogonal matrix of the corresponding eigenvectors. Let \(a_{i(j)}\) be the \(j\)-th diagonal element of \(A_i\) for \(j = 1, \ldots, p\ (i = 1, 2)\). Let \(\tilde{x}_{1j} = A_i^{-1/2}(x_{1j} - \mu_i)\) and \(\tilde{x}_{2j} = A_i^{-1/2}A_i^{-1}(x_{2j} - \mu_2)\) for \(j = 1, \ldots, n_i\). Let \(\Sigma_1 = A_i^{-1/2}\Sigma_1A_i^{-1/2}\),
\[ \Sigma_2 = \mathcal{A}_1^{1/2} \mathcal{A}_2^{-1} \mathcal{A}_2 \mathcal{A}_1^{1/2}, \quad \tilde{\Gamma}_1 = [\tilde{\gamma}_{11}, \ldots, \tilde{\gamma}_{1q_1}] = \mathcal{A}_1^{-1/2} \Gamma_1 \text{ and } \tilde{\Gamma}_2 = [\tilde{\gamma}_{21}, \ldots, \tilde{\gamma}_{2q_2}] = \mathcal{A}_1^{1/2} \mathcal{A}_2^{-1} \Gamma_2. \] Note that Var(\(x_{ij}\)) = \(\tilde{\Gamma}_i \tilde{\Gamma}_i^T = \sum_{s=1}^{q_i} \tilde{\gamma}_{is} \tilde{\gamma}_{is}^T = \Sigma_i, \ i = 1, 2. \] Let \(B_i = \mathcal{A}_i^{-1} - \mathcal{A}_1^{-1}\) for \(i = 1, 2.\)

**Proof of Proposition 1.1.** We can write that tr\((\mathcal{A}_i \mathcal{A}_j^{-1}) = \sum_{s=1}^{p} h_{js}(A_i) h_{js}(A_j)/\lambda_{js}(A_i).\) Note that \(\sum_{s=1}^{p} h_{js}(A_i) h_{js}(A_j) \geq \text{tr}(\mathcal{A}_j)\) and \(\sum_{s=1}^{p} (\mathcal{A}_i h_{js}(A_j) - \lambda_{is}(A_i)) \leq 0\) for any \(i \in \{1, \ldots, p\}.\) Then, by noting that \(\lambda_{j1}(A) \geq \cdots \geq \lambda_{jp}(A) > 0,\) we have that

\[
\text{tr}(\mathcal{A}_i \mathcal{A}_j^{-1}) = \frac{\lambda_{i1}(A)}{\lambda_{j1}(A)} + \frac{h_{j1}(A) - \lambda_{i1}(A)}{\lambda_{j1}(A)} + \sum_{s=2}^{p} \frac{h_{js}(A) h_{js}(A) - \lambda_{is}(A)}{\lambda_{js}(A)}
\]

\[
\geq \sum_{s=1}^{2} \frac{\lambda_{is}(A)}{\lambda_{js}(A)} + \sum_{s=1}^{p} \frac{h_{js}(A) h_{js}(A) - \lambda_{is}(A)}{\lambda_{js}(A)} = \sum_{s=1}^{p} \frac{\lambda_{is}(A)}{\lambda_{js}(A)}.
\]

Thus, when \(\text{tr}\{\Sigma_i(\mathcal{A}_j^{-1} - \mathcal{A}_j^{-1})\} = \text{tr}(\mathcal{A}_i \mathcal{A}_j^{-1}) - p,\) it holds that

\[
\Delta_i \geq \sum_{s=1}^{p} \frac{\lambda_{is}(A)}{\lambda_{js}(A)} - 1 + \log(\lambda_{js}(A)/\lambda_{is}(A)) \geq 0
\]

from the fact that \(c - 1 + \log c^{-1} \geq 0\) for any positive constant \(c.\) Note that \(\lambda_{is}(A) \neq \lambda_{js}(A)\) or \(h_{js}(A) h_{js}(A) < \lambda_{is}(A)\) for some \(s\) when \(\mathcal{A}_1 \neq \mathcal{A}_2.\) Since \(c - 1 + \log c^{-1} > 0\) when \(c \neq 1,\) it holds that \(\Delta_i > 0\) when \(\lambda_{is}(A) \neq \lambda_{js}(A)\) for some \(s.\) From (B.1), if \(h_{js}(A) h_{js}(A) < \lambda_{is}(A)\) for some \(s,\) it follows that \(\text{tr}(\mathcal{A}_i \mathcal{A}_j^{-1}) > \sum_{s=1}^{p} \{\lambda_{is}(A)/\lambda_{js}(A)\},\) so that \(\Delta_i > 0.\) When \(\mu_1 \neq \mu_2,\) it holds that \(\Delta_i \geq \mu_{12}^{-1} A^{-1} - \mu_{12} > 0.\) Hence, it concludes the results. \(\square\)

**Proof of Proposition 2.1.** We note that

\[
\Delta_{i} \leq \mu_{12}^{-1} A_{j}^{-1} \lambda_{i} \Sigma_1^{1/2} A_{j}^{1/2} \Sigma_1 A_{j}^{-1/2} \leq \Delta_{i} \lambda_{i} \Sigma_1^{1/2} A_{j}^{1/2} \Sigma_1 A_{j}^{-1/2} \leq \Delta_{i} \lambda_{j} \Sigma_1 A_{j}^{-1} \Sigma_1 A_{j}^{-1/2} \leq \text{tr}\{(\Sigma_1 A_{j}^{-1})^2\}^{1/2} \text{tr}\{(\Sigma_1 A_{j}^{-1})^2\}^{1/2} \text{tr}(\Sigma_1 A_{j}^{-1}) \text{tr}(\Sigma_1 A_{j}^{-1})^{1/2}.
\]

When \(\liminf_{p \to \infty} \lambda_{i} > 0, \ i = 1, 2,\) it holds that

\[
\lambda_{i} = \min_{\lambda_{i}}(A_{j}^{-1} \Sigma_i A_{j}^{-1/2}) \leq \lambda_{i} \lambda_{i} \lambda_{i} = \lambda_{i} \lambda_{i} = (\lambda_{i}) \lambda_{i} = O(\lambda_{i}) \text{ and } \lambda_{i} \lambda_{i} \lambda_{i} = O(\lambda_{i}).
\]

\[
\text{tr}\{(\Sigma_1 A_{j}^{-1})^2\} \leq \text{tr}(\Sigma_1 A_{j}^{-1}) \lambda_{i} \lambda_{i} \lambda_{i} = \text{tr}(\Sigma_1 A_{j}^{-1}) \lambda_{i} \lambda_{i} = O(\text{tr}(\Sigma_1 A_{j}^{-1}))
\]

(B.3)
for all $i', j'$. By combining (B.2) with (B.3), when $\liminf_{p \to \infty} \lambda_{ip}(A) > 0$, $i = 1, 2$, (C-i') and (C-ii') imply (C-i) and (C-ii).

Next, for (C-iii), it holds that $\text{tr}\{[\sigma_i(A_1^{-1} - A_2^{-1})]2\} \leq \lambda_{11} \text{tr}\{|(A_1^{-1} - A_2^{-1})\Sigma_i(A_1^{-1} - A_2^{-1})\}$. When $A_i$s are diagonal matrices such as $A_i = \text{diag}(a_{i(1)}, \ldots, a_{i(p)})$, $i = 1, 2$, it holds that $\Delta_i \geq \sum_{s=1}^p \{a_{i(s)}/a_{j(s)} - 1 - \log(a_{i(s)}/a_{j(s)}) \}$ and $\text{tr}\{[(A_1^{-1} - A_2^{-1})\Sigma_i(A_1^{-1} - A_2^{-1})]\} = \sum_{s=1}^p \sigma_i(s)(a_{1(s)} - a_{2(s)})^2/(a_{1(s)}a_{2(s)})^2$. Note that $a_{i(s)} \in (0, \infty)$ as $p \to \infty$ for all $i, s$, under $\lambda(A_i) \in (0, \infty)$ as $p \to \infty$ for $i = 1, 2$. By Taylor expansion, we claim that $a_{i(s)}/a_{j(s)} - 1 - \log(a_{i(s)}/a_{j(s)}) \geq 2 a_{1(s)}/a_{2(s)}^2/(2 \max\{1, a_{i(s)/a_{j(s)}^2}\}$.

Then, it follows that $\sum_{s=1}^p \sigma_i(s)(a_{1(s)} - a_{2(s)})^2/(a_{1(s)}a_{2(s)})^2 = O(\Delta_i)$ because $\sigma_i(s) \in (0, \infty)$ as $p \to \infty$ for all $i, s$. Thus we have that $\text{tr}\{[\sigma_i(A_1^{-1} - A_2^{-1})^2\} = O(\Delta_i \lambda_{11})$. It concludes the results.

**Proofs of Corollaries 2.1 and 2.2.** From Theorem 2.1 and Proposition 2.1, we can claim Corollaries 2.1 and 2.2 straightforwardly.

**Proof of Proposition 2.2.** We first consider the case when $\liminf_{p \to \infty} \sum_{s=1}^p |\lambda_{is}/\lambda_{js} - 1|/p > 0$. When $c_{1s} < |\lambda_{is}/\lambda_{js} - 1| < c_{2s}$ for some constants $c_{1s} (> 0)$ and $c_{2s} (< \infty)$, by Taylor expansion, it holds that

$$\lambda_{is}/\lambda_{js} - 1 - \log(\lambda_{is}/\lambda_{js}) \geq \frac{(\lambda_{is}/\lambda_{js} - 1)^2}{2 \max\{1, \lambda_{is}/\lambda_{js}^2\}} = \frac{c_{1s}|\lambda_{is}/\lambda_{js} - 1|}{2(c_{2s} + 1)^2}.$$ 

When $\lambda_{is}/\lambda_{js} \to \infty$ as $p \to \infty$, it holds that for sufficiently large $p$

$$\lambda_{is}/\lambda_{js} - 1 - \log(\lambda_{is}/\lambda_{js}) > |\lambda_{is}/\lambda_{js} - 1|/2.$$ 

Thus, when $\liminf_{p \to \infty} \sum_{s=1}^p |\lambda_{is}/\lambda_{js} - 1|/p > 0$, it follows that $\liminf_{p \to \infty} \Delta_i(I\!V)/p \geq \liminf_{p \to \infty} \sum_{s=1}^p |\lambda_{is}/\lambda_{js} - 1 - \log(\lambda_{is}/\lambda_{js})|/p > 0$ from (B.1).

Next, we consider the case when $\liminf_{p \to \infty} |\text{tr}(\Sigma_i \Sigma_j^{-1})|/p - 1 \geq 0$. Note that $\text{tr}(\Sigma_i \Sigma_j^{-1}) \geq \sum_{s=1}^p \lambda_{is}/\lambda_{js}$ from (B.1). When $\text{tr}(\Sigma_i \Sigma_j^{-1})/(\sum_{s=1}^p \lambda_{is}/\lambda_{js}) \to 1$ as $p \to \infty$, it holds that $\liminf_{p \to \infty} |\sum_{s=1}^p (\lambda_{is}/\lambda_{js})/p - 1| > 0$ under $\liminf_{p \to \infty} \Delta_i(I\!V)/p > 0$. It follows that $\liminf_{p \to \infty} \Delta_i(I\!V)/p > 0$ from the fact that $\sum_{s=1}^p |\lambda_{is}/\lambda_{js} - 1|/p \geq |\sum_{s=1}^p (\lambda_{is}/\lambda_{js})/p - 1|$. On the other hand, we note that

$$\Delta_i(I\!V) \geq \text{tr}(\Sigma_i \Sigma_j^{-1}) - p - \sum_{s=1}^p \log(\lambda_{is}/\lambda_{js}) \geq \text{tr}(\Sigma_i \Sigma_j^{-1}) - \sum_{s=1}^p (\lambda_{is}/\lambda_{js})$$

because $\sum_{s=1}^p (\lambda_{is}/\lambda_{js} - 1 - \log(\lambda_{is}/\lambda_{js})) \geq 0$. Thus, when $\sum_{s=1}^p (\lambda_{is}/\lambda_{js})/p - 1 \to 0$ as $p \to \infty$ and $\liminf_{p \to \infty} \text{tr}(\Sigma_i \Sigma_j^{-1})/(\sum_{s=1}^p \lambda_{is}/\lambda_{js}) > 1$, we have that $\liminf_{p \to \infty} \Delta_i(I\!V)/p > 0$. Hence, it concludes the results.
Proof of Proposition 3.1. Under \( \liminf_{p \to \infty} \lambda_{ip(A)} > 0 \) for \( i = 1, 2 \), we have that \( \text{tr}\{(\Sigma_i A_j^{-1})^2\} = O(\text{tr}(\Sigma_i^2)) \) and

\[
\text{tr}\{(\Sigma_i A_i^{-1} \Sigma_i A_i^{-1})^2\} = \text{tr}\{(\Sigma_i^{1/2} A_i^{-1} \Sigma_i A_i^{-1} \Sigma_i^{1/2})^2\}
\]

\[
\leq \lambda_{\text{max}}(\Sigma_i^{1/2} A_i^{-1} \Sigma_i A_i^{-1} \Sigma_i^{1/2})\text{tr}(\Sigma_i^{1/2} A_i^{-1} \Sigma_i A_i^{-1} \Sigma_i^{1/2})
\]

\[
\leq \lambda_{\text{max}}(\Sigma_i^{1/2} A_i^{-2} \Sigma_i^{1/2})\lambda_{i1} \xi_i^2 n_t = O(\lambda_{i1} \xi_i^2 n_t); \text{ and}
\]

\[
\mu_{12}^T A_j^{-1} \Sigma_i A_j^{-1} \mu_{12} \leq ||\mu_{12}||^2 \lambda_{\text{max}}(A_j^{-1} \Sigma_i A_j^{-1}) = O(||\mu_{12}||^2 \lambda_{i1}) \text{ for } l = i, j.
\]

Then, when \( \liminf_{p \to \infty} \lambda_{ip(A)} > 0 \), \( i = 1, 2 \), (C-iv') and (C-vi') imply (C-iv) and (C-vi), respectively. Similar to Proof of Proposition 2.1, we can claim the result for (C-v') from \( \text{tr}\{A_1 - A_2\}^2\} = \sum_{s=1}^{p}(a_1(s) - a_2(s))^2 \) when \( A_i \)'s are diagonal matrices. The proof is completed.

Lemma B.1. Let \( \omega_i = \{\text{tr}\{(\Sigma_i A_i^{-1})^2\}/n_i + \text{tr}(\Sigma_i A_i^{-1} \Sigma_i A_j^{-1})/n_j\}^{1/2} \) for \( i = 1, 2 \) \((j \neq i)\). Then, under (A-i), (C-iv) and (C-vi), we have that

\[
(x_0 - \mu_i)^T \{A_i^{-1}(\bar{x}_{in_i} - \mu_i) - A_j^{-1}(\bar{x}_{jn_j} - \mu_j)\}/\omega_i \Rightarrow N(0,1) \quad \text{as } m \to \infty
\]

when \( x_0 \in \pi_i \) for \( i = 1, 2 \) \((j \neq i)\).

Proof of Lemma B.1. We consider the case when \( i = 1 \) \((j = 2)\) and \( x_0 \in \pi_1 \). Let \( \bar{x}_0 = A_i^{-1/2}(x_0 - \mu_i) \). Then, it holds that \( \text{Var}(\bar{x}_0|x_0 \in \pi_1) = \text{Var}(\bar{x}_{i1}) = \Sigma_1 \). Let

\[
v_l = \bar{x}_0^T \bar{x}_{1l}/(n_1 \omega_1), \quad l = 1, ..., n_1, \text{ and } v_{n_1+l} = -\bar{x}_0^T \bar{x}_{2l}/(n_2 \omega_1), \quad l = 1, ..., n_2.
\]

Note that \( \sum_{l=1}^{n_1+n_2} E(v_l^2) = 1 \) and

\[
\sum_{l=1}^{n_1+n_2} v_l = (x_0 - \mu_1)^T \{A_1^{-1}(\bar{x}_{1n_1} - \mu_1) - A_2^{-1}(\bar{x}_{2n_2} - \mu_2)\}/\omega_1.
\]

Then, it holds that \( E(v_l|v_{l-1},...,v_1) = 0 \) for \( l = 2, ..., n_1 + n_2 \). We consider applying the martingale central limit theorem given by McLeish [19]. Under (A-i), we can write that \( \bar{x}_{1l} = \Gamma_1 y_{1l} \) and \( \bar{x}_{2l} = \Gamma_2 y_{2l} \). Then, in a way similar to the equations (23) and (24) in Aoshima and Yata [2], we can evaluate that under (A-i)

\[
(n_s \omega_1)^4 E(v_l^4) = 3\text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_s)^2 + O[\text{tr}\{(\bar{\Sigma}_1 \bar{\Sigma}_s)^2\}] \quad \text{and}
\]

\[
(n_s n_s')^2 \omega_1^2 E(v_l^2 v_l^2) = \text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_s)\text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_{s'}) + 2\text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_s \bar{\Sigma}_1 \bar{\Sigma}_{s'})
\]

\[
+ O[\text{tr}\{(\bar{\Sigma}_1 \bar{\Sigma}_s \bar{\Sigma}_1 \bar{\Sigma}_{s'})\}]^{1/2}
\]

\[
(n_s n_s')^2 \omega_1^2 E(v_l^2 v_l^2) = \text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_s)\text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_{s'}) + 2\text{tr}(\bar{\Sigma}_1 \bar{\Sigma}_s \bar{\Sigma}_1 \bar{\Sigma}_{s'})
\]

\[
+ O[\text{tr}\{(\bar{\Sigma}_1 \bar{\Sigma}_s \bar{\Sigma}_1 \bar{\Sigma}_{s'})\}]^{1/2}
\]
for $l \neq l'$, where $s = 1$ for $l \in [1, ..., n_1]$, $s = 2$ for $l \in [n_1 + 1, ..., n_1 + n_2]$, $s' = 1$ for $l' \in [1, ..., n_1]$, and $s' = 2$ for $l' \in [n_1 + 1, ..., n_1 + n_2]$. Note that $\text{tr}(\tilde{\Sigma}_1^4) \leq \text{tr}(\tilde{\Sigma}_1^2)^2$ and $\text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\} \leq \text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2$. Then, by using Chebyshev’s inequality and Schwarz’s inequality, from (B.4), under (A-i), we have that for Lindeberg’s condition
\[
\sum_{l=1}^{n_1 + n_2} E\{v_l^2 I(v_l^2 \geq \tau)\} \leq \sum_{l=1}^{n_1 + n_2} E(v_l^4) = O\left[\frac{\text{tr}(\tilde{\Sigma}_1^2)^2}{n_1^3} + \frac{\text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2}{n_1^3} \right] \tau \rightarrow 0
\]
as $m \rightarrow \infty$ for any $\tau > 0$, where $I(\cdot)$ is the indicator function. Since $2\omega_1/\delta_1 = 1 + o(1)$ under (C-vi), we note that
\[
\frac{\text{tr}(\tilde{\Sigma}_1^4)}{n_1^2 \omega_1^4} \rightarrow 0, \quad \frac{\text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}}{n_2^2 \omega_1^4} \rightarrow 0, \\
\text{and} \quad \frac{\text{tr}(\tilde{\Sigma}_1^4)}{n_1 n_2 \omega_1^4} \leq \frac{\text{tr}(\tilde{\Sigma}_1^4)^{1/2} \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}^{1/2}}{n_1 n_2 \omega_1^4} \rightarrow 0
\]
under (C-iv). Then, by using Chebyshev’s inequality, from (B.4) and (B.5), under (A-i), (C-iv) and (C-vi), we have that for any $\tau > 0$
\[
P\left(\left|\sum_{l=1}^{n_1 + n_2} v_l^2 - 1\right| \geq \tau\right) \leq O\left[\frac{\text{tr}(\tilde{\Sigma}_1^4)/n_1^2 + (\tilde{\Sigma}_1^4)^{1/2} \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}^{1/2}/(n_1 n_2) + (\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2/n_2^3}{\omega_1^4}\right] \rightarrow 0
\]
as $m \rightarrow \infty$, so that $\sum_{l=1}^{n_1 + n_2} v_l^2 = 1 + o_P(1)$. Hence, by using the martingale central limit theorem, we obtain that $\sum_{l=1}^{n_1 + n_2} v_l \Rightarrow N(0, 1)$ as $m \rightarrow \infty$ under (A-i), (C-iv) and (C-vi). Hence, we conclude the result when $i = 1$. For the case when $i = 2$, we can have the same arguments. The proof is completed. \(\square\)

**Lemma B.2.** Under (A-ii), (C-iv) and (C-vii), we have that
\[
2(x_0 - \mu_i)^T \{A_i^{-1}(\pi_{i m} - \mu_i) - A_j^{-1}(\pi_{j n} - \mu_j) + (-1)^i \mu_{12}\}/\delta_i \Rightarrow N(0, 1)
\]
as $m \rightarrow \infty$ when $x_0 \in \pi_i$ for $i = 1, 2$ ($j \neq i$).
Proof of Lemma B.2. We consider the case when \( i = 1 \) \( (j = 2) \) and \( x_0 \in \pi_1 \). Let \( x_0 - \mu_1 = \Gamma_1 y_0 \) and \( y_0 = (y_{01}, \ldots, y_{0q_1})^T \). Under (A-ii), \( y_{0s}, \ s = 1, \ldots, q_1 \), are independent. Let \( \overline{x}_i n' = \sum_{j' = 1}^{n'} \overline{x}_{i'j'}/n' \), \( i', j' = 1, 2 \),
\[
\hat{\mu} = A_1^{1/2} A_2^{-1} \mu_{12} \text{ and }
\]
\[
w_s = 2y_{0s} \gamma_{1s}^T \{\overline{x}_{1n_1} - \overline{x}_{2n_2} + \hat{\mu}\}/\delta_1, \ s = 1, \ldots, q_1.
\]
Note that \( q_1 \geq p \), \( E(w_s) = 0 \), \( s = 1, \ldots, q_1 \), \( \sum_{s=1}^{q_1} E(w_s^2) = 1 \) and
\[
\sum_{s=1}^{q_1} w_s = 2(x_0 - \mu_1)^T \{A_1^{-1}(\overline{x}_{1n_1} - \mu_1) - A_2^{-1}(\overline{x}_{2n_2} - \mu_2 - \mu_{12})\}/\delta_1.
\]
Also, note that \( E(w_s|w_{s-1}, \ldots, w_1) = 0 \) for \( s = 2, \ldots, q_1 \), under (A-ii). We consider applying the martingale central limit theorem. Let \( M_{i's} = E(y_{i's}^2) \) for all \( i', j' \). Note that \( \lim sup_{p \to \infty} |M_{i's}| < \infty \) for all \( i', s \), under (A-ii) because \( \lim sup_{p \to \infty} E(y_{i's}^4) < \infty \). Then, by using Schwarz’s inequality and the arithmetic mean-geometric mean inequality, we can evaluate that under (A-ii)
\[
E\{(\gamma_{1s}^T \overline{x}_{1n_1})^2(\hat{\gamma}_{1s}^T \overline{x}_{1n_1})^2\} = [1 + o(1)] \gamma_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1l}/n_l^2
\]
\[+ O(\{\gamma_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1l}/n_l\}^2); \text{ and}
\]
\[
|E\{(\gamma_{1s}^T \overline{x}_{1n_1})^2(\hat{\gamma}_{1s}^T \overline{x}_{1n_1} \hat{\gamma}_{1l}^T \hat{\mu})| = \sum_{u=1}^{q_1} (\gamma_{1s}^T \hat{\gamma}_{1lu}/n_l^2)^2 \gamma_{1s}^T \hat{\gamma}_{1lu} \hat{\gamma}_{1l}^T \hat{\mu} M_{lu}/n_l^2
\]
\[
\leq E\{(\gamma_{1s}^T \overline{x}_{1n_1})^4\}^{1/2} E\{(\gamma_{1s}^T \overline{x}_{1n_1} \gamma_{1l}^T \hat{\mu})^2\}^{1/2}
\]
\[
= O\{\gamma_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1s} (\gamma_{1l}^T \hat{\Sigma}_l \hat{\gamma}_{1l}/n_l)^{1/2} \gamma_{1l}^T \hat{\mu}/n_l\}
\]
\[
= O\{\gamma_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1s} (\gamma_{1l}^T \hat{\Sigma}_l \hat{\gamma}_{1l}/n_l + (\gamma_{1l}^T \hat{\mu})^2)/n_l\}
\]
\[
= O\{\gamma_{1s}^T \hat{\Sigma}_l \hat{\gamma}_{1s}/n_l\}^2 + (\gamma_{1l}^T \hat{\Sigma}_l \hat{\gamma}_{1l}/n_l)^2 + (\gamma_{1l}^T \hat{\mu})^4\}, \ l = 1, 2
\]
for all \(s, t\). Then, we have that

\[
\delta_1^l E(w_s^4) = O\left[ \sum_{l=1}^{2} \{ \hat{\gamma}_{1s}^{\top} \hat{\Sigma}_{1s} \hat{\gamma}_{1s}/n_l \}^2 \right] \quad \text{and}
\]

\[
(\delta_1/2)^l \frac{E(w_s^2 w_t^2)}{E(y_{0s}^2 y_{0t}^2)} - \hat{\gamma}_{1s}^{\top} \left( \sum_{l=1}^{2} \hat{\Sigma}_{l} / n_l + \hat{\mu}^{\top} \hat{\Sigma}_{1s} \hat{\gamma}_{1s} \right) \hat{\gamma}_{1t} \hat{\gamma}_{1t}^{\top} \left( \sum_{l=1}^{2} \hat{\Sigma}_{l} / n_l + \hat{\mu}^{\top} \hat{\Sigma}_{1s} \hat{\gamma}_{1s} \right) \hat{\gamma}_{1t} / n_l^2
\]

\[
= 2 \sum_{l=1}^{2} \sum_{u=1}^{q_l} \left\{ \left( \frac{\hat{\gamma}_{1s} \hat{\gamma}_{lu}}{\hat{\gamma}_{1t}} \right)^2 \hat{\gamma}_{1u} \hat{\gamma}_{1u}^{\top} \hat{\gamma}_{1t} \right\} \hat{\mu}_{lu} / n_l^2 + O\left( \sum_{l=1}^{2} \{ \hat{\gamma}_{1s}^{\top} \hat{\Sigma}_{1s} \hat{\gamma}_{1s}/n_l \}^2 \right).
\]

(B.6)

Here, under (C-iv), we can evaluate that

\[
\sum_{s, l=1}^{q_l} \sum_{u=1}^{q_l} \left( \frac{\hat{\gamma}_{1s} \hat{\gamma}_{lu}}{\hat{\gamma}_{1t}} \right)^2 \left( \frac{\hat{\gamma}_{1u} \hat{\gamma}_{1u}^{\top} \hat{\gamma}_{1t}}{\hat{\gamma}_{1t} \hat{\gamma}_{1t}^{\top} \hat{\gamma}_{1t}} \right) \hat{\mu}_{lu} / n_l^2
\]

\[
= O\left[ ||\hat{\mu}^{\top} \hat{\Sigma}_{1s}^{1/2}| \left| \sum_{u=1}^{q_l} \left( \hat{\gamma}_{1u} \hat{\gamma}_{1u}^{\top} \hat{\gamma}_{1t} \right)^{1/2} \right| \right]
\]

\[
= O\left[ ||\hat{\mu}^{\top} \hat{\Sigma}_{1s}^{1/2}| \text{tr}(\hat{\Sigma}_{1s} \hat{\Sigma}_{1s}^{\top})^{1/2} \left| \sum_{u=1}^{q_l} \left( \hat{\gamma}_{1u} \hat{\gamma}_{1u}^{\top} \hat{\gamma}_{1t} \right)^{1/2} / n_l^2 \right| \right]
\]

(B.8)

\[
= O\left[ ||\hat{\mu}^{\top} \hat{\Sigma}_{1s} \hat{\mu} + \text{tr}(\hat{\Sigma}_{1s} \hat{\Sigma}_{1s}^{\top}) \text{tr}(\hat{\Sigma}_{1s} \hat{\Sigma}_{1s}^{\top})^{1/2} / n_l^2 \right] = o(\delta_1^l), \quad l = 1, 2
\]

from the fact that \(\sum_{u=1}^{q_l} \left( \frac{\hat{\gamma}_{1s} \hat{\gamma}_{lu}}{\hat{\gamma}_{1t}} \right)^2 \leq \sum_{u, w=1}^{q_l} \left( \frac{\hat{\gamma}_{1u} \hat{\gamma}_{1u}^{\top} \hat{\gamma}_{1t}}{\hat{\gamma}_{1t} \hat{\gamma}_{1t}^{\top} \hat{\gamma}_{1t}} \right)^2 = \text{tr}(\hat{\Sigma}_{1s} \hat{\Sigma}_{1s}^{\top}) = o(n_l^2 \delta_1^l) \) under (C-iv). Then, by combining (B.6) and (B.7) with (B.8), under (A-ii), (C-iv) and (C-vii), for any \(\tau > 0\), we have that as \(m \to \infty\)

\[
\sum_{s=1}^{q_l} E(w_s^2) = O\left[ \frac{\sum_{s, l=1}^{q_l} \text{tr}(\hat{\Sigma}_{1s} \hat{\Sigma}_{1s}^{\top}) / n_l^2 + \sum_{s=1}^{q_l} (\hat{\gamma}_{1s}^{\top} \hat{\mu})^4 / n_l^2} \delta_1^l \right] \to 0
\]

and

\[
P\left( \sum_{s=1}^{q_l} w_s^2 - 1 \geq \tau \right) \leq \frac{\sum_{s=1}^{q_l} E(w_s^2 w_t^2) - 1}{\tau^2} \to 0
\]

so that \(\sum_{s=1}^{q_l} E\{w_s^2 I(w_s^2 \geq \tau)\} \leq \sum_{s=1}^{q_l} E(w_s^2) / \tau \to 0\) and \(\sum_{s=1}^{q_l} w_s^2 = 1 + o_P(1)\). Hence, by using the martingale central limit theorem, we obtain that \(\sum_{s=1}^{q_l} w_s \Rightarrow N(0, 1)\) as \(m \to \infty\) under (A-ii), (C-iv) and (C-vii). We conclude the result when \(i = 1\). For the case when \(i = 2\), we can have the same arguments. The proof is completed. \(\square\)
Proofs of Corollaries 3.1 and 3.2. From Theorems 3.1 and 3.2 and Proposition 3.1, we can claim Corollaries 3.1 and 3.2 straightforwardly. □

Lemma B.3. Assume that when \( x_0 \in \pi_i \) for \( i = 1, 2 \)

\[
\begin{align*}
(B.9) & \quad \text{tr} \{ (x_0 - \mu_i)(x_0 - \mu_i)^T - \Sigma_i \} (\hat{B}_1 - \hat{B}_2) \} = o_P(\kappa); \\
(B.10) & \quad \text{tr} \{ \Sigma_i (\hat{B}_1 - \hat{B}_2) \} + \log |\hat{A}_1 A_1^{-1}| - \log |\hat{A}_2 A_2^{-1}| = o_P(\kappa); \quad \text{and} \\
(B.11) & \quad \{2(x_0 - \mu_i) + (-1)^{i+1} \mu_{12}\}^T \hat{B}_j \mu_{12} = o_P(\kappa) \quad (j \neq i) \\
& \quad \text{and} \quad (p/n_1^{1/2})||\hat{B}_1|| = o_P(\kappa), \quad l = 1, 2,
\end{align*}
\]

where \( \kappa = \Delta_{\min} \) or \( \kappa = \delta_{\min} \). Then, (4.1) holds.

Proof of Lemma B.3. We consider the case when \( x_0 \in \pi_1 \). We have that

\[
W_i(\hat{A}_1) - W_i(A_1) - W_j(A_2) + W_j(A_2) \\
= \text{tr} \{ (x_0 - \mu_1)(x_0 - \mu_1)^T - \Sigma_1 \} (\hat{B}_1 - \hat{B}_2) \\
+ \text{tr} \{ \Sigma_1 (\hat{B}_1 - \hat{B}_2) \} + \log |\hat{A}_1 A_1^{-1}| - \log |\hat{A}_2 A_2^{-1}| \\
+ \frac{2}{l=1}(-1)^{l+1}\text{tr}\{2(x_0 - \mu_1 - (\overline{x}_{in} - \mu_1)/2)(\mu_l - \overline{x}_{in})^T - S_{in}/n_l\} \hat{B}_l.
\]

Note that \( \text{tr}(S_{in}) = O_P(p) \), \( ||\overline{x}_{in} - \mu_1||^2 \leq ||\overline{x}_{in} - \mu_l||^2 + ||\mu_l - \mu_1||^2 = ||\mu_l - \mu_1||^2 + O_P(p/n_l) \) and \( ||x_0 - \mu_1 - (\overline{x}_{in} - \mu_1)/2||^2 \leq ||x_0 - \mu_1||^2 + ||\overline{x}_{in} - \mu_1||^2 + ||\mu_1 - \mu_l||^2 = O_P(p), \) \( l = 1, 2 \), from the facts that \( E(||x_0 - \mu_1||^2) = \text{tr}(\Sigma_1) \), \( E\{\text{tr}(S_{in})\} = \text{tr}(\Sigma_i) \), \( E(||\overline{x}_{in} - \mu_l||^2) = \text{tr}(\Sigma_i)/n_l \), \( \text{tr}(\Sigma_i) = O(p), \) \( i, 2, \) and \( ||\mu_{12}||^2 = O(P) \). Then, we have that for \( l = 1, 2 \)

\[
||x_0 - \mu_1 - (\overline{x}_{in} - \mu_1)/2|| \cdot ||\overline{x}_{in} - \mu_l|| \cdot ||\hat{B}_l|| + \text{tr}(S_{in})||\hat{B}_l||/n_l \\
= O_P\{p/n_1^{1/2}\}||\hat{B}_l||.
\]

Also, we have that \( ||\overline{x}_{2n_2} - \mu_{12}|| \hat{B}_2 \mu_{12} = O_P\{p/n_1^{1/2}\}||\hat{B}_l|| \). Thus it holds that \( \sum_{l=1}^{2}(-1)^{l+1}\text{tr}\{2(x_0 - \mu_1 - (\overline{x}_{in} - \mu_1)/2)(\mu_l - \overline{x}_{in})^T - S_{in}/n_l\} \hat{B}_l = -\{2(x_0 - \mu_1) + \mu_{12}\}^T \hat{B}_2 \mu_{12} + O_P\{p/n_1^{1/2}\}||\hat{B}_l|| \}. Hence, it concludes the result when \( x_0 \in \pi_1 \). For the case when \( x_0 \in \pi_2 \), we can have the same arguments. The proof is completed. □

Proof of Propositions 4.1 and 4.2. We consider the case when \( x_0 \in \pi_1 \). Similar to Proof of Lemma B.3, we can claim that \( ||\{2(x_0 - \mu_1) + \)}
\[ \begin{align*}
& \mu_{12}^T \hat{B}_2 \mu_{12} \leq \|2(x_0 - \mu_1) + \mu_{12} \| \cdot \| \mu_{12} \| = \| \hat{B}_2 \| = O_P(\|B_2\|) = O_P(p^{1/2}) \] because \( \| \mu_{12} \|^2 = O(p) \) and \( \|2(x_0 - \mu_1) + \mu_{12}\|^2 = O_P(p) \), so that (B.11) implies (C-viii) and (C-ix). Note that (B.9) and (B.10) naturally hold when \( \hat{A}_1 = \hat{A}_2 \) and \( A_1 = A_2 \). Hence, from Lemma B.3, it concludes the result of Proposition 4.2 when \( x_0 \in \pi_1 \).

Next, we consider (B.9) and the first term of (B.10). We have that for \( l = 1, 2 \)
\[ |\text{tr}\{\Sigma_1 \hat{B}_l\}| \leq \text{tr}(\Sigma_1) \| \hat{B}_l \| = O_P(\|B_l\|) \] and
\[ |\text{tr}\{\{x_0 - \mu_1\}(x_0 - \mu_1)^T - \Sigma_1\} \hat{B}_l\}| \leq \|x_0 - \mu_1\|^2 \| \hat{B}_l \| + \text{tr}(\Sigma_1) \| \hat{B}_l \| = O_P(\|B_l\|). \]

Finally, we consider log \( \| \hat{A}_l A_l^{-1} \|, l = 1, 2 \), in (B.10). Let \( e_p \) be an arbitrary (random) \( p \)-vector such that \( \| e_p \| = 1 \). Note that \( \| e_p^T A_l^{1/2} \| \in (0, \infty) \) as \( p \to \infty \) under \( A(A_l) \in (0, \infty) \) as \( p \to \infty \). Thus we have that
\[ e_p^T A_l^{1/2} \hat{B}_l A_l^{1/2} e_p = e_p^T A_l^{1/2} \hat{A}_l^{-1} A_l^{1/2} e_p - 1 = O_P(\|\hat{B}_l\|), \]
so that \( \lambda_{\min}(A_l^{1/2} \hat{A}_l^{-1} A_l^{1/2}) - 1 = O_P(\|B_l\|) \) and \( \lambda_{\max}(A_l^{1/2} \hat{A}_l^{-1} A_l^{1/2}) - 1 = O_P(\|B_l\|) \). Hence, under \( \|B_l\| = o_P(1) \), it holds that for \( l = 1, 2 \)
\[ \log \| \hat{A}_l A_l^{-1} \| = \log \| A_l^{1/2} \hat{A}_l^{-1} A_l^{1/2} \| = O_P(\|B_l\|). \]
Note that \( \Delta_{\min} = O(p) \) and \( \delta_{\min} = O(p) \) under \( \lambda(A_l) \in (0, \infty) \) as \( p \to \infty \) for \( i = 1, 2 \). Then, under (C-viii), it holds that \( \| B_l \| = o_P(1) \) for \( l = 1, 2 \). Hence, (C-viii) implies (B.9) and (B.10). It concludes the result of Proposition 4.1 when \( x_0 \in \pi_1 \). For the case when \( x_0 \in \pi_2 \), we can have the same arguments. The proof is completed.

**Proof of Corollary 4.1.** Under (A-i), we have that \( \text{Var}\{\text{tr}(S_{tn})\} = O(\text{tr}(\Sigma_i^2)/n_t) \), \( l = 1, 2 \), so that \( \text{tr}(S_{tn}) = \text{tr}(\Sigma_i) + o_P\{\text{tr}(\Sigma_i^2)/n_t\} \). Then, it holds under (C-i) that \( \text{tr}(S_{tn}) = \text{tr}(\Sigma_i) + o_P(\Delta_{\min(I_l)}) = \text{tr}(\Sigma_i)\{1 + o_P(1)\} \) and \( \text{tr}(\Sigma_i^2)/n_t = o(\Delta_{\min(I_l)}/p^2) = o(1) \) for \( l = 1, 2 \) because \( \Delta_{\min(I_l)} = O(p) \). Thus, we have that under (A-i) and (C-i)
\[ \|\hat{B}_l\| = \|p/\text{tr}(S_{tn}) - p/\text{tr}(\Sigma_i)\| = \frac{p|\text{tr}(S_{tn}) - \text{tr}(\Sigma_i)|}{\text{tr}(S_{tn})\text{tr}(\Sigma_i)} \]
(B.12) \[ = O_P\{\text{tr}(\Sigma_i^2)/n_t\}^{1/2}/\text{tr}(S_{tn}) = o_P(\Delta_{\min(I_l)}/p) = o_P(1), \]
so that \( p\|\hat{B}_l\| = o_P(\Delta_{\min(I_l)}) \). Note that \( \lambda_{\max}(A_i) = \lambda_{\min}(A_i) = \text{tr}(\Sigma_i)/p \in (0, \infty) \) as \( p \to \infty \). Thus, from Corollary 2.1 and Proposition 4.1, it concludes the result.

\[ \square \]
Proof of Corollary 4.2. We consider the case when \( x_0 \in \pi_i \). Note that \( \text{tr}(S_{jn_i})/\text{tr}(\Sigma_j) = 1 + O_P\{|(\text{tr}(\Sigma_j^2)/n_j)^{1/2}/p\} = 1 + o_P(1), \ j = 1, 2, \) as \( m \to \infty \) and \( \text{tr}\{(x_0 - \mu_i)(x_0 - \mu_i)^T - \Sigma_i\} = O_P(\text{tr}(\Sigma_i^2)^{1/2}) \) under (A-i). Also, note that \( \text{tr}(\Sigma_i^2)\text{tr}(\Sigma_j^2) \leq \lambda_{ij}\lambda_{ji}\text{tr}(\Sigma_i)\text{tr}(\Sigma_j) = O(n_{\min}^\delta_{m(II)}^b)^2), \ l = 1, 2 \) under (C-iv'). Then, from (B.12), it holds that for \( l = 1, 2 \)

\[
\text{tr}\{(x_0 - \mu_i)(x_0 - \mu_i)^T - \Sigma_i\} B_i \]

\[
= p\frac{\text{tr}(\Sigma_i) - \text{tr}(S_{jn_i})}{\text{tr}(\Sigma_i)\text{tr}(S_{jn_i})}\text{tr}\{(x_0 - \mu_i)(x_0 - \mu_i)^T - \Sigma_i\}
\]

(B.13) \[
= O_P\{(\text{tr}(\Sigma_i^2)/n_i)^{1/2}/p\} = o_P(\delta_{\min(II)}), \text{ and}
\]

\[
p\|\hat{B}_i\|/n_i^{1/2} = O_P(\text{tr}(\Sigma_i^2)^{1/2}/n_i) = o_P(\delta_{\min(II)}).\]

Similarly, from (B.12), under (A-i) and (C-iv'), we have that for \( j \neq i \)

\[
\{2(x_0 - \mu_i) + (-1)^{i+1}\mu_{12}\}^T \hat{B}_j \mu_{12}
\]

\[
= O_P\{(\mu_{12}^T \Sigma_i \mu_{12}/n_j)^{1/2}\} + O_P\{|(\text{tr}(\Sigma_i^2)/n_j)^{1/2}||\mu_{12}\|^2/p\}
\]

\[
= O_P\{(\lambda_{i1}||\mu_{12}\|^2/n_j)^{1/2}\} + O_P\{(\lambda_{j1}||\mu_{12}\|^2/n_j)^{1/2}\} = o_P(\delta_{\min(II)})\]

from the facts that \( \mu_{12}^T \Sigma_i \mu_{12} \leq \lambda_{i1}||\mu_{12}\|^2, \text{tr}(\Sigma_i^2) = O(\lambda_{j1}p) \) and \( ||\mu_{12}\|^2 = O(p) \). On the other hand, under (A-i) and (C-iv'), from (B.12), we have that for \( l = 1, 2 \)

\[
\log\{\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i})\} = (\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) - 1) + O_P\{|(\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) - 1)^2\}
\]

\[
= (\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) - 1) + O_P\{|(\text{tr}(\Sigma_i^2)/(np^2)\}
\]

from the facts that \( \text{tr}(\Sigma_i^2)/p = O(\lambda_{i1}) \) and \( \text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) = 1 + o_P(1) \). Then, under (A-i) and (C-iv'), it holds that

\[
\text{tr}(\Sigma_i \hat{B}_i) - \log|A_i \hat{A}_i^{-1}|
\]

\[
= p\{\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) - 1\} - p\log\{\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i})\} = o_P(\delta_{\min(II)}).
\]

Similarly, under (A-i) and (C-iv'), we have that

\[
\text{tr}(\Sigma_i \hat{B}_j) - \log|A_j \hat{A}_j^{-1}|
\]

\[
= p\{\text{tr}(\Sigma_i)/\text{tr}(S_{jn_i}) - 1\}(\text{tr}(\Sigma_j)/\text{tr}(S_{jn_j}) - 1) + o_P(\delta_{\min(II)})
\]

(B.14) \[
= O_P\{|(\text{tr}(\Sigma_i)/\text{tr}(S_{jn_j}) - 1)(\text{tr}(\Sigma_j^2)/n_j)^{1/2}\} + o_P(\delta_{\min(II)}).
\]

By combining (B.13) to (B.14) with Lemma B.3 and Corollary 3.1, we can claim the result. \qed
Proofs of Corollaries 4.3 and 4.4. We can write that

\[(B.15) \quad s_{i o i n (j)} = n_i s_{o i n (j)}/(n_i - 1) - n_i (\bar{x}_{ijn_i} - \mu_{ij})^2/(n_i - 1),\]

where \(s_{o i n (j)} = \sum_{t=1}^{n_i} (x_{ijt} - \mu_{ij})^2/n_i\). Then, under (A-iii), for any \(x\) satisfying \(x \to \infty\) and \(x = o(n_i^{1/2})\) as \(n_i \to \infty\), we have that as \(n_i \to \infty\)

\[P(n_i^{1/2}|s_{o i n (j)} - \sigma_{i (j)}/\eta_{i (j)}^{1/2} \geq x) = \exp \left( - \frac{x^2}{2} \{1 + o(1)\} \right).\]

Refer to Chapter 6 in De la Peña [10] for the details of this result. Let \(\tau_{ij} = M(\eta_{i (j)} n_i^{-1} \log p)^{1/2}\) for \(j = 1, \ldots, p\), where \(M > 2^{1/2}\). Then, under \(n_i^{-1} \log p = o(1)\), it holds that as \(p \to \infty\)

\[\sum_{j=1}^{p} P(|s_{o i n (j)} - \sigma_{i (j)}| \geq \tau_{ij}) = \sum_{j=1}^{p} P(n_i^{1/2}|s_{o i n (j)} - \sigma_{i (j)}|/\eta_{i (j)}^{1/2} \geq M(\log p)^{1/2}) \]

\[(B.16) \quad = \sum_{s=1}^{p} \exp \left( - \frac{M^2 \log p}{2} \{1 + o(1)\} \right) \to 0.\]

Next, we consider the second term of (B.15). Let \(u_{ij} = t_{ij}(\sigma_{i (j)}/\eta_{i (j)})^{1/2}\) and \(x_{o i j l} = |x_{ijl} - \mu_{ij}|\) for \(j = 1, \ldots, p\) \((l = 1, \ldots, n_i)\). Then, we have that for \(j = 1, \ldots, p\)

\[E\{\exp(u_{ij} x_{o i j l}/\sigma_{i (j)}/\sigma_{i (j)})\} \]

\[= E\{\exp(u_{ij} x_{o i j l}/\sigma_{i (j)})I(x_{o i j l} \leq 1)\} + E\{\exp(u_{ij} x_{o i j l}/\sigma_{i (j)})I(x_{o i j l} > 1)\} \]

\[\leq \exp(u_{ij}/\sigma_{i (j)}) + E\{\exp(u_{ij} x_{o i j l}^2/\sigma_{i (j)})\} \]

\[\leq \exp(u_{ij}/\sigma_{i (j)}) + E\{\exp(t_{is} x_{o i j l}/\eta_{i (j)})\},\]

so that \(\limsup_{p \to \infty} E\{\exp(u_{ij} x_{o i j l}/\sigma_{i (j)})\} < \infty\) under (A-iii). Thus, in a way similar to (B.16), we have that

\[(B.17) \quad \sum_{j=1}^{p} P(|\bar{x}_{ijn_i} - \mu_{ij}| \geq \tau_{2j}) = \sum_{j=1}^{p} P(n_i^{1/2}|\bar{x}_{ijn_i} - \mu_{ij}|/\sigma_{i (j)}/\sigma_{i (j)} \geq M(\log p)^{1/2}) \to 0\]

for \(\tau_{2j} = M(\sigma_{i (j)} n_i^{-1} \log p)^{1/2}, j = 1, \ldots, p\). By combining (B.16) and (B.17)
with (B.15), under $n_i^{-1} \log p = o(1)$ and (A-iii), we have that
\[
\sum_{j=1}^{p} P\{|s_{in_i(j)} - n_i \sigma_i(j)/(n_i - 1)| \geq n_i(\tau_{1j} + \tau_{2j}^2)/(n_i - 1)\} \\
\leq \sum_{j=1}^{p} P(|s_{in_i(j)} - \sigma_i(j)| + |\overline{x}_{ijn_i} - \mu_{ij}|^2 \geq \tau_{1j} + \tau_{2j}^2) \\
\leq \sum_{j=1}^{p} P(|s_{in_i(j)} - \sigma_i(j)| \geq \tau_{1j}) + \sum_{j=1}^{p} P(|\overline{x}_{ijn_i} - \mu_{ij}|^2 \geq \tau_{2j}^2) \to 0.
\]

Note that $n_i \sigma_i(j)/(n_i - 1) = \sigma_i(j) + o(n_i^{-1/2})$ and $\tau_{2j}^2 = o(\tau_{1j})$ under $n_i^{-1} \log p = o(1)$. Thus we have that $\max_{j=1,...,p}\{|s_{in_i(j)} - \sigma_i(j)|\} = O_P(\max_{j=1,...,p} \tau_{1j})$ under $n_i^{-1} \log p = o(1)$ and (A-iii), so that
\[
\max_{j=1,...,p} \{|s_{in_i(j)} - \sigma_i(j)|\} = O_P\{(n_i^{-1} \log p)^{1/2}\}.
\]

Then, for $i = 1, 2$, it holds that under $n_i^{-1} \log p = o(1)$
\[
||\hat{B}_i|| = ||S_{i(d)}^{-1} - \Sigma_{i(d)}^{-1}|| = \max_{j=1,...,p} \{|s_{in_i(j)} - \sigma_i(j)|/(s_{in_i(j)} \sigma_i(j))\} \\
= O_P\{(n_i^{-1} \log p)^{1/2}\} = o_P(1).
\]

Then, it follows that (C-i') holds under (4.4), and (C-iv') holds under $p^2 \log p/(n_{\min} \delta_{\min(III)}) = o(1)$. From the facts that $\Delta_{\min(III)} = O(p)$ and $\delta_{\min(III)} = O(p)$, note that $n_i^{-1} \log p = o(1)$ under either (4.4) or $p^2 \log p/(n_{\min} \delta_{\min(III)}) = o(1)$. Then, by combining (B.19) with Proposition 4.1, Corollaries 2.1 and 3.1, we can claim the results of Corollaries 4.3 and 4.4. \]

**Proofs of Corollary 4.5.** First, note that $s_{n(j)} - \sigma(j) = \sum_{i=1}^{n_i} (n_i - 1)\{s_{in_i(j)} - \sigma_i(j)/\sum_{i=1}^{n_i} n_i - 2\}$. From (B.18), we obtain that $\max_{j=1,...,p}\{|s_{n(j)} - \sigma(j)|\} = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}$ under $n_{\min}^{-1} \log p = o(1)$ and (A-iii). Thus it follows that $||S_{n(d)}^{-1} - \Sigma_{n(d)}^{-1}|| = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}$. Note that $\Delta_{(III)}/||\mu_{12}||^2 \in (0, \infty)$ as $p \to \infty$. Then, by combining Theorem 2.1 with Propositions 2.1 and 4.2, we can claim the results of Corollary 4.5. \]

**Proofs of Corollary 4.6.** Let $S_{oin_i} = \sum_{i=1}^{n_i} (x_{il} - \mu_i)(x_{il} - \mu_i)^T/n_i$ and denote its $(r, s)$ element by $s_{oin_i(rs)}$ for $r, s = 1, ..., p$ ($i = 1, 2$). Let $u_{i(rs)} = \min\{t_{ir}/t_{i(s), t_{is}/t_{i(s)}}\}^{1/2}$ for $r, s = 1, ..., p$ ($i = 1, 2$), where $x_{oijl} = \ldots$
\[ |x_{ij} - \mu_{ij}|, \; j = 1, \ldots, p. \] Then, we have that for \( r, s = 1, \ldots, p \)
\[
\mathbb{E}\{\exp(u_{is}x_{oirl}x_{oisl} - \sigma_{i(sl)})/\eta_i(r, s)^{1/2}\}
\] \[
\leq \mathbb{E}\{\exp(u_{is}x_{oirl}^2 + x_{oisl}^2 + \sigma_i(sl))/\eta_i(r, s)^{1/2}\}
\] \[
\leq \exp(u_{is}\sigma_i(sl)/\eta_i(r, s))\mathbb{E}\{\exp(t_{ir}x_{oirl}^2/(2\eta_i(r)^{1/2}))\exp(t_{is}x_{oisl}^2/(2\eta_i(s)^{1/2}))\}
\] \[
\leq \exp(u_{is}\sigma_i(sl)/\eta_i(r, s))\mathbb{E}\{\exp(t_{ir}x_{oirl}^2)/\eta_i(r)^{1/2}\}\mathbb{E}\{\exp(t_{is}x_{oisl}^2)/\eta_i(s)^{1/2}\},
\]
so that \( \limsup_{p \to \infty} \mathbb{E}\{\exp(u_{is}x_{oirl}x_{oisl} - \sigma_i(sl))/\eta_i(r, s)^{1/2}\} < \infty \) under (A-iii).

Note that \( s_{im_i(sl)} = n_is_{oin_i(sl)}/(n_i - 1) - n_i(\overline{x}_{irm_i} - \mu_{ir})(\overline{x}_{ism_i} - \mu_{is})/(n_i - 1), \) where \( s_{im_i(sl)} \) is the \((r, s)\) element of \( S_{im_i}. \) Also, note that \( \eta_i(sl) \in (0, \infty) \) as \( p \to \infty \) under (A-iii) and \( \liminf_{p \to \infty} \eta_i(sl) > 0 \) for all \( r, s, \) from the fact that \( \eta_i(sl) \leq \{(\eta_i(r) + \sigma_i^2(sl))(\eta_i(s) + \sigma_i^2(sl))\}^{1/2}. \) In a way similar to (B.16) and (B.17), under \( n_i^{-1}\log p = o(1), \) (A-iii) and \( \liminf_{p \to \infty} \eta_i(sl) > 0 \) for all \( r, s, \) we have that
\[
\sum_{r,s=1}^{p} P\{|s_{im_i(sl)} - n_i\sigma_i(sl)/(n_i - 1)| \geq n_i(\tau_1(sl) + \tau_2(sl))/(n_i - 1)\}
\leq \sum_{r,s=1}^{p} \{P(|s_{oin_i(sl)} - \sigma_i(sl)| \geq \tau_1(sl)) + P(|\overline{x}_{irm_i} - \mu_{ir}|/\sigma_i(sl)_{ir} \geq \tau_1(sl) + \tau_2(sl))\}
\leq \sum_{r,s=1}^{p} P(|\overline{x}_{irm_i} - \mu_{ir}|^2 + |\overline{x}_{ism_i} - \mu_{sr}|^2 \geq \tau_2(sl))\} + o(1) \to 0
\]
for \( \tau_1(sl) = M(\eta_i(sl)n_i^{-1}\log p)^{1/2} \) and \( \tau_2(sl) = M^2\{(\sigma_i(sl)_{ir} + \sigma_i(sl)_{is})n_i^{-1}\log p\}, \) \( r, s, 1, \ldots, p, \) where \( M > 2^{1/2}. \) Thus it holds that \( \max_{r,s=1,\ldots,p}|s_{im_i(sl)} - \sigma_i(sl)| = O_P(\max_{r,s=1,\ldots,p}\tau_1(sl)) \) because \( \tau_2(sl) = o(\tau_1(sl)), \) so that
\[
\max_{r,s=1,\ldots,p}\{|s_{im_i(sl)} - \sigma_i(sl)| = O_P\{n_i^{-1}\log p\}^{1/2}\}.
\]

Here, from the equations (A1) and (A2) in Bickel and Levina [7], we have that \( \|M\| \leq \max_{s,t=1,\ldots,p} |m_{st}| \) for any symmetric matrix \( M, \) where \( m_{st} \) is the \((s, t)\) element of \( M. \) From (B.20), we have that
\[
\|S_{im_i} - \Sigma_i\| = O_P\{p(n_i^{-1}\log p)^{1/2}\} = o_P(1)
\]
under \( n_i^{-1}\log p = o(1), \) (A-iii) and \( \liminf_{p \to \infty} \eta_i(sl) > 0 \) for all \( r, s, \) Then, under \( \lambda(\Sigma_i) \in (0, \infty) \) as \( p \to \infty, \) we can claim that \( \lambda(S_{im_i}) \in (0, \infty) \) in probability. Thus it holds that \( \|e_p^T\Sigma_i^{-1}\| \in (0, \infty) \) and \( \|e_p^T S_{im_i}^{-1}\| \in (0, \infty) \) in probability, where \( e_p \) is an arbitrary (random) \( p \)-vector such that \( \|e_p\| = 1. \) Then,
from (B.21), we have that 
\[ e_p^T \Sigma_i^{-1} (S_{in_i} - \Sigma_i) S_{in_i}^{-1} e_p = e_p^T (\Sigma_i^{-1} - S_{in_i}^{-1}) e_p = O_P \{ p(n_i^{-1} \log p)^{1/2} \} \]
under \( n_i^{-1} p^2 \log p = o(1) \), (A-iii) and \( \liminf_{p \to \infty} \eta_i (r_s) > 0 \) for all \( r, s \), so that \( \| \hat{B}_i \| = O_P \{ p(n_i^{-1} \log p)^{1/2} \} = o_P(1) \). Note that (C-i') and (C-ii') hold under the conditions of Corollary 4.6. Also, note that \( \text{tr} \{ (I_p - \Sigma_2 \Sigma_j^{-1})^2 \} = O(p) \) \( (j \neq i) \) under \( \lambda (\Sigma_i) \in (0, \infty) \) as \( p \to \infty \). By combining Corollary 3.2 with Proposition 4.1, we can claim the result of Corollary 4.6.

Proof of Corollary 5.1. By using Theorem 5.1, we can claim the result straightforwardly.

Proof of Corollary 5.2. Let \( \Sigma_j(d)_* = \text{diag}(\sigma_j(r_1), ..., \sigma_j(r_{p_*}))) \) and \( S_j(d)_* = \text{diag}(s_{j1}(r_1), ..., s_{jn}(r_{p_*})) \) for \( j = 1, 2 \), where \( D = \{ r_1, ..., r_{p_*} \} \) and \( \hat{D} = \{ \hat{r}_1, ..., \hat{r}_{p_*} \} \). Let us write that for \( j = 1, 2 \)

\[ W_j(\Sigma_j(d)_*) = \sum_{t \in \hat{D}} \{ (x_{0t} - \bar{x}_{jn(\hat{r})})^2 / \sigma_j(t) - s_{jn(\hat{r})}/(\sigma_j(t))n_j + \log \sigma_j(t) \}. \]

Note that \( E \{ W_j(\Sigma_j(d)_*) \} - E \{ W_i(\Sigma_i(d)_*) \} = \Delta_{\text{III}}(j \neq i) \) when \( x_0 \in \pi_i \). Also note that \( \liminf_{p \to \infty} \Delta_{\text{III}} \) \( (j \neq i) \) \( p_* > 0 \) under \( \liminf_{p \to \infty} \theta_r > 0 \) for all \( r \in D \). If \( \lambda_{\text{max}}(\Sigma_i) = o(p_*) \), (C-i') and (C-ii') hold for \( \Sigma_j, j = 1, 2 \). On the other hand, in a way similar to (B.19), when \( \hat{D} = D \), under \( n_j^{-1} \log p = o(1) \) and (A-iii), it holds that \( \| S_j^{-1}(d)_* - \Sigma^{-1}_j(d)_* \| = O_P \{ (n_j^{-1} \log p)^{1/2} \} \). Hence, we have that \( p_* \| S_j^{-1}(d)_* - \Sigma^{-1}_j(d)_* \| = O_P \{ (n_j \log p)^{1/2} \} \). By combining Corollary 5.1 with Propositions 2.1 and 4.1, we can claim the result.

REFERENCES

[1] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. Sequential Anal. (Editor’s special invited paper) 30, 356–399.

[2] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. Ann. Inst. Statist. Math. 66, 983–1010.

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

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

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

[6] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. Ann. Statist. 36, 2577–2604.
[7] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* 36, 199–227.

[8] Cai, T.T., Liu, W. and Luo, X. (2011). A constrained \( \ell_1 \) minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* 106, 594–607.

[9] Cai, T.T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* 106, 1566–1577.

[10] De la Peña, V.H., Lai, T.L. and Shao, Q.M. (2009). *Self-Normalized Processes*. Springer, Berlin Heidelberg.

[11] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97, 77–87.

[12] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* 36, 2605–2637.

[13] Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Statist. Soc. Ser. B* 74, 745–771.

[14] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

[15] Hall, P., Marron, J.S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. Ser. B* 67, 427–444.

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

[17] Li, Q. and Shao, J. (2013). Sparse quadratic discriminant analysis for high dimensional data. *Statist. Sinica*, to appear. doi: 10.5705/ss.2013.150.

[18] Marron, J.S., Todd, M.J. and Ahn, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* 102, 1267–1271.

[19] McLeish, D.L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* 2 620–628.

[20] Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* 39 1241–1265.

[21] Tan, A.K., Naiman, D.Q., Xu, L., Winslow, R.L. and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21, 3896–3904.

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

[23] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.* 122 334–354.